STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 8
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (i): Planning and management of statistical data editing

NEED FOR A HIGH-LEVEL AUXILIARY DATA SERVICE TO IMPROVE THE QUALITY OF EDITING AND IMPUTATION

Contributed Paper

Submitted by Statistics Finland[1]

Abstract: The paper focuses on discussing the need for good auxiliary data when dealing with editing and imputation. A typology of 10 different types of auxiliary variables is given. These variables may be derived both from external sources such as registers and other surveys, and from internal sources, that is, from the same survey considered. Some auxiliary data are aggregated, others are from the micro level. When dealing with editing and imputation micro-level auxiliary data are of highest interest. Some auxiliary data should be available in the beginning of the editing process, for example, when deciding how to exploit selective editing. At the same time, the so-called pre-imputations may be done for facilitating editing. Later, at the estimation phase, final imputations and weightings will be done. It is important to note that available auxiliary data are usually poorer in the beginning of the process than in the end of it. Each time, best possible data should have been exploited. This does not seem to be a common house style in most survey institutes, quite often the same (initial) auxiliary data are tried to use in each step of the process. However, many other alternative variables may be available, and the initial data could be updated for the reference period of the survey. This may lead to fatal biases in estimates, especially in business surveys, where even the basic statistical units (businesses) may be changed dramatically after the sampling selection. In order to exploit exhaustively such data, the paper proposes a specific auxiliary data service system for survey institutes. How to organize such a service should have been next discussed and implemented? The author has no exact proposal to this question but the success in this activity requires some centralization, for example.

The examples in the paper are mainly from business surveys but the problems and issues are fairly similar in other surveys.

I. INTRODUCTION

1. There are various ways to classify tasks needed when providing survey data for users. The following is a list which focuses on requirements in editing and imputation: (i) users' needs, (ii) survey design, (iii) sampling design, (iv) data collection, (v) editing and imputation (pre-imputation, automatic and manual editing, final imputation), (vi) initial weighting (design weights, basic weights), (vii) re-weighting (post-stratification, response propensity modelling, g-weighting, outlier weighting, calibration), (viii) output data (aggregated macro data and micro data for special users), (ix) data integration, e.g. linking and matching files together (post-editing, post-imputation), (x) dissemination.

2. We will not look in detail at all these steps of a survey, but the rest of this paper is concerned mainly with *step (v)*, that is, standard editing and imputation, its focus being more or less between editing and imputation. It

[1] Prepared by Seppo Laaksonen (seppo.laaksonen@stat.fi).

is important, however, to notice that the impact of the editing and imputation work needs to take into account all the following steps. For example, the output data should have been *flagged* with all special operations done for the data, both at aggregate and individual level. Moreover, attention should be paid to the fact that when continuing to exploit several files including the initial survey file, new post-editing and imputation steps may be necessary because otherwise the data may be too 'dirty' to use. This topic needs to receive special attention and will hopefully be studied in more detail at future editing meetings.

3. The rest of this paper is organized so that, in Section II, a typology is given for so-called auxiliary variables, used in the various steps of a survey. Some examples on business data are included as examples in this typology table. The typology gives a systematization for the next sections where the motivation to an auxiliary service is given. We do not give an exact solution to the administration of such a service, because it is institution-dependent. Section IV gives some examples of how to operate with such a system for business survey data. The final section, V, concludes with a few crucial points.

II. A TYPOLOGY OF AUXILIARY VARIABLES

4. The author of this paper has been dealing with various surveys, both concerning human and business entities, over the past 15 years. When trying to improve the quality of survey data, it is sensible to exploit all available information exhaustively. Usually, we denote target variables of a survey by *Y*, whereas the so-called auxiliary variables or covariates are symbolized by *X*. These latter variables may also be survey variables *Y*, in which case, these certain *Y* variables have been exploited for improving the quality of some other survey variables. During those 15 years, I have tried to systematize these auxiliary or X variables. Before this paper I have only once published any systematization[2]. In this earlier paper, I had a typology of 8 types of auxiliary variables, but now I consider it important to extend this number to 10 (see Table 1).

Table 1. A Typology of Auxiliary Variables in Surveys with Business Survey Examples
(*t* = survey period, *t-1* = earlier available period: some months, one year or maybe many years earlier)

Type of Auxiliary Data	Examples (period)	Use
1. Sampling design variables from population level	Sizeband (t-1), Industry class (t-1), Region (t-1).	Designing, Design weighting for sampled units
2. Non-updated sampling design variables from population level	The same as in type 1, new strata may be done (post-strata);	Initial or post-stratified weights for respondents excl. over-coverage based on sample information
3. Updated sampling design variables from population level	The same as the previous but from period t;	Better weights as in the previous, sample and population over-coverage, under-coverage, deaths, births, mergers, splits, re-constructions
4. Other population level data from registers or recent surveys (estimated)	Aggregated register turnover, employment (t-1, t); aggregated turnover from parallel short-term survey (around t)	*Macro editing*, *Macro imputation* G-Weights based on ratio estimation or advanced (modelling) methods (Calibration)
5. Micro data at sample level (respondents, over-coverage, non-respondents) from registers, independent surveys and other *external*	Categorical: sizeband and industry (t, t-1); Continuous: register turnover (t, t-1), register employment (t, t-1), parallel survey turnover (around t) The above	*Micro editing*: error localization, selective editing, … *Imputation*: modelling and task for key variables with missingness Re-weighting: GREG, response

[2] Laaksonen, S. (1999). Weighting and Auxiliary Variables in Sample Surveys. In: G. Brossier and A-M. Dussaix (eds). "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches". Dunod. Paris. pp. 168-180.

sources	ones are available soon (designing time), but some others maybe later	propensity modelling
6. Micro data at respondents' level from *internal sources* (same survey)	In addition to group 5: whatever survey variables from t , e.g. survey turnover, survey employment, survey value added, total output, imputed y value	*Editing* incl. selective editing using 'best guess' (preliminary imputed value = *pre-imputation*); *Imputation*: modelling using auxiliary vbles either independently for each imputation task or sequentially (imputing first missing values of one vble, then the next)
7. Micro data as a sub-sample of non-respondents or respondents	In addition to standard vbles: key variables of the survey concerned	Quality checking Re-weighting, *Imputation*
8. Micro data from the previous waves of the same repeated survey (panel)	Any categorical and continuous variables for the same unit (if unit changed, this should be taken into account) from t-1, t-2,… Note: also changes in weights	*Micro editing* *Imputation* Re-weighting if need for longitudinal analysis (longitudinal weighting)
9.'Super-auxiliary' variables for specific small groups at micro level if possible	Big and other unique businesses are often so special that from the same survey cannot be found reasonably observations for modelling or donors. Hence multi-national data or other super data should be used	*Micro editing*: plausibility checking *Imputation* Outlier weights
10. Hypotheses on the behaviour of variables, based on previous experiences from the same survey, international harmonization purpose, etc.	Distributions (normal, log-normal, binomial, Poisson), link functions, conditions (CMAR, MAR, NMAR), sensitivity, bounds, relevant time series	Models for editing, *imputation*, weighting, outlier detection

5. The examples of Table 1 are from business surveys, but the corresponding variables may be given for social surveys, too. However, when speaking about editing and imputation, as in this paper, some of these variables are not as much focused in social surveys. For example, the changes in statistical units themselves are not so dramatic in standard social surveys than in business surveys. Nevertheless, the households may be changing their composition essentially over a longer period, and in some surveys, a large unit problem may arise, for example, in income and wealth surveys (extremely rich person/household, very poor). A correct handling of outliers is important in all surveys for continuous variables, but especially in business surveys.

III. NEED FOR AUXILIARY DATA SERVICE

6. One may think that the use of auxiliary variables exhaustively is natural for survey statisticians. Or, at least, their own survey has exploited those completely. Or, they say that we have not many auxiliary variables available, because we have no register data, for example. These are good points for discussion, but in my experience, almost always after a more detailed consideration, some new variables or their specifications may be found. I give some examples, anonymously.

An international social survey: the sampling frame of administrative sources (registers, election list, etc.) was available, not up-to-date but not poor. The interviewers collected useful information about over-coverage and non-respondents, but none of this information was saved in any electronic file in order to try to exploit such

data in quality checking or adjustments. The reason: it was their house style, and they thought that any advantage of these auxiliary data cannot be taken.

A household survey: the interviewers contacted most households but no information was saved from refusals, although many of these were willing to tell their household composition, at least. This information would be very useful for analyzing non-respondents. Also, partially completed questionnaires were not exploited, not even saved in the file.

A household panel survey: a quite good population register was available in a country. A certain number of over-coverage and non-responding units was found and saved for the year *t* file, but in the next year, no follow-up for the year *t* non-respondents was done in year *t+1*. This task could be done with low costs at the same time as the normal updating of the survey, when linking register and survey data together. Even, the unit non-response rate for year *t+1* was calculated as average of non-response of initial non-response rate of t and the additional non-response of the second year. This method, naturally, gives too low non-response rates. Respectively, the panel effect cannot be exploited in editing, imputation and weighting.

A survey for elderly people : this kind of survey is sensitive to changes in population, because older people may be moving to service houses, hospitals and the mortality rate for them is rather high. Hence, it is useful to collect as much auxiliary variables both via the interviewing system, and from registers, if available. In our case example, many useful variables were not requested from the register authority although these were available without additional cost. The reason was that these variables were not asked earlier and they thought that there is no need for this survey either. Secondly, the refusals and the non-contacted people were not checked from the fresh register although this was available.

An annual business survey without rotating panel: register turnover and register employment were available for sample designing. The information was more or less old (a delay of a few months for all but for most small businesses much more). This cross-sectional information could be excellent for editing, imputation and weighting, if was up-to-date for the survey period. Such an updating could be done using the newest register data at the estimation phase, but was not done for some reasons. A reason may be administrative because this updating necessitates contacting another department within the NSI. Maybe some additional costs may also be needed if not done at the same time as some other downloading. It should be noted that all information is not as important. It is most useful to check changes in businesses such as real births vs. artificial births, real deaths vs. artificial deaths, mergers and splits, and to concentrate on large and maybe medium-sized businesses.

An annual business survey with rotating panel: longitudinal information has been forgotten, although this could help a lot in both editing and imputation. This also requires checking changes in businesses, so these changes should be considered as key auxiliary variables.

A two-stage business survey of wage and salary earners . This kind of survey uses businesses of certain sectors as a sampling frame, but the smallest businesses are not included in the frame. In principle, all workers of each sample unit are to be surveyed from a certain period. There will be several error sources in such a survey: first-stage unit non-response, second-stage unit non-response, and item non-response, over-coverage and under-coverage. To completely code all these cases into the data file may be an impossible task. This leads to difficulties in using completely auxiliary data from the same source, that is, from the business register and the data collection system. In some countries, the types of other sources may also be exploited: (i) register information from the businesses in annual or short-term surveys, and (ii) register variables from workers derived from a taxation or another register. Both sources are not maybe fully consistent with each other because of the different reference periods, or the concepts of variables may not be exactly the same. But it is possible to find the relationships between these differences and then to exploit those in adjustments.

A monthly (short-term) business survey with few questions such as sales and employment. A quite long time series of the harmonized variables was available, in principle. Some auxiliary information was also used,

mainly from the 1-2 previous months, maybe also from the corresponding month of the previous year. But a systematic evaluation of a longer series was not done. Secondly, and more problematically, they had difficulties with unit changes and births and deaths. This kind of information was only available from the same survey, that is, for responding businesses. And although it was available, it was difficult to construct a consistent time series. In an optimal situation, the business demographics should be obtainable from the central business register, for example.

Any individual social or business survey has been handled independently of other surveys and censuses. Thus the surveyors have not linked or matched possible useful information from registers, censuses or other surveys with this survey. There are several reasons for this situation: (i) administrative problems, (ii) the linking is not easy, for example, because the different identity codes are used in linkable databases, (iii) after the data linking, new editing checks are often needed, and (iv) there is no time and resources for these extra operations.

7. I cannot say how typical the above-mentioned examples are in various countries, but I have met with these continually even in Finland where there are relatively good possibilities to exploit auxiliary variables from various sources. But this has not been done systematically, it seems to depend on the competence and willingness of the responsible persons of a survey. Hence, in my opinion, it is beneficial to establish some type of auxiliary service system within a survey institute like an NSI. In practice, because business surveys and social surveys are so different, the two sub-systems may be a better solution. This service system does not perhaps need any special unit, but some responsible persons, necessarily. The development of such a system requires an evaluation of the needs of each survey from this point of view, thus what improvements in editing, imputation, weighting and data analysis may be achieved with more complete and qualified data files. There is also a need to harmonize identity codes, variable labels, classifications and other metadata. A special problem is how to harmonize such operators as non-applicable value, missing value, fatal error, another error, warning, and initial, edited and imputed value. The success in these tasks, on the other hand, is supposed to be good IT solutions, with as much automated mechanisms as possible.

8. Finally, I want to discuss the standards of data files under a good system of auxiliary data services. In my experience, a typical micro file available for users consists of the respondents only, and variables Y and sampling weights are only included. Maybe this is enough for some users, but not for sophisticated users such as (i) a methodologist who tries to check and improve the data quality, and (ii) an advanced analyst who wants to exploit exhaustively the available data. Thus, the maximum number of variables X, also for unit non-respondents and over-coverage units, should be linked (or linkable) with the proper survey file. The easiest for a user may be such a solution that aggregate data are included in the same file or system (e.g. regional totals of X would be put in each region). Naturally, the full information of sampling design should be included, not only the sampling weights. Attention should also be paid to flagging the special values such as imputed and others mentioned above. The maximal file may be very big, especially if various levels such as household plus household members or business unit plus employees are in the same wholeness. This may be passed over correctly with various technical tools, but I do exclude this discussion from this paper. In practice, a user does not analyse the full data set in any separate operation but instead chooses just an optimal reduced set for each handling. This is easier if the good indicators for such a purpose are included in the file.

IV. USE OF AUXILIARY DATA FOR EDITING AND IMPUTATION: Illustrative Model-Based Examples For Structural Business Survey

9. The examples of this section do not show real empirical results, although the author has done a lot of such analogous work in practice. The purpose of this more general presentation is to illustrate how the opportunities to exploit auxiliary data vary during the survey process which in this case concerns an annual business survey.

10. In Section I we have mentioned that the editing and imputation process includes 3 main steps: (i) pre-imputation, (ii) editing and (iii) final imputation. Of course, the number of steps may be much longer if more details are wanted to give, but we accept these main steps especially when speaking about auxiliary data service. Under those steps, the three major tasks are needed, that is, A. model building, B. Error localization (and editing) and C. imputation. We next discuss these tasks.

A. Model Building

11. The common feature in each step is *model building* and this is thus the first step after the data collection. So, we have to build a model both for pre-imputation, editing and final imputation. Each model takes advantage of best possible auxiliary variables available at that time of the process. The model type depends, of course, on various factors, but whatever linear or non-linear model may be considered.

Examples on models: good guess, known function, edit rules (gates), linear regression model with constant term, linear regression with noise term, linear regression with constant and noise term, linear regression with slope (and noise), linear regression with constant, slopes and noise, logistic regression, Poisson regression, multi-level modelling, non-parametric regression models, regression tree, classification tree, neural nets such as SOM (self-organizing maps), MLP (multi-layer-perception), CMM (correlation matrix memory), SVM (support vector machine).

12. In the editing step, the errors have been localized as well as possible, and then corrected manually or automatically. If an error is minor with high probability, it may be allowed.

13. How to specify such a model? When considering the missingness or erroneousness of a single variable y it simply means either
(a) That the dependent variable is just this y or its good transformation (e.g. logarithm) and all possible auxiliary variables x have been attempted as explanatory variables. The model has been fitted for the data set without missing and erroneous values, thus for the so-called clean data set. Note, however, that variables x should be available for the full data set, although not used in modelling. If there is a missing x value, this may be imputed first, or the respective unit have not included in the model building.
or
(b) That the dependent variable is categorical so that
= 1 if the value is missing or erroneous and
= 0 if the value is correct
(more categories may be used too but I have not done such exercises until now).
The model is fitted for the clean data set in the other sense, that is, variables x should be correct and the categories of variable y as well. Note that in case (b) the data set for fitting is larger.

14. The following are examples for business survey data:
Case (a)
variable y = survey turnover or log(turnover+1) from period t,
variables x (all these may be re-scaled):
= register turnover from period *t-1*, later from period *t* if available,
= register turnover from period *t-2* if the business existed already at time *t-2*,
= survey turnover from period *t-2* if the panel used,
= turnover from a parallel monthly survey, period, e.g., some months from *t*,
= industry class from period *t-1*, maybe from t,
= region from period *t*,
= register employment from *t-1*, *t-2* and maybe from *t*, respectively,
= taxes and possible other information from registers,
= wages paid from register,
= purchases from the same and previous survey, maybe not available for all responded units,

= business demographics indicators such as = new business in *t*, dormant in *t-1* (*t-2*), not in *t*, split after *t-1*, merged after *t-1*, other re-construction

Case (b)

variable *y* = 1 if survey turnover differs from initially coded value of turnover in the training data set, that is, it is erroneous; else = 0;
this is for the error localisation model,
for the missingness model, respectively.

variables *x*: there are similar opportunities as in case (a) but the specification and scaling may be different.

B. Error localization

15. Case (a): A simple way to continue from the estimated model, that has been fitted for the cleaned data set, is to estimate also the confidence intervals for that model and then to predict this model with confidence intervals within the same clean data set, say *y*low* and y**high*. We may consider the values outside this confidence interval as erroneous. On the other hand, we find from the same data set those initial values of variable *y* which are really edited, say $y,^{e}$ and the corresponding units. If those modelled edits and the real edits are equal, our error localisation has been successful in that clean data set. Consequently, we may find the three other cases: no error based on the model and in reality (success case), no error based on the model but error in reality (no success) and error based on the model and no error in reality (no success). Because we may use the training data set, we may benchmark the confidence intervals so that the optimal result may be achieved.

16. Next, we look at the real dirty data and to make the same predictions with the benchmarked confidence intervals to this data set, and will get the predictions for errors. The success in this operation depends, in particular, on how similar are to be the dirty data set on one hand and the training data set on the other. It is assumed that variables *x* are correct; if this not the case, these should have been first to correct. It is possible to add the number possible errors and error checking by extending the confidence interval.

17. Case (b): The model gives now the predicted values for error probability. Typically, logistic regression is used in this estimation. This probability may be compared as in case (a) against the real data, and the decision on error checking done, consequently. The higher the error probability will be used, the more values will become for checking.

Examples for business survey data

18. In principle, the similar variables as given in examples of model building may be attempted, but in practice, less variables with real *x* values are available. This because these *x* values should be available both for the training data set and for the initial survey data set. I suppose that register turnover from *t-1* and *t-2* (maybe from *t*) could be available, at least, and the respective values for register employment and register taxes. Some categorical variables should, in addition, be applicable.

C. Imputation

19. The estimated model gives some type of predicted values for use in the imputation task. These values may be very simple or complex. Basically, there are only three types values: (i) imputation cells or groups, (ii) pure predicted values, (iii) predicted values with (random) noise terms. An imputation cell may be very simple like the whole population or based on a complex multi-dimensional non-parametric model, like in SOM technology.

20. When exploiting the estimated imputation model, the following two alternatives for the imputation task may be attempted (or the mixture of both methods)[3]:

[3] See Laaksonen, S. (2000). Regression-Based Nearest Neighbour Hot Decking. *Computational Statistics* 15, 1, 65-71.

- In case of *model-donor imputation* the imputed values are *directly* derived from a (behavioural) model.

- In case of *real-donor imputation* the imputed values are *directly* derived from a set of observed values, from a real donor respondent, but still are *indirectly* derived from a more or less exactly defined model.

21. Alternative 1: an imputed value is a predicted value of the model, adding a noise term if necessary. Alternative 2: how to choose a donor, it is the big issue:
- Generalising: it is always a value from the neighbourhood.
- Many terms for this method are used, such as random hot decking (random raw with or without replacement), sequential hot decking, nearest neighbour, near neighbour.

22. All these methods thus have either a *deterministic* or *stochastic* feature, or both. A stochastic feature may be included both in the model or the imputation task itself.

23. When using a nearest or another near neighbour method, the nearness metrics may be constructed in various ways, including the exploitation of edit rules as done in software NIM, for example [4]. When using an explicit model like regression model, it is logical to continue to use this information as I have presented in the above-mentioned paper from 2000. Thus: exploit these predicted values with or without noise term for determining the nearness metrics as well (Regression-Based Nearest Neighbour = RBNN). The advantage of this method is its objectivity, the weights for explanatory variables of the model are estimated from the clean data, they are not taken from a black box or from the brain of an imputer. This as all nearness methods may be problematic if a reasonable number of (real) donors is not available in the neighbourhood of the units with missing values. This, for example, is often the case when trying to impute the values for large firms. This just leads to the need for 'super-auxiliary' variables (Table 1).

24. The *model-based nearest neighbour technique* (real-donor method) may be used both for continuous and categorical variables. The former gives the 'continuous' predicted values of variable *y* for all the units, with or without missing values. The latter case also provides the continuous values but for the probability (propensity) of missingness. Also, from these values, the nearest real donor for the missing one can be found without problems, and the values taken to replace the missing one, consequently.

Examples for business survey data

25. Now, imputation may be tried both in an early stage which is called *pre-imputation*, and at the end of the validation process which we call *final imputation*. Pre-imputation gives some preliminary values to help in the editing process. I think that, in principle, all the values of the key variables of the sampled units, may be pre-imputed in order to get some preliminary understanding on the final results at aggregate and at individual level. I have until today not met any NSI where this type of system would be introduced into use. *Please tell me if you know.*

26. I believe that in some regular business surveys, especially, this kind of large-scale pre-imputation system could be useful if too much effort will not be devoted to this task. The system thus should be rather simple and automatic, but maybe not as simple as used in some *selective editing* (or *significance editing*) exercises where a preliminary value is a mean imputed value at industry level. If the survey is panel-based, the previous available value of the same unit is somewhat better (this method is called *cold decking)* unless the business unit has changed essentially. Lawrence and McKenzie (2000, 245-246)[5] use the term 'expected amended value' that may be taken using the editing model as I have demonstrated in this paper. In any case, it

[4] See, e.g. the paper by Claude Poirier in the Cardiff UNECE meeting 2000: A Functional Evaluation of Edit and Imputation Tools.
[5] The General Application of Significance Editing. *Journal of Official Statistics* 16, 243-253.

is best to try with a simple robust model-donor imputation method in this pre-imputation step. When using the respective explanatory variables for turnover as above demonstrated for error localization, it is possible to get the rough imputed values for the significance editing process. I think that this needs more research, I personally, for example, have not finalized any full exercise.

27. For the final imputation, more effort is needed, including:
- Better model specification
- Including attempts to apply various transformations (log, logit, ratios, …)
- Updated auxiliary variables
- New auxiliary variables from other surveys as well
- Careful imputation

28. What Careful Imputation in the case of regression model may mean?
- Large businesses are useful to impute but not necessarily to use these values in the final data set (except for non-key variables). It is best to contact again those, in order obtain the real values.
- The predicted values as in pre-imputation are not maybe reasonable as imputed values, some noise (but with care) should be added.
- But: the predicted values (with or without noise term) may be used as metrics for nearest neighbours, not in the case of large businesses where good real donors are not available as often than for small and medium sized businesses.
- Final imputation is often useful to do within homogenous imputation cells.
- Sampling weights should have been taken into account in final imputation.
- Sequential imputation is becoming more common in business surveys, for example, so that the key variables have been first imputed and the imputed values of these have been used as explanatory variables when imputing non-key variables.
- Make results consistent to each other using edit rules.
- Check the completed results against available benchmarking data (aggregate level).

V. CONCLUSION

29. All available and 'useful' internal or external data related to each survey may and should be used as auxiliary information. It is not always clear how these data could be exploited in the best way because those are not often in the same database. A solution is to establish a system of the so-called *Auxiliary Data Service (ADS)* within an NSI or another survey institute. Currently, such a system exists in each NSI, implicitly at least, but I hope that it could be found more explicitly. I am interested in hearing about this kind of implementation attempts and solutions in future UNECE and other meetings.