**STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

Working Paper No. 5
English only

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**
(Cardiff, United Kingdom, 18-20 October 2000)

Topic I: Management and evaluation of editing and imputation procedures

# EVALUATION CRITERIA FOR EDITING AND IMPUTATION IN EUREDIT

Submitted by Department of Social Statistics, University of Southampton, United Kingdom[1]

**Invited paper**

## I. THE EUREDIT PROJECT

1. EUREDIT, or to give it its full title "The Development and Evaluation of New Methods for Editing and Imputation" is a large research project funded under the FP5 Information Societies Technology Programme of the European Commission. The objectives of EUREDIT are quite ambitious, and include the development of new methods for editing and imputation, the comparative evaluation of these methods together with "standard" methods and the dissemination of these results via a software product as well as publications.

2. The main focus of EUREDIT is editing and imputation for the types of data collected by national statistical agencies. Reflecting this, the participants in EURDEDIT include the UK Office for National Statistics, the Netherlands Central Bureau of Statistics, the Italian National Statistical Institute, Statistics Denmark, Statistic Finland and the Swiss Federal Statistics Office. From a methodological point of view, the main "new" methods that will be investigated within EUREDIT will be those based on the application of neural net and related computationally intensive methods, as well as the application of modern outlier robust statistical methods. The neural net and computationally intensive methods will be developed by the University of York, the University of Jyvaeskylae and Royal Holloway University, with Statistics Denmark, while the outlier robust methods will be largely developed by the University of Southampton and Insiders GmbH, with the Swiss Federal Statistics Office and the Netherlands Central Bureau of Statistics. Numerical Algorithms Group (NAG) will be responsible for creating the software product that will form the key dissemination vehicle for the research output of the project.

3. Since EUREDIT will involve a large number of different institutions all developing methods for editing and imputation, it was recognised right from the project's inception that there would need to be a common core of evaluation procedures that all EUREDIT developers would have to apply to a common core of data sets to ensure comparability of performance. The core of "test" data sets has been put together by the ONS from contributions by different members of the EUREDIT consortium. The purpose of this paper is to describe the core of evaluation criteria that will be applied to edit and imputation outcomes for these data that will be generated by the different methodologies being investigated within EUREDIT. At this stage it is envisaged that a substantial number of these criteria will be incorporated into the EUREDIT software product.

[1] Prepared by Ray Chambers.

4. In what follows we first discuss assessment of editing performance within EUREDIT, followed by imputation performance. Ancillary issues related to special types of data structures commonly found in official statistics data sets are then discussed, as well as the very important issue of practical implementation of an edit and imputation method. A more extensive document setting out the detail of the evaluation formulae to be used in EUREDIT is available from its website (http://www.cs.york.ac.uk/euredit).

## II. WHAT IS EDITING?

5. Editing is the process of detecting errors in statistical data. An error is the difference between a *measured* value for a datum and the corresponding *true* value of this datum. The true value is defined as the value that would have been recorded if an ideal (and expensive) measurement procedure had been used in the data collection process.

6. Editing can be of two different types. *Logical* editing is where the data values of interest have to obey certain pre-defined rules, and editing is the process of checking to see whether this is the case. A data value that fails a logical edit must be wrong. *Statistical* editing on the other hand is concerned with identification of data values that might be wrong. Ideally, it should be highly likely that a data value that fails a statistical edit is wrong, but there is always the chance that in fact it is correct. Since the context in which the edit is applied (e.g. the presence or absence of external information, and its associated quality) modifies the way an edit is classified, EUREDIT will not attempt to distinguish between evaluation of logical editing performance and evaluation of statistical editing performance. It will only be concerned with evaluation of overall editing performance (i.e. detection of data fields with errors).

7. One can also distinguish editing from *error localisation*. The latter corresponds to the process of deciding which of the fields in a particular record that "fail" the edit process should be modified (e.g. parent/child age above). The key aspect of performance here is finding the "smallest" set of fields in a record such that at least one of these fields is in error. This type of evaluation depends on application of the Felligi-Holt principle of minimum change and requires access to the full set of edit rules for the data set of interest. Since it is infeasible to include every possible edit rule with the evaluation data sets being developed in EUREDIT, the evaluation procedures used by EUREDIT will not be concerned with the localisation aspects of editing.

### II.1 Performance Requirements for Statistical Editing

8. Two basic requirements for a good statistical editing procedure have been identified in EUREDIT.

(i) Efficient Error Detection: Subject to constraints on the cost of editing, the editing process should be able to detect virtually all errors in the data set of interest.

(ii) Influential Error Detection: The editing process should be able to detect those errors in the data set that would lead to significant errors in analysis if they were ignored.

### II.2 Efficient Error Detection

9. Typically, the concern is detection of the maximum number of true errors (measured value ≠ true value) in the data set for a specified detection cost. This detection cost rises as the number of incorrect detections (measured value = true value) made increases, while the number of true errors detected obviously decreases as the number of undetected true errors increases. Consequently, EUREDIT will evaluate the error detection performance of an editing procedure in terms of the both the number of

incorrect detections it makes as well as the number of correct detections that it fails to make for each variable in the data set of interest.

10. In many situations, a data record that has at least one variable value flagged as "suspicious" will have all its data values flagged in the same way. This is equivalent to defining a *case-level* editing process. EUREDIT will also apply evaluation measures to correct/incorrect case level detections.

11. The above measures are essentially "averages", and so will vary across subgroups of the data set of interest. An important part of the evaluation of an editing procedure in EUREDIT will therefore be to show how these measures vary across identifiable subgroups of this data set. For example, in a business survey application, the performance of an editing procedure may well vary across different industry groups.

## II.3 Influential Error Reduction

12. In this case the aim in editing is not so much to find as many errors as possible, but to find the errors that matter (i.e. the influential errors) and then to correct them. From this point of view the size of the error in the measured data (measured value - true value) is the important characteristic, and the aim of the editing process is to detect measured data values that have a high probability of being "far" from their associated true values.

13. In order to evaluate the error reduction brought about by editing, EUREDIT will assume that all values flagged as suspicious by the editing process are checked, and their actual true values determined. Suppose the variable Y is scalar. Then the editing procedure leads to a set of post-edit values defined by $\hat{Y}_i = E_i Y_i + (1 - E_i) Y_i^*$ where $Y_i$ is the measured value for this variable for the ith case and $Y_i^*$ is the corresponding "true" value. The key performance criterion in this situation is the "distance" between the distribution of the true values $Y_i^*$ and the distribution of the post-edited values $\hat{Y}_i$. The aim is to have an editing procedure where these two distributions are as close as possible, or equivalently where the difference between the two distributions is as close to zero as possible.

14. When Y is scalar, the errors in the post-edited data are $D_i = \hat{Y}_i - Y_i^* = E_i (Y_i - Y_i^*)$. For intrinsically positive variables, the evaluation measures that will be used in EUREDIT will be the average of both the $D_i$ and the square of the $D_i$. For strictly positive variables these averages will be expressed as a proportion of the average of the true values (suitably sample weighted if required). In addition, other, more "distributional" measures related to the spread of the $D_i$ will be computed, for example the ratio of the range of the $D_i$ to the interquartile distance of the corresponding true values.

15. With a categorical variable one cannot define an error by simple differencing. Instead EUREDIT will tabulate the joint distribution of the post-edit and true values, and a "good" editing procedure is then one such that the weighted frequency of "mismatches" in this joint distribution is small. When the variable of interest is ordinal, this weighted frequency measures will be modified by the "distance" between the categories that contribute to a mismatch.

16. A slightly different approach to evaluating whether an editing procedure has found the "errors that matter" is to check whether any remaining errors in the post-edited survey data do not lead to estimates that are significantly different from what would be obtained if editing was "perfect". EUREDIT will check this by comparing estimates based on the post-edited data with those based on the true data, e.g. via calculation of appropriately standardised differences. In the case of linear estimates for scalar variables this is equivalent to calculating a t-statistic for testing the hypothesis that the weighted mean of the $D_i$ is not significantly different from zero.

17. When the variable of interest is categorical, with A categories, the $D_i$ values above will be computed as

$$D_i = \sum_a \sum_{b \neq a} I(\hat{Y}_i = a) I(Y_i^* = b) .$$

Here $I(Y_i = a)$ is the indicator function for when case i takes category a. If Y is ordinal rather than nominal then

$$D_i = \sum_a \sum_{b \neq a} d(a,b) I(\hat{Y}_i = a) I(Y_i^* = b)$$

where d(a,b) is a measure of the distance between category a and category b. Evaluation then proceeds as outlined above.

### II.4 Evaluating the Outlier Detection Performance of an Editing Procedure

18. Statistical outlier detection can be considered a form of editing. As with "standard" editing, the aim is to identify data values that are inconsistent with what is expected, or what the majority of the data values indicate should be the case. However, in this case there are no true values that can be ascertained. Instead, the aim is to remove these values from the data being analysed, in the hope that the outputs from this analysis will then be closer to the truth than an analysis that includes these values (i.e. with the detected outliers included).

19. In order to evaluate how well an editing procedure detects outliers, the moments and distribution of the outlier-free data values will be compared with the corresponding moments and distribution of the true values. Similarly, the distribution of the "outlier-free" values will be compared with that of the true values over a range of values that covers the distribution of the true values, e.g. the deciles of the distribution of the true values.

## III. WHAT IS IMPUTATION?

20. Imputation is the process by which values in a data set that are missing or suspicious (e.g. edit failures) are replaced by known acceptable values. EUREDIT will not distinguish between imputation due to missingness or imputation as a method for correcting for edit failure, since imputation is carried out for any variable for which true values are missing. Reasons for imputation vary, but typically it is because the data processing system has been designed to work with a complete dataset, i.e. one where all values are acceptable (satisfy edits) and there are no "holes".

21. Methods of imputation for missing data vary considerably depending on the type of data set, its extent and the characteristics of the missingness in the data. However, there are two broad classes of missingness for which different imputation methods are typically applied. These are *unit missingness*, where all the data for a case are missing, and *item missingness*, where part of the data for a case are missing. The extent of item missingness may well (and often does) vary between different records.

22. An important characteristic of missingness is *identifiability*. Missingness is identifiable if we know which records in the dataset are missing, even though we do not know the values contained in these records. Missingness due to edit failure is always identifiable. Missingness brought about through underenumeration (as in a population census) or undercoverage (as in a sample survey) is typically not identifiable. The importance of identifiability is that it allows one at least in theory to cross-classify the missing records according to their true and imputed values, and hence evaluate the efficacy of the imputation process. EUREDIT will only be concerned with imputation of identifiable missingness.

### III.1 Performance Requirements for Imputation

23. Ideally, an imputation procedure should be capable of effectively reproducing the key outputs from a "complete data" statistical analysis of the data set of interest. However, this is usually impossible, so

alternative measures of performance are of interest. The basis for these measures is set out in the following list of desirable properties for an imputation procedure. The list itself is ranked from properties that are hardest to achieve to those that are easiest. <u>This does NOT mean that the ordering reflects desirability</u>. In fact, in most uses of imputation within national statistical agencies the aim is to produce aggregated estimates from a data set and criteria (i) and (ii) below will be irrelevant. On the other hand, if the data set is to be publicly released or used for development of prediction models, then (i) and (ii) become rather more relevant.

(i) <u>Predictive Accuracy</u>: The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are "close" as possible to the true values.

(ii) <u>Ranking Accuracy</u>: The imputation procedure should maximise preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.

(iii) <u>Distributional Accuracy</u>: The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.

(iv) <u>Estimation Accuracy</u>: The imputation procedure should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).

(v) <u>Imputation Plausibility</u>: The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

24. It should be noted that not all the above properties are meant to apply to every variable that is imputed. In particular, property (ii) requires that the variable be at least ordinal, while property (iv) is only distinguishable from property (iii) when the variable being imputed is scalar. Consequently the imputation evaluation measures used in EUREDIT will depend on the scale of measurement of the variable being imputed.

25. An additional point to note about property (4) above is that it represents a compromise. Ideally, this property should correspond to "preservation of analysis", in the sense that the results of any statistical analysis of the imputed data should lead to the same conclusions as the same analysis of the complete data. However, since it is impossible to a priori identify all possible analyses that could be carried out on a data set containing imputed data, this criterion has been modified to focus on preservation of estimated moments of the variables making up the data set of interest.

26. Furthermore in all cases performance relative to property (5) above ("plausibility") can be checked by treating the imputed values as measured values and assessing how well they perform relative to the statistical editing criteria described earlier in this paper.

27. Finally, unless specifically stated to the contrary below, all imputation evaluation measures used in EUREDIT will be defined with respect to the set of n imputed values within a data set, rather than the set of all values making up this set.

### III.2 Imputation Performance Measures for a Nominal Categorical Variable

28. The extent to which an imputation procedure preserves the marginal distribution of a categorical variable with c+1 categories can be assessed by calculating the value of a Wald-type statistic that compares the imputed and true distributions of the variable across these categories. This statistic is the extension (Stuart, 1955) of McNemar’s statistic (without a continuity correction) for marginal homogeneity in a 2 × 2 table. It is given by

$$W = (\mathbf{R} - \mathbf{S})^t \left[\mathrm{diag}(\mathbf{R} + \mathbf{S}) - \mathbf{T} - \mathbf{T}^t\right]^{-1} (\mathbf{R} - \mathbf{S}).$$

Here **R** is the c-vector of imputed counts for the first c categories of the variable, **S** is the c-vector of actual counts for these categories and **T** is the square matrix of order c corresponding to the crossclassification of actual vs. imputed counts for these categories. Assuming some form of stochastic imputation is used, the large sample distribution of W is chi-square with c degrees of freedom, and so a statistical test of whether the imputation method preserves the distribution of the categorical variable of interest can be carried out.

29. Note that adding any number of “correct” imputations to the set of imputed values being tested does not alter the value of W. That is, it is only the extent of the "incorrect" imputations in the data set that determines whether the hypothesis of preservation of marginal distributions is supported or rejected.

30. The extension of W to the case where more than one categorical variable is being imputed is straightforward. One just defines Y as the single categorical variable corresponding to all possible outcomes from the joint distribution of these categorical variables and then computes W as above.

31. It is also important in EUREDIT to assess how well an imputation process preserves true values for a categorical variable Y with c+1 categories. An obvious measure of how closely the imputed values “track” the true values for this variable is given by the proportion of off-diagonal entries for the square table $\mathbf{T}^+$ of order c+1 obtained by cross-classifying these imputed and actual values. This is

$$D = 1 - n^{-1}\sum_{i=1}^{n} I(\hat{Y}_i = Y_i^*)$$

where $\hat{Y}_i$ denotes the imputed version of Y and $Y_i^*$ is its true value.

32. Provided the hypothesis that the imputation method preserves the marginal distribution of Y cannot be rejected, the variance of D can be estimated by

$$\hat{V}(D) = n^{-1} - n^{-2}\mathbf{1}^t\{\mathrm{diag}(\mathbf{R}+\mathbf{S}) - \mathbf{T} - \mathrm{diag}(\mathbf{T})\}\mathbf{1} = n^{-1}(1 - D)$$

where **1** denotes a c-vector of ones. If the imputation method preserves individual values, D should be identically zero. To allow for the fact that the imputation method may "almost" preserve true values, one can test whether the expected value of D is significantly greater than a small positive constant $\varepsilon$. That is, one is willing to allow up to a maximum expected proportion $\varepsilon$ of incorrect imputations and still declare that the imputation method preserves true values. Consequently, if

$$D > \varepsilon + 2\sqrt{\hat{V}(D)}$$

one can say that the imputation method has an expected incorrect imputation rate that is significantly larger than $\varepsilon$ and hence does not preserve true values. The choice of $\varepsilon$ will depend on the application. In EUREDIT this constant is set to

$$\varepsilon^* = \max\left(0, D - 2\sqrt{\hat{V}(D)}\right).$$

The smaller this value, the better the imputation process is at preserving true values. An imputation method that generates a value of zero for $\varepsilon^*$ for a particular data set will be said to have preserved true values in that data set.

### III.3 Imputation Performance Measures for an Ordinal Categorical Variable

33. When Y is ordinal, preservation of order becomes an issue. To illustrate, consider the following 4 imputed by actual cross-classifications for an ordinal variable Y taking values 1, 2 and 3. In all cases the value of W is zero, so the issue is one of preserving values, not distributions. The D statistic value for each table is also shown. Using a subscript to denote a particular table it can be seen that $D_a < D_b < D_c < D_d$ so the imputation method underlying table (a) appears “best”.

33. However, one could question whether this actually means method (a) IS better than methods (b), (c) and (d). Thus method (a) twice imputes a value of 1 when the actual value is 3, and similarly twice imputes a value of 3 when the actual value is 1, a total of 4 “major” errors. In comparison, method (b) only makes 2 corresponding major errors, but also makes an additional 4 “minor” errors. The total error count (6) for (b) is clearly larger than that of (a), but its “major error count” (2) is smaller. The corresponding count for (c) is smaller still (0). It may well be that method (c) is in fact the best of all the four methods!

| (a) | $Y^* = 1$ | $Y^* = 2$ | $Y^* = 3$ | |
|---|---|---|---|---|
| $\hat{Y} = 1$ | 3 | 0 | 2 | **5** |
| $\hat{Y} = 2$ | 0 | 5 | 0 | **5** |
| $\hat{Y} = 3$ | 2 | 0 | 3 | **5** |
| | **5** | **5** | **5** | D = 4/15 |

| (b) | $Y^* = 1$ | $Y^* = 2$ | $Y^* = 3$ | |
|---|---|---|---|---|
| $\hat{Y} = 1$ | 3 | 1 | 1 | **5** |
| $\hat{Y} = 2$ | 1 | 3 | 1 | **5** |
| $\hat{Y} = 3$ | 1 | 1 | 3 | **5** |
| | **5** | **5** | **5** | D = 6/15 |

| (c) | $Y^* = 1$ | $Y^* = 2$ | $Y^* = 3$ | |
|---|---|---|---|---|
| $\hat{Y} = 1$ | 3 | 2 | 0 | **5** |
| $\hat{Y} = 2$ | 2 | 1 | 2 | **5** |
| $\hat{Y} = 3$ | 0 | 2 | 3 | **5** |
| | **5** | **5** | **5** | D = 8/15 |

| (d) | $Y^* = 1$ | $Y^* = 2$ | $Y^* = 3$ | |
|---|---|---|---|---|
| $\hat{Y} = 1$ | 0 | 0 | 5 | **5** |
| $\hat{Y} = 2$ | 0 | 5 | 0 | **5** |
| $\hat{Y} = 3$ | 5 | 0 | 0 | **5** |
| | **5** | **5** | **5** | D = 10/15 |

35. A way of allowing not only the absolute number of imputation errors, but also their “size” to influence assessment, is to compute a generalised version of D, where the "distance" between imputed and true values is taken into account. That is, we compute

$$D = n^{-1} \sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*)$$

where d(a, b) is the "distance" from category a to category b. Thus, if we put d(a, b) equal to the “block metric” distance function, then $d(\hat{Y}_i, Y_i^*)$ = 1 if $\hat{Y}$ = a and $Y^*$ = a-1 or a+1 and $d(\hat{Y}_i, Y_i^*)$ = 2 if $\hat{Y}$ = a and $Y^*$ = a-2 or a+2. With this definition we see that $D_a = D_b = D_c = 8/15$ and $D_d = 20/15$. That is, there is in fact nothing to choose between (a), (b) and (c). On the other hand, suppose that $d(\hat{Y}_i, Y_i^*)$ = 1 if $\hat{Y}$ = a and $Y^*$ = a-1 or j+1 and $d(\hat{Y}_i, Y_i^*)$ = 4 if $\hat{Y}$ = a and $Y^*$ = a-2 or a+2. That is, major errors are four times as bad as minor errors (a squared error rule). Then $D_a = 16/15$, $D_b = 12/15$, $D_c = 8/15$ and $D_d = 40/15$. Here we see that method (c) is the best of the four.

36. Assuming the categories are numbered from 1 to A, the following compromise definition will be used by EUREDIT:

$$d(a, b) = \frac{1}{2}\left[\frac{|a-b|}{A-1} + I(a \neq b)\right].$$

### III.4 Imputation Performance Measures for a Scalar Variable

37. The statistics W and D can be easily extended to evaluating imputation for a continuous scalar variable by first categorising that variable. If the variable is integer-valued the categorisation is obvious, though “rare” tail values may need to be grouped. For truly continuous variables (e.g. income) a more appropriate categorisation would be on the basis of the actual distribution of the variable in the population (e.g. decile groups). Here again, “rare” groups may need special attention. However, EUREDIT will avoid the arbitrariness of categorising a continuous variable by using imputation performance measures that are directly applicable to scalar variables.

38. To start, consider preservation of true values. If this property holds, then $\hat{Y}$ should be close to $Y^*$ for cases where imputation has been carried out. One way this "closeness" can be assessed therefore is by calculating the weighted Pearson moment correlation between $\hat{Y}$ and $Y^*$ for imputed cases. For data that are reasonably "normal" looking this should give a good measure of imputation performance. For data that are highly skewed however, the non-robustness of the correlation coefficient means that it is preferable to focus on estimates of the regression of $Y^*$ on $\hat{Y}$, particular those that are robust to outliers and influential values.

39. The regression approach evaluates the performance of the imputation procedure by fitting a linear model of the form $Y^* = \beta\hat{Y} + \varepsilon$ to the imputed data values using a (preferably sample weighted) robust estimation method. Let b denote the fitted value of $\beta$ that results. Evaluation then proceeds by testing whether $\beta = 1$. If this test does not indicate a significant difference, then a measure of the regression mean square error

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i (Y_i^* - b\hat{Y}_i)^2$$

can be computed. A good imputation method will have a non-significant p-value for the test of $\beta = 1$ as well as a low value of $\hat{\sigma}^2$.

40. Underlying the above regression-based approach to evaluation is the idea of measuring the performance of an imputation method by the distance $d(\hat{\mathbf{Y}}, \mathbf{Y}^*)$ between the n-vector $\hat{\mathbf{Y}}$ of imputed values and the corresponding n-vector $\mathbf{Y}^*$ of true values. An important class of such measures is defined by the weighted $L^p$ distance between $\hat{\mathbf{Y}}$ and $\mathbf{Y}^*$. EUREDIT will calculate these measures for p =1 and p = 2.

41. Preservation of ordering for a scalar variable will be evaluated by replacing $Y^*$ and $\hat{Y}$ above by their <u>ranks</u> in the full data set (not just in the set of imputed cases).

42. The distance between the weighted empirical distribution functions $F_{Y^*n}(t)$ and $F_{\hat{Y}n}(t)$ defined by the true and imputed values respectively provides a measure of how well the imputation procedure "preserves distributions". Such a distance is

$$d_\alpha(F_{Y^*n}, F_{\hat{Y}n}) = \frac{1}{t_{2n} - t_0}\sum_{j=1}^{2n}(t_j - t_{j-1})\left|F_{Y^*n}(t_j) - F_{\hat{Y}n}(t_j)\right|^\alpha$$

where the $\{t_j\}$ values are the 2n jointly ordered true and imputed values of Y with $t_0$ equal to the largest integer smaller than or equal to $t_1$.and $\alpha$ is a "suitable" positive constant. Larger values of $\alpha$ attach more importance to larger differences between $F_{Y^*n}(t)$ and $F_{\hat{Y}n}(t)$. EUREDIT will use $\alpha = 1$ and $\alpha = 2$.

43. Finally, one can consider preservation of aggregates when imputing values of a scalar variable. The most important case here is preservation of the raw moments of the empirical distribution of the true values. For k = 1, 2, ..., EUREDIT will measure how well these are preserved by computing

$$m_k = \left|\sum_{i=1}^{n} w_i (Y_i^{*k} - \hat{Y}_i^k) \Big/ \sum_{i=1}^{n} w_i\right| = \left|m(Y^{*k}) - m(\hat{Y}^k)\right|.$$

Preservation of derived moments, particularly moments around the mean, is also of interest in EUREDIT. In this case the data values (true and imputed) will be replaced by the corresponding differences. For example, preservation of moments around the mean will be assessed by calculating $m_k$ above with $Y_i^*$ replaced by $Y_i^* - m(\mathbf{Y}^*)$ and $\hat{Y}_i$ replaced by $\hat{Y}_i - m(\hat{\mathbf{Y}})$. Similarly, preservation of joint second order moments for two variables $Y_1$ and $Y_2$ will be measured by calculating $m_k$, but now replacing $Y_i^*$ by $\left(Y_{1i}^* - m(\mathbf{Y}_1^*)\right)\left(Y_{2i}^* - m(\mathbf{Y}_2^*)\right)$, and $\hat{Y}_i$ by $\left(\hat{Y}_{1i} - m(\hat{\mathbf{Y}}_1)\right)\left(\hat{Y}_{2i} - m(\hat{\mathbf{Y}}_2)\right)$.

### III.5 Evaluating Outlier Robust Imputation

44. The outlier robustness of an imputation procedure can be assessed by the "robustness" of the analyses based on the imputed values, compared to the analyses based on the true data (which can contain outliers). This is a rather different type of performance criterion from that investigated so far, in that the aim here is not to get "close" to the unknown true values but to enable analyses that are more "efficient" than would be the case if they were based on the true data values.

45. For the EUREDIT project the emphasis will be on assessing efficiency in terms of mean squared error for estimating the corresponding population mean using a weighted mean based on the imputed data values. Note that this measure uses all N data values in the data set rather than just the n imputed values, and is given by

$$\mathrm{MSE} = \left(\sum_{i=1}^{N} w_i\right)^{-1} \left(\sum_{i=1}^{N} w_i I(\hat{Y}_i = Y_i^*)\right)^{-1} \sum_{i=1}^{N} w_i^2 (\hat{Y}_i - m_N(\hat{Y}))^2 + \left[m_N(\hat{Y}) - m_N(Y^*)\right]^2 .$$

Here $m_N(Y)$ refers to the weighted mean of the variable Y defined over all N values in the data set of interest. Observe that the variance term in (30) includes a penalty for excessive imputation.

## IV. EVALUATING IMPUTATION PERFORMANCE FOR MIXTURE TYPE VARIABLES

46. Mixture type variables occur regularly in most official statistics collections. These are scalar variables that can take exact values with non-zero probability and are continuously distributed otherwise. An example is a non-negative variable that takes the value zero with positive probability $\pi$, and is distributed over the positive real line with probability $1 - \pi$. The most straightforward way to evaluate imputation performance for such variables is to evaluate this performance separately for the "mixing" variable and for the actual values at each level of the mixing variable.

## V. EVALUATING IMPUTATION PERFORMANCE IN PANEL AND TIME SERIES DATA

47. A panel data structure exists when there are repeated observations made on the same set of cases. Typically these are at regularly spaced intervals, but they do not have to be. The vector of repeated observations on a variable Y in this type of data set can therefore be considered as a realisation of a multivariate random variable. EUREDIT will therefore calculate multivariate versions of the measures described earlier in order to evaluate imputation performance for this type of data.

48. For time series data the situation is a little different. Here i = 1, ..., n indexes the different time series of interest, with each series corresponding to a multivariate observation indexed by time. For such data most methods of analysis are based on the estimated autocorrelation structure of the different series. Hence an important evaluation measure where imputed values are present is preservation of these estimated autocorrelations. Let $r_{ik}^*$ denote the true value of the estimated autocorrelation at lag k for the series defined by variable $Y_i$, with $\hat{r}_{ik}$ the corresponding estimated lag k autocorrelation based on the imputed data. The EUREDIT measure of the relative discrepancy between the estimated lag k autocorrelations for the true and imputed versions of these series is then

$$R_k = \left| \sum_{i=1}^{n} \left(r_{ik}^* - \hat{r}_{ik}\right) / \sum_{i=1}^{n} r_{ik}^* \right| .$$

## VI. COMPARING TWO (OR MORE) IMPUTATION METHODS

49. A key analysis in the EUREDIT project will be the comparison of a number of imputation methods. Simple tabular and graphical analyses will often be sufficient in this regard. For example, Madsen and Larson (2000) compare MLP neural networks with logistic regression at different levels of "error probabilities" and display their results in tabular and graphical format showing how various performance measures for error detection for these methods vary with these probability levels.

50. A more sophisticated statistical analysis would involve the independent application of the different methods to distinct subsets of the data set (e.g. industry or regional groups) and then computing the performance measure of interest for each of the different methods within each of these groups. A repeated measures ANCOVA analysis of these values (maybe after suitable transformation) with imputation method as a factor will then be carried out.

51. Alternatively, pairwise "global" comparisons of imputation methods will be carried out in EUREDIT using a measure of the form

$$r_{1L\alpha}(\hat{Y},\tilde{Y}) = \frac{d_{L\alpha}(\hat{\mathbf{Y}},\mathbf{Y}^*)}{d_{L\alpha}(\tilde{\mathbf{Y}},\mathbf{Y}^*)} = \left\{ \sum_{i=1}^{n} w_i \left|\hat{Y}_i - Y_i^*\right|^{\alpha} \Big/ \sum_{i=1}^{n} w_i \left|\tilde{Y}_i - Y_i^*\right|^{\alpha} \right\}^{1/\alpha} .$$

Here $\hat{Y}$ and $\tilde{Y}$ are two imputed versions of Y, and $\alpha = 1$ or 2.

## VII. OPERATIONAL EFFICIENCY

52. Editing and imputation methods have to be operationally efficient in order for them to be attractive to most "large scale" users. This means that an important aspect of assessment for an editing and imputation method is the ease with which it can be implemented, maintained and applied to large scale data sets. The following criteria will be used in EUREDIT to determine the operational efficiency of an editing and imputation (E&I) method:

(a) What resources are needed to implement the E&I method in the production process?
(b) What resources are needed to maintain the E&I method?
(c) What is the required expertise needed to apply the E&I method in practice?
(d) What are the hardware and software requirements?
(e) Are there any data limitations (e.g. size/complexity of data set to be imputed)?
(f) What feedback does the E&I method produce? Can this feedback be used to "tune" the process in order to improve its efficiency?
(g) What resources are required to modify the operation of the E&I method? Is it possible to quickly change its operating characteristics and rerun it?
(h) A key aspect of maintaining an E&I system is its *transparency*. Is the underlying methodology intuitive? Are the algorithms and code accessible and well documented?

## VIII. PLAUSIBILITY

53. The plausibility of the imputed values is a *binding requirement* for an imputation procedure, in the sense that an imputation procedure is unacceptable if it generates implausible values. This is particularly important for applications within NSIs. Within EUREDIT plausibility will be assessed by the imputed data passing all "fatal" edits, where these are defined. The *degree* of plausibility will be measured by calculating the edit performance measures described earlier, treating the imputed data values as the pre-edit "raw" values.

## IX. QUALITY MEASUREMENT

54. In the experimental situations that will be explored in EUREDIT, it is possible to use simulation methods to assess the quality of the E&I method, by varying the experimental conditions and observing the change in E&I performance. However, in real life applications the true values for the missing/incorrect data are unknown, and so this approach is not feasible. In particular, information on the quality of the editing and imputation outcomes in such cases can only be based on the data available for imputation.

55. In this context editing quality can be assessed by treating the imputed values as the true values and computing the different edit performance measures described earlier. Of course, the quality of these quality measures is rather suspect if the imputed values are themselves unreliable. Consequently an important property of an imputation method should be that it produces measures of the quality of its imputations. One important measure (assuming that the imputation method preserves distributions) is the so-called imputation variance. This is the *additional* variability, over and above the "complete data variability", associated with inference based on the imputed data. It is caused by the extra uncertainty associated with randomness in the imputation method. This additional variability can be measured by repeating the imputation process and applying multiple imputation theory. Repeatability of the imputation process is therefore an important quality measure.

## References

Madsen, B. and Bjoern Steen Larsen, B. S. (2000). The uses of neural networks in data editing. Invited paper, International Conference on Establishment Surveys (ICES II), June 2000, Buffalo, N.Y.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42, pg. 412.