WP. 30
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

METHODS AND SOFTWARE FOR EDITING AND IMPUTATION: RECENT ADVANCEMENTS AT ISTAT

Invited Paper

Submitted by ISTAT, Italy[1]

Abstract: In the paper the most recent methodological and technological advancements at ISTAT in the area of editing and imputation are described. A recently developed model-based method for localizing systematic unity measure errors and some parametric and non-parametric approaches for imputing either categorical or continuous data are illustrated. Available generalized software is also briefly presented. Recent developments concerning the DIESIS software for editing and imputing demographic census data are also discussed.

I. INTRODUCTION

1. National Statistical Institutes (NSIs) are in charge of producing public use data that are generally employed by Governments for their social/economic policies. For this reason, NSIs have to provide high quality statistical information. Particularly in large-scale surveys, it is common practice to adopt more or less complex procedures to guarantee the completeness and the coherence of the published data. Data verification aiming at identifying unacceptable information in surveyed data can be performed at different stages of the overall survey process. The set of actions performed on data at the post-data capturing stage, indicated as *editing and imputation*, aim at eliminating non-sampling errors that are residual from the previous survey phases. In order to improve the quality of editing and imputation, NSIs are focussing their work by: anticipating data verification at the early stages of the survey process (at the data capturing and/or data entry stages), exploiting as much as possible available auxiliary information on respondents/target phenomena, using more efficient strategies in the post-data collection stage. In this paper we focus on the latter aspect.

2. Due to the costs of editing and imputation, some aspects are to be carefully considered when planning and implementing an editing and imputation procedure: the need to balance between resources spent and data accuracy, the need to optimize the treatment of each specific class of data/error problems by identifying appropriate methods/approaches, the need to provide survey practitioners with generalized solutions in order to facilitate the adoption of such methods. As for the former aspect, it is recognized that, in order to provide high quality statistical information, it is in general not necessary to identify all the errors affecting data (Granquist *et al.*, 1997): efforts are to be concentrated on errors having the

[1] Prepared by Marco Di Zio (dizio@ISTAT.it), Ugo Guarnera (guarnera@ISTAT.it), Orietta Luzi (luzi@ISTAT.it), Antonia Manzari (manzari@ISTAT.it)

highest impact on figures to be published. As for the second aspect, in spite of the strong specificity of surveys, some common, "optimal" solutions can be identified for some types of error problems. For example, probabilistic approaches can be considered as optimal for identifying non-influential errors generated by completely random mechanisms. Allowing survey practitioners to adopt such solutions through the use of generalized algorithms gives guarantees on the quality of data, supports the standardization of data treatment, and reduces costs through to the development of ad hoc software.

3. For an acceptable trade-off between process quality, process costs and process standardization, investments in both the areas of methodological research and software development are needed, and an organizational effort has also to be made to standardize as much as possible the data processing strategies adopted in the survey production processes. During the last decade the Italian Institute of Statistics (ISTAT) has spent a great amount of resources on improving the quality of its statistical products and increasing the degree of standardization of statistical survey processes. For a long time, the situation in ISTAT has not been uniform, because either different methods were used to face the same statistical problems, or a same approach was adopted by developing (hence duplicating) algorithms and software. In order to reduce this waste of resources, and at the same time to disseminate methodologies and generalized tools giving guarantees on data quality, a considerable effort has been made at ISTAT in the following areas:

(i) editing;
(ii) imputation;
(iii) evaluation and documentation of editing and imputation.

4. In the area of error localization, research and development activities first concentrated on the problem of dealing with large amounts of statistical categorical data affected by random errors. In 1992 ISTAT carried out the first application on Census Population data of the Fellegi-Holt methodology (Fellegi *et al.*, 1976) as implemented in the SCIA software (Riccini *et al.*, 1995). This tool has been progressively adopted by all the main ISTAT social surveys dealing with data on households and individuals, including the Labour Force Survey, the Household Budget Survey, the Multipurpose Survey, the European Panel Survey (ECHP). More recently, in the context of the methodological and operational activities for the 2001 Italian Population Census, ISTAT started developing the DIESIS software implementing a *data-driven* approach to deal with random errors in mixed demographical household data (Bruni *et al.*, 2001), and inspired by the Canadian NIM/Canceis (Bankier, 2000)

5. Concerning editing of continuous data, in 1998 ISTAT acquired from Statistics Canada the Generalized Editing and Imputation Software - *GEIS/Banff* (Kovar *et al.*, 1988) in order to provide survey managers of the economic area with a Fellegi-Holt-based methodology to deal with stochastic non-influential errors in continuous data. This tool has been successfully adopted in the Labour Cost Survey and in the Survey on Structure and Production of Agricultural Firms. Nevertheless, it is well known that one of the most critical problems in business surveys is represented by influential errors, originated by either stochastic or systematic mechanisms. The relevance of these errors depends on their potential biasing effects on the survey target parameters. In this area, significance/selective editing (Latouche *et al.*, 1992; Lawrence *et al.*, 2000) and graphical editing (Tukey, 1977) are highly effective non-parametric approaches for identifying critical observations, and for balancing between data accuracy and data processing costs.

6. In the area of imputation, ISTAT is confronted with two main problems. Firstly, all the generalized tools mentioned above essentially use donor-based approaches, probably due to the operational simplicity of hot-deck methods, but in spite of the theoretical availability of techniques sometimes more complex but in some circumstances more effective from a statistical point of view. Furthermore, concerning the adopted imputation approaches, ISTAT surveys are not homogeneous at all, particularly in the business area. This fact only partially depends on the nature of investigated phenomena and on the available resources, but it is mainly due to the absence of guidelines and minimal standards in the area of imputation. The result is that a variety of naive, tailor-made approaches are adopted in surveys. In addition, often the same method is implemented in different environments and/or programming languages, with duplication of software and waste of resources. For all these reasons,

research and implementation efforts were recently concentrated on providing survey managers with imputation methodologies, suitable for a broad range of statistical purposes and implemented in generalized tools, which imply different assumptions on data and that exploit in different ways the statistical relationships among data.

7. In the area of evaluating and documenting editing and imputation, a significant effort has been made at ISTAT during the last decade to standardize these tasks and stimulate subject matter experts in analysing their data after data processing (Di Zio *et al.*, 2002). Concerning the problem of evaluating the quality of editing and imputation in terms of capability of correctly identifying errors and properly recovering *true* data, a solution based on simulation has been proposed. The prototypal system ESSE (*Editing System Standard Evaluation*) (Della Rocca *et al.*, 2000) providing algorithms for the artificial generation of errors and missing data and for the calculation of some quality indicators has also been developed. ESSE represents a starting point for future improvements. For documentation, the generalized tool *IDEA* (*Indices for Data Editing Assessment*) has been developed for the analysis of the impact and the statistical effects of editing and imputation on survey data (Della Rocca *et al.*, 2004). A number of different indicators are computed by comparing raw and clean data, including the standard quality indicators required by the *Information System for Survey Documentation–SIDI* (Brancato *et al.*, 1998).

8. In this paper, the most recent methodological and technological advancements at ISTAT in the field of editing and imputation are illustrated and discussed. In particular, the paper focuses on the following issues:

1) *in the area of editing*:
   - Finite Mixture Models for the detection of unity measure errors;
2) *in the area of imputation*:
   - Bayesian Networks for the imputation of categorical data;
   - EM-based methods for the imputation of continuous data;
3) *in the area of editing and imputation:*
   - recent advancements of the DIESIS software.

9. For editing continuous data, a model-based approach to the detection of systematic unity measure errors is proposed. In particular, we show how suitably constrained Finite Mixture Models can be used in order to identify these errors based on a classification approach grouping units according to their error pattern. An important feature of this approach is that suitable diagnostics are obtained that can be used for the identification of critical data in a selective/significance editing perspective. The Finite Mixture model approach is described in section II.A.

10. In the second part of the paper we deal with recent advancements in the area of imputation. In this context, two different issues are discussed: using Bayesian Networks for categorical data imputation, and imputing continuous data using EM and other parametric and non-parametric methods. Concerning Bayesian Networks, this technique is shown to be particularly useful to preserve the multivariate relationships among variables. Software allowing the application of the method has been developed at ISTAT. Bayesian Networks and related issues are dealt with in section III.A. As for continuous data, the generalized tool *QUIS* (QUick Imputation System) has been recently implemented at ISTAT. Imputation based on the use of the EM algorithm and multivariate predictive mean matching can be performed on incomplete quantitative survey data. Nearest-neighbour techniques with different distance measures are also available, as well as a multiple imputation module. QUIS and the available methods are described in section III.B.

11. In section IV a summary description of recent developments of the Diesis software for error localization and imputation of demographic Census data is performed. Improvements mainly relate to the use of a clustering approach for improving the selection of donors in the editing and imputation process.

II. RECENT DEVELOPMENTS IN THE EDITING AREA

12. Statistical editing consists of identifying implausible data with respect to some plausibility rules or statistical models defined by subject matter experts. Identifying the most appropriate rules/models for efficiently identifying the different types of errors is not a simple task: using not well-designed edits or miss-specified models can dramatically compromise the effectiveness of the editing process, hence the validity of subsequent statistical analyses. In addition, particularly in business surveys, editing strategies have to be designed based on the evidence that errors have not the same importance.

13. The well-established Fellegi-Holt methodology, as well as the *data-driven* approach and other automatic algorithms (see for example De Waal *et al.*, 2003) represent valid answers to deal with huge amounts of irrelevant random errors for both continuous and categorical data. In this area, they currently represent standard approaches in many NSIs, although some problems still remain from both a theoretical and an operational point of view (e.g., how to identify the "optimal" set of edits, how to efficiently use auxiliary information, which are the statistical consequences implied by their use, which are their statistical effects on data characteristics, and so on).

14. On the other hand, dealing with (random or systematic) errors that are influential on survey figures is a more specific survey-dependent problem. In this area, several approaches have been proposed in literature, which can be roughly classified in non-parametric and parametric, depending on the assumptions made on data models and/or on data relations. Typically non-parametric approaches are macro-editing, graphical editing, selective and significance editing, classification/regression trees (see for example Chambers *et al*., 2004). Model-based approaches are generally adopted for identifying errors originating anomalous situations with respect to the assumed data models, like non-representative outliers (among recent papers, see for example Hulliger *et al.*, 2004).

15. In the next section a model-based approach developed at ISTAT for identifying systematic unity measure errors in continuous data is described.

A. Finite mixture models for detecting of systematic unity measure errors

16. In the area of editing and imputation a common error classification leads to define two broad error typologies: systematic error and random error. In the family of systematic errors, one that has a high impact on final estimates and that frequently affects data in statistical surveys measuring quantitative characteristics (e.g. business surveys) is the *unity measure error times a constant factor* (e.g. 100 or 1,000). This error is due to the erroneous choice, by some respondents, of the unity measure in reporting the amount of some variables. The unity measure error (*UME* in the following) is generally treated through ad hoc procedures using essentially graphical representations of marginal or bivariate distributions, and *ratio edits*. These approaches have some drawbacks: firstly, ratio edits are effective for UME when one of the two variables is error free; furthermore, with ratio edits no more than pairwise analyses between variables can be performed, disregarding more complex interactions; finally, adopting pairwise analyses implies that variables are to be treated in a pre-defined hierarchy, thus increasing the complexity of the error localization phase.

17. These drawbacks can be overcome by adopting a probabilistic formalisation of the problem though finite mixture models (McLachlan *et al.*, 1988*;* McLachlan *et al.*, 2000). This modelling can provide a principled statistical approach, allowing an estimate of the conditional probability that an observation be affected by *UME*. This approach has the advantages, with respect to the traditional ones, to formally state the problem in a multivariate context, to be easily implemented in generalised software, and to naturally provide useful diagnostics for prioritising doubtful units possibly containing influential errors.

18. The *UME* generally acts multiplying variables by a constant factor. Hence data in error appear in log-scale as translated by a vector of constants, that depends on which items are in error (*error pattern*), while the covariance structure is the same for each error pattern. Since in business surveys variables are

frequently considered log-normal, in logarithmic scale the Gaussian setting can be adopted. Our goal is to assign each single observation to a specific error pattern, i.e. to localise items in error: by interpreting each single error pattern as a "cluster", the error localisation problem is transformed in a cluster analysis problem, and we can exploit experiences from the model-based cluster analysis theory (Fraley *et al.*, 2002).

19. Let $\boldsymbol{X}_i = (X_{i1},\ldots,X_{iq})$ be the q-dimensional vectors of values of q survey target variables, with p.d.f. $f(x_1,\ldots,x_q;\boldsymbol{q})$, such that $E(X_1,\ldots,X_q)=(\mu_1,\ldots,\mu_q)=\boldsymbol{\mu}$, and $Var(X_1,\ldots,X_q) = \Sigma$. Based on the assumption that systematic errors affect the random vector $\boldsymbol{X}$ only by transforming its expected value $\boldsymbol{\mu}$ into $\psi_g(\boldsymbol{\mu})$, where $\psi_g(\cdot) : R^q \to R^q$, for $g=1,\ldots,h$, are a set of known functions, the functions ψ_g characterise univocally h distinct clusters (*error patterns*), differing each other only on the location parameter. For instance, if the systematic error possibly affects all the variables X_s for $s=1,..,q$, in the same manner by transforming their expected values μ_s according to $\mu_s \to \mu_s + C$, where C is a known constant, the number of clusters will be $h = 2^q$, i.e. the number of different combinations of error occurrence on the q variables (including the case of no error). In this case, each function ψ_g and each corresponding cluster, is associated with one of the 2^q possible sub-sets of variables affected by the error. Observed data can therefore be thought of as realizations of a random vector $\boldsymbol{Y}$ whose distribution, conditional on $\boldsymbol{X}$, depends on the systematic error mechanism.

20. For the error localisation purpose we follow a model-based approach based on finite mixture models, where each mixture component G_g, $g=1,..h$, represents a single error pattern. Formally, we assume that $\boldsymbol{Y}_i = (Y_{i1},\ldots, Y_{iq})$, for $i=1,\ldots,n$, are iid with respect to $\sum_{t=1}^{h} \pi_t f_t(\cdot;\boldsymbol{\theta}_t)$, where $\sum_t \pi_t = 1$ and $\pi_t \geq 0$. The mixing parameter π_t represents the probability that an observation belongs to the t-th mixture component.

21. We assume that $f_g(\boldsymbol{y}; \boldsymbol{q}_t) \sim MN(\boldsymbol{\mu}_g, \Sigma)$ and that each function $\psi_g(\cdot)$ acts on the mean vector $\boldsymbol{\mu}$ as a translation: $\psi_g(\boldsymbol{\mu}) = \boldsymbol{\mu} + \boldsymbol{C}_g$, where $\boldsymbol{C}_g$ is the known translation vector for the mean of the g-th cluster. The algorithm used to compute the likelihood estimates is a modified version of the EM algorithm as suggested in McLachlan *et al.* (1988), adapted in order to meet our particular situation: in fact, while in the non-constrained case (McLachlan *et al.*, 1988) a different mean vector has to be estimated for each mixture component, in our constrained situation only one mean vector has to be estimated.

22. Once the maximum likelihood estimates of the parameters $(\boldsymbol{q}_g, \pi_g)$ have been obtained, each observation $\boldsymbol{y}_i$ is classified in one of the h groups based on its posterior probability:

$$\tau_g(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}},\hat{\pi}) = pr(i\text{-th observation } \in G_g \mid \boldsymbol{y}_i; \boldsymbol{q}, \pi) = \hat{\pi}_g f_g(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_g) \Big/ \sum_{t=1}^{h} \hat{\pi}_t f_t(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_t) \quad g=1,\ldots,h$$

The *i-th* observation is assigned to the cluster G_t if $\tau_t(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}},\hat{\pi}) > \tau_g(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}},\hat{\pi})$, $g=1,\ldots,h;\ g \neq t$. This allocation rule is the optimal solution for the classification problem, in the sense that it minimises the overall error rate (Anderson, 1984, Chapter 6). Once data have been classified into the clusters, for each observation we can assess whether it is in error or not, and which variables are in error.

23. The main problems arising from the normality assumption are assessing the validity of the assumption itself, and analysing the robustness of the model when data depart from normality. The problem of assessing normality in mixture models is well described in McLachlan *et al.* (1988). It is based on the quantities $\hat{\alpha}_{gi}$ (called *atipicality* index) directly provided by the model (for more details, see McLachlan *et al.*, 1988, Chapter 2). Under the normality assumption, $\hat{\alpha}_{gi}$ for $i=1,\ldots,\hat{m}_g$ is approximately uniformly distributed on (0,1). Following McLachlan *et al.* (1988), we used the the Anderson-Darling statistic for assessing the uniform distribution of $\hat{\alpha}_{gi}$ (Di Zio *et al.*, 2005). Concerning the model robustness for departures from the normality assumption, some experiments on this aspect have been performed by Di Zio *et al.* (2005). The model performance (in terms of number of correctly

classified units) has been evaluated with respect to populations characterized by different degrees of skewness (e.g. a bivariate t distribution and a bivariate skew-t distribution). Results showed that, even if the number of correctly classified units decreases with the departure from normality, it seems acceptable also in the most critical case. Nevertheless, it is worth mentionning that for extreme departures from normality the method is expected to fail (e.g. when true data contain different clusters, for instance differences in men and women income might cause a bimodal distribution for the income itself). In some cases the problem could be overcome by stratifying data with respect to some explicative variables, e.g. sex in the previous example. An alternative approach to this specific problem could be based on modelling each cluster in turn as a Gaussian mixture, thus obtaining a "mixture of mixture models" (Di Zio *et al.*, 2004).

24. An important characteristic of mixture models relates to the possibility of using diagnostics directly provided by the models for prioritising the different types of doubtful or critical units possibly containing influential errors:

(i) possibly misclassified observations, i.e. units classified in a cluster, but having a non-negligible probability of belonging to another cluster (units that are in the regions where the mixture components overlap each other). In order to measure the degree of belief in the classification of an observation $\boldsymbol{y}_i$ we can consider the corresponding posterior probability: observations for which this probability is not very close to one, have a non-negligible probability to belong to another cluster;

(ii) outliers with respect to the model, i.e. observations that are far from all the clusters. Outliers can be identified using the above mentioned atypicality index $\hat{a}_{gi}$: the lower is $\hat{a}_{gi}$ the higher is the probability of $\boldsymbol{y}_{gi}$ of being atypical, thus all observations with $\hat{a}_{gi} < \alpha$, where α is a specified threshold, are classified as atypical. Results of experiments on mixture models performance for different levels of α for different types of populations can be found in Di Zio *et al.* (2005).

25. A clerical review is needed on critical observations in order to increase data accuracy. Hence, it is important to optimize the selection of critical observations in order to save time and costs. To this aim, classification probability and atypicality index can be used, according to a selective/significance editing approach (Latouche *et al.*, 1992; Lawrence *et al.*, 2000) to build up score functions to prioritise doubtful units. Di Zio *et al.* (2005) propose different global scores, and carry out a test on real survey data. For each unit, this function takes into account the potential UME error (by using the posterior probabilities and the expected magnitude of the error), as well as the impact of this error on the target parameter estimate.

26. Relating to the model complexity in terms of number of parameters to be estimated, it is worthwhile noting that even if the number of clusters and then the number of mixing parameters p_t can have an exponential growth with respect to the number of variables, the number of parameters related to the mean vector and covariance matrix increases much slower than in a usual mixture problem, due to the location constraints characterising our model. Actually, the higher the number of variables analysed, the bigger is this difference (for instance in the case of three variables and 8 clusters we need to estimate 16 parameters instead of 37). However, attention has to be paid to this aspect, e.g. by limiting the analysis to target variables, and/or group variables based on their statistical association.

III. RECENT DEVELOPMENTS IN THE IMPUTATION AREA

27. The imputation is a phase in the production of statistics that is discussed in-depth in literature, see for instance Little and Rubin (2003).

28. One issue that is worth dealing with is to find techniques that preserve as much as possible the joint distribution of the analysed variables. Furthermore, due to the nature of surveys treated by National Statistical Institutes, an important aspect of the imputation methods is also the capability of dealing with complex situations, both in terms of number of variables and in terms of number of observations.

29. Hot deck techniques are a valid answer to these problems. However, in spite of their versatility in handling complex situations, they still present aspects to be further improved, in particular the preservation of the joint distributions. Vice-versa parametric models can be useful for the preservation of joint distributions but can be hardly used in practice when dealing with complex surveys.

A. Using Bayesian Networks for the imputation of categorical data

30. Methods studied in the artificial intelligence and machine-learning contexts can be a promising way to deal with practical problems like complexity, and statistical issues like the preservation of joint distributions. In particular, Bayesian Networks (BNs) are shown through empirical studies to have a good behaviour in the imputation context. The first idea of using Bayesian Networks is by Thibaudeau and Winkler (2002). Then Di Zio *et al.* (2004a) have implemented the algorithm with some minor variations and tested it on data artificially contaminated with missing data.

31. The interest in this technique is that it is possible to express the joint probability distribution with a dramatic decrease of parameters to be estimated, with the consequence of a simplification of the problem. Actually, BNs may estimate the necessary relationships between variables that are really informative for predicting values, thus discarding all the interactions that are not useful for the prediction. In hot deck terminology, it is like a system that finds for each variable to be imputed the corresponding most important matching variables (*covariates*) to consider. However this choice is made in a way that all the conditional distributions used to impute each variable (conditional on the covariates) are compatible and the joint distribution is an estimate of the unknown distribution generating the process.

32. For instance, let us suppose that variables C1, C2, .., C7 are observed. Let us suppose that the BN estimated on the data is like that in Figure 1. This net says that if we want to impute variable C3, we must consider the conditional distribution P(C3|C2, C1) (estimated from the BN technique). C1 and C2 are named "parents" of C3. If one of them (or both) are missing, they previously must be imputed by their distribution conditional to their parents, and since they are at the root of the net (they have no parents), they will be imputed according to their marginal distribution, for instance if C2 is missing the missing values will be imputed generating a value from P(C2), (estimated from the BN).

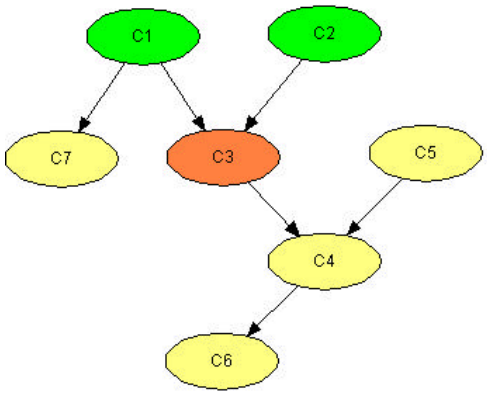

Figure 1: An example of Bayesian Network

33. After these first experiments, a change was made to the algorithm. The modification relates to the choice of conditional distribution to consider for imputing each variable. It uses not only information coming from the parents, but also from the children. Actually, more variables than the parents are directly linked to the variables to impute. More formally, the covariates chosen are the variables that make the variable to impute independent of the rest. This is called Markov Blanket and in the previous example the variable C3 should be imputed according to the distribution P(C3|C1,C2,C4,C5) (see Figure 2).

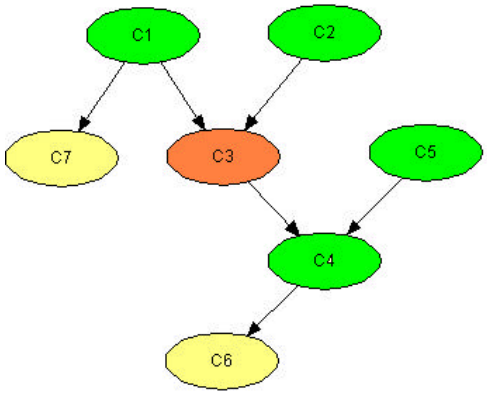

Figure 2: Markov blanket of C3

34. The available software gives only the probability distribution for each node conditional on the parents. In order to impute by Markov Blanket a trick is used, suggested by considering that the Markov Blanket of the variables without children is composed of only their parents. Hence, one BN for each variable to be imputed is estimated in a way that it is at the end of the network (node without children). Impute each single variable with respect to the corresponding graph, iterating the algorithm discussed at the beginning of the section. This method performed better than the first, at the cost of introduction of a more demanding algorithm.

35. In Di Zio *et al.* (2004a) the first algorithm has been compared with the stratified random hot deck. In general the results show that joint distributions are better preserved by the BN method. In particular the two methods have the same behaviour in the case that the hot deck is stratified according to variables explaining exactly the missing mechanism (in a MAR context). Of course this *a priori* knowledge is rarely available, while the BN algorithm acts equally satisfactorily without the need of such knowledge. In Di Zio *et al.* (2004b), the second algorithm has been compared to the first one. The empirical results show a general better behaviour of the second algorithm with respect to the first in terms of preservation of joint distributions, due mainly to the improvements obtained in the imputation of the variables at the root of the network. This is explained by the fact that in the first algorithm these variables are imputed according to a random draw from the marginal distribution, while in the second algorithm they are imputed with a random draw from their conditional distributions with respect to the Markov Blanket.

36. The software used for this algorithm are: 1) Hugin (www.hugin.com; see also Madsen *et al.*, 2003) for the part devoted to the network estimation and the conditional probability distributions, 2) an ISTAT software that implements the two algorithms so far described, and that also presents the possibility, in a simulative context, of evaluating through simple indicators the performance of the imputation. These indicators compare an initial *true* data set, the data set with missing items, and a data set with imputed values.

37. In the studies so far introduced, BNs are only used for the imputation of categorical variables. This is a limit that should deserve further study.

B. Methods and tools for the imputation of continuous data

38. Due to the lack of generalized software for applying some of the most common methods for item-non-response treatment, the software *QUIS* (QUick Imputation System) has been recently implemented in ISTAT in order to provide people involved in Editing & Imputation activities with an easy tool for quantitative data imputation within a unified environment (Guarnera, 2004).

39. In *QUIS* three imputation methods for numerical variables are available:

(i) Regression Imputation via EM algorithm;
(ii) Nearest Neighbour Donor Imputation (NND);
(iii) Multivariate Predictive Mean Matching (PMM).

Moreover, based on the SAS-Macro implemented by Allison and available on the website, (http://www.ssc.upenn.edu/~allison/), the Shafer's Multiple Imputation procedure for Normal Multivariate data (Schafer 1997) is also available.

40. In *QUIS* the EM algorithm is used for estimating the parameters of the multivariate normal model that is utilized to impute missing values. This is easily done through the use of "sweep operators", a family of operators that allows transforming the parameters of the joint distribution for normal data into those of the conditional distributions (and vice versa). The conditional distributions corresponding to the different error patterns are used in the E-step of the EM algorithm for computing the expectations of the sufficient statistics of the model parameters conditional on the observed data and the current estimates. The maximization step is then performed by equating these quantities to the expected values of the sufficient statistics expressed in terms of the updated values of the model parameters (Shafer, 1997).

41. Once the Maximum Likelihood Estimates (MLE) have been computed, imputation can be performed in two different ways: missing data can be imputed through expectations of missing values conditional on observed ones (predictive means), or, alternatively, a normal random residual can be added to the predictive means. The best option depends on the specific research objectives: in particular, if some linear quantities have to be estimated such as means or totals, the first method should be preferred for its optimality properties, while if one is interested in preserving the distributional characteristics of the data, imputing with a random residual seems to be more appropriate. The stopping criterion of the EM algorithm consists, as usually, of checking the estimates change between two consecutive runs and stopping the algorithm when this change is below some prefixed threshold. However, the user is requested to provide the maximum number of runs allowed (default is 200).

42. The version of the NND method implemented in QUIS consists of simultaneous imputation of all the missing values in any incomplete record, with values taken from the same (nearest) donor. Three distance functions are available for the selection of the nearest neighbour: *euclidean*, *manhattan* and *min-max*. The matching variables can be chosen even among variables affected by missing values: whenever a matching variable is not observed for one recipient record, it is simply ignored in the computation of the distance. In any case, the donors are taken from the set of records that are complete with respect to all of the variables of interest. If no matching variables are selected or if none of the selected matching variables is observed for a given incomplete record, a donor is chosen at random from the donors set. Unlike the regression imputation method, based on EM algorithm, NND does not require any normality assumption.

43. Also based on the normal model, but somewhat less sensitive to the model assumptions, is the multivariate predictive mean matching: PMM, like the regression method, uses EM in order to estimate the parameters of a multivariate normal model. However, in this case a "live" value rather than a normal random variable is used for adding a random residual to the predictive mean. More in detail, once the parameters of the underlying model have been estimated, for each incomplete record, the predictive means of missing variables, conditional on the observed ones, are computed. The same predictive means are estimated for all the complete units (donors) so that the values resulting from the regression can be used as "matching variables" in order to find a nearest neighbour donor. As for the EM algorithm, the parameters needed for regression are derived from the model parameters through the sweep operators.

44. Since more than one missing value can occur in incomplete records, the predictive means are in general multidimensional-vectors. Thus the problem arises of defining a metric to be used for the donor selection. For each pattern of missing values a natural metric is given by the Mahalanobis distance defined through the residual variance-covariance matrix corresponding to that pattern. In case of slight

departure from normality, this procedure is hoped to be more robust than standard regression based on the normal model, in particular it is expected to perform better in case of heteroskedasticity.

45. The Multiple Imputation (MI) procedure, based on the algorithm described in Schafer (1997), provides as output a unique data set where one further variable is added indicating the imputed data-set number (imputation-number). The number of imputations is selected by the user in the MI window as well as some technical parameters (max number of runs of EM for initialising the parameters, number of iterations needed by the data-augmentation algorithm to reach stationary, etc.).

46. The availability in a unified environment of the above-described methods allows using different approaches for different typologies of data. Moreover, it is particularly useful whenever different techniques are suitable for different groups of variables within the same application.

47. *QUIS* has been developed in the environment SAS SYSTEM V8 and consists of some routines (SAS Macro) mainly written in SAS-IML code. The user interface, made of interactive windows, has been implemented using the *Screen Control Language*, a programming language specifically designed for building interactive applications in the SAS-environment.

IV. RECENT DEVELOPMENTS IN DIESIS SOFTWARE FOR DEMOGRAPHIC DATA

48. The Data Imputation and Edit System - Italian Software (DIESIS) (Bruni *et al*., 2001) is the new system developed and used by ISTAT for the treatment of the demographic variables from the 2001 Population Census. The editing and imputation methodology implemented in DIESIS is based on the identification of a subset of *potential donors* for each failed edit record (erroneous record). The potential donors should be the passed edit records as similar as possible to the erroneous record. The similarity between each erroneous record e and each passed edit record d is calculated by a function $f(e, d) \in [0,1]$ defined as the weighted sum of the distances (for quantitative variables) or similarities (for categorical variables) for each demographic variable over all individuals within the household. The set of potential donors contains only the nearest k passed edit records (where k is a pre-specified value) provided that their distance is below a pre-specified threshold.

49. For each erroneous record e, the identification of the potential donors should be made by searching within the set of all possible passed edit records D. However, when D is very large, as in the case of a Census, the computation of the distance between each e and all $d \hat{I} D$ could require unacceptable computational time. To face this problem a sub-optimal solution has been adopted consisting in arresting the search before examining the entire set D, according to some stopping criteria (Bianchi *et al.*, 2004). The stopping criteria are based on distance thresholds and guarantee that the distances between the selected potential donors and the erroneous record are below a pre-defined threshold. It must be stressed that this solution might reduce the imputation quality, because the criteria could not guarantee the selection of the subset of potential donors having minimum distance from the erroneous record.

50. In order to overcome this drawback, a new approach for reducing the number of passed edit records to be examined has been proposed (Bianchi *et al.*, 2005). This approach consists in preventively dividing the set of passed edit records D into a collection of smaller subsets $\{D_1, \ldots, D_n\}$ in such a way that $D_1 \cup \ldots \cup D_n = D$, and that all elements of the same subset D_j have similar characteristics. Such subdivision is obtained by solving an unsupervised clustering problem (Hastie *et al*., 2001; Jain *et al*., 1999) because no a priori information about such subdivision is known. The search for the potential donors is then conducted, for each erroneous record e, by examining only the passed edit records within the cluster(s) more similar to e. The algorithm for the assignment of the passed edit records to the clusters, called *algorithm of spherical neighborhoods*, uses a distance function based on the similarity between the joint distributions of the demographic variables over all individuals within the household (Reale *et al*., 2004). This distance will soon be available in DIESIS also to identify the potential donors for each erroneous record e for all the searching methods.

51. The clustering algorithm progressively selects some passed edit records, considering around each one a sphere of pre-specified radius r using the just-mentioned distance function. In particular the algorithm consists of two main phases. In the first phase, the clusters are formed in an iterative way. A passed edit record $d_s \in D$ is randomly selected and a cluster D_s is formed for d_s by taking all other passed edit records $d \in D$ having distance $f(d_s, d) = r$ (the *spherical neighborhood*). Then another passed edit record, external to the just formed cluster D_s, is selected and a new cluster is formed. The process goes on until the set D has been completely examined. Record d_s is the centroid of the cluster D_s. Each passed edit record which is not a centroid may belong to more than one cluster, since the spherical neighborhoods may overlap. In the second phase each cluster that contains a high number of passed edit records is subdivided in smaller clusters by gradually reducing the radius. In this phase each passed edit record may belong to only one cluster. Note that final clusters can have different radius depending on their density (in terms of number of units belonging to the cluster). Moreover, the algorithm is computationally inexpensive, and may therefore be used for very large data sets.

52. The centroids of each cluster allows associating one or more clusters to each erroneous record by means of the distance function: only the clusters having the centroid nearest to the erroneous record are selected. Then, for each erroneous record, the potential donors are searched for only inside the selected clusters. The described procedure has been implemented in C++.

53. Two types of initial tests have been performed on large data sets of four-person and six-person household records. Persons within households have been ordered by decreasing age.

- The first test compares the editing and imputation quality obtained by the exhaustive search for the potential donors within all D with the one obtained by the search guided by the clustering approach. This test aims at verifying whether the use of clustering would decrease the editing and imputation quality with respect to the "ideal" search. The evaluation has been performed in an error-simulation context, by comparing some accuracy indicators (Manzari *et al.*, 2002) computed for each variable and each searching methods. The results show no sensible difference for the two donor selection methods for both household sizes. This suggests that clustering does not reduce editing and imputation quality, although it drastically reduces the number of computations of $f(e,d)$, and hence the computational time.

- The second test compares the selection of the potential donors obtained by the sub-optimal search implemented in DIESIS with the one obtained by the search guided by the clustering approach. In both cases only a pre-defined number of computations of $f(e,d)$, has been allowed. This test aims at studying whether the use of clustering would increase the donor selection quality (in terms of minimum distance) with respect to the sub-optimal search. The evaluation has been performed by comparing the percentage of the actual minimum distance donors which have been correctly selected as potential donors by each searching methods. Note that the actual minimum distance donors have been preliminarily identified, for each erroneous record, by using the exhaustive search for the potential donors within all D. The results show relevant differences. In particular, for four-person households, a percentage of about 100% of the actual minimum distance donors has been obtained by using clustering, this percentage decreases to about 70% when no clustering is used. On the other hand, for six-person households, a percentage of about 95% has been obtained by using clustering, whereas the percentage decreases below 5% when no clustering is used.

54. The results obtained, in particular the large gap observed in the second test between the percentages of six-person households, suggest that the search guided by clustering could be especially useful for treating households having uncommon structure. For this type of households, in fact, few donors are generally available and often they are not very similar to the failed edit household. The search guided by clustering could reduce the number of changes required by the data driven approach and hence

help in preserving the collected information. Therefore, further tests and improvements of the algorithm are worthwhile to be performed.

55. At present, ISTAT is developing a generalized version of the DIESIS system, to extend the use of the implemented methodologies to other surveys. To this aim, it is first necessary to simplify the definition of rules and the acquisition of data, and to develop a graphical interface. Moreover, more efficient algorithms have been realized to reduce the memory allocation and the processing time, in particular in the search of potential donors and in the editing phase (Bianchi *et al.*, 2003). In addition, the application will be made multi-platform, in the sense that it will be possible to use the software in different environments (Windows, Linux and Unix).

References

Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis.* Second Edition. New York:Wiley.

Bankier M. (2000). Canadian Census Minimum change Donor imputation methodology. *Proceedings of the Workshop on Data Editing, UN/ECE Work Session on Statistical Data Editing,* UK, Cardiff.

Bianchi G., Saporito G. (2003). Enumerate Permutations Problem. *Technical report (in Italian),* ISTAT.

Bianchi G., Pezone A., Reale A., Saporito G. (2004). Metodi e Procedure per il Controllo e la Correzione delle Variabili Demografiche Familiari del Censimento della Popolazione 2001, *Technical report* (*in italian),* ISTAT.

Bianchi G., Bruni R., Nucara R., Reale A. (2005). Data Clustering for Improving the Selection of Donor for Data Imputation, *to be presented at the fifth Conference on Classification and Data Analysis (CLADAG), June 6-8, 2005,* Parma, Italy.

Brancato G., D'Angiolini G., Signore M. (1998). Building up The Quality Profile of ISTAT Surveys. *Proceedings of the Joint IASS-INEGI-IAOS Conference "Statistics for Economic and Social Development"*, Aguascaliente, Messico, 1-4 September, CD ROM.

Bruni R., Reale A., Torelli R. (2001). Optimization Techniques for Edit Validation and Data Imputation, *Proceedings of the Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" XVIIIth International Symposium on Methodological Issues.*

Chambers R., Hentgesand A., Xinqiang Z. (2004). Robust automatic methods for outlier and error detection, *Journal of the Royal Statistical Society*, Vol. 167, Issue 2.

Della Rocca G., Di Zio M., Manzari A., Luzi O. (2000). E.S.S.E. Editing System Standard Evaluation, *Proceedings of the SEUGI 18,* Dublin, June 20-23.

Della Rocca G., Di Zio M., Luzi O. (2004). "Assessing editing and imputation effects on statistical survey data", *Proceedings of the International Conference on Quality in Official Statistics* (First version), Mainz , May 24-26.

De Waal T., Quere R. (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics*, Vol. 19, N. 4.

Di Zio M., Manzari A., Luzi O. (2001). Evaluating Editing and Imputation Processes: the Italian Experience. *UN/ECE Work Session on Statistical Data Editing,* Helsinky, Finland, May 27-29.

Di Zio M., Guarnera U., Rocci R. (2004). A Mixture of Mixture Models to detect Unity Measure Error. *Proceedings in Computational Statistics, Antoch Jaromir (Ed.)*, 919-927, Physica Verlag, Prague, August 23-28.

Di Zio M., Scanu M., Coppola L., Luzi O., Ponti P. (2004a). Bayesian networks for imputation. *Journal of the Royal Statistical Society, Series A*, 167(2), 309 – 322.

Di Zio M., Sacco G., Scanu M., Vicard P. (2004b). Multivariate Techniques for Imputation based on Bayesian Networks, in J. Antoch (Editor): *Proceedings Compstat 2004, 16th Symposium of IASC*, Prague, 23-27 August 2004, Physica Verlag – Springer Verlag, 928-934.

Di Zio M., Guarnera U., Luzi O. (2005). Editing Systematic Unity Measure Errors Through Mixture Modelling. *Survey Methodology*, Vol. 31, No. 1 (*to appear*)

Encyclopedia of Statistical Sciences (1999). Update Vol. 3, 621-629. J. Wiley & Sons, Inc.

Fellegi I. P. and Holt D. (1976) A systematic approach to edit e imputation, *Journal of the American Statistical Association*, vol.71, pp. 17-35.

Granquist L. (1995). Improving the Traditional Editing Process. In *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott, New York,: Wiley, 385-401.

Granquist L. and Kovar J. (1997). Editing of Survey Data: How Much is Enough? In *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, 415-435.

Groves R.M., Dillman D.A., Eltinge J.L., Little J. (2002). *Survey Nonresponse.* Wiley, New York.

Guarnera U. (2004). Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi. Il software QUIS. *Contributi ISTAT* n. 1/2004 (in Italian).

Hastie T., Tibshirani R., Friedman J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, US.

Hulliger B., Béguin C. (2004). Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society: Series A*, Vol. 167: Issue 2, 275-294.

Kovar, J.G., MacMillan, J., Whitridge, P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Statistics Canada, *Methodology Branch Working Paper, BSMD-88-007E/F*, Ottawa.

Jain A.K., Murty M.N., Flynn P.J. (1999). Data Clustering: A Review, *ACM Computing Surveys*, 31:3.

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

Lawrence, D., McKenzie, R. (2000). The General Application of Significance Editing. *Journal of Official Statistics*, Vol. 16, 243-253.

Little J., Rubin D. (2002). *Statistical Analysis with Missing data.* Second Edition.Wiley, New York.

McLachlan G. J., Basford K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

Madsen A.L., Lang M., Kjaerulff U.B., Jensen F. (2003). The Hugin Tool for learning Bayesian Networks. In T.D. Nielsen and N.L. Zhang (Eds) *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings of the 7th European Conference,* ECSQARU 2003, Aalborg, Denmark, July 2003.

Manzari A., Reale A. (2002). Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology. *Proceedings of the 53rd Session of The International Statistical Institute,* August 22-29, 2001, 634-655. Sydney: International Statistical Institute.

McLachlan G. J., Peel D. (2000). *Finite Mixture Models.* New York: Wiley.

Schafer J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Reale A., Bianchi G., Manzari A., Pezone A., Saporito G. (2004). Proposta di un metodo per il calcolo della distanza mista tra donatore ed errato, *Technical report* (*in italian),* ISTAT.

Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995). The Methodology of Editing and Imputation by Qualitative Variables Implemented in SCIA. *ISTAT Technical Report.*

Thibaudeau Y., Winkler W.E. (2002). Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. *Technical report RRS2002/9,* U.S. Bureau of the Census.

Tukey J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, London.
