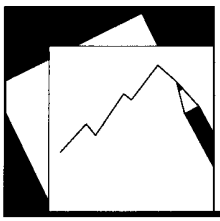
Working Paper

INTERNATIONAL MONETARY FUND

WP/11/151



IMF Working Paper

Avoid Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series

Michaela Denk, Michael Weber

INTERNATIONAL MONETARY FUND

 WP/11/151

IMF Working Paper

STA

Avoid Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series

Prepared by Michaela Denk, Michael Weber

Authorized for distribution by Ann McPhail

June 2011

Abstract

International organizations collect data from national authorities to create multivariate cross-sectional time series for their analyses. As data from countries with not yet well-established statistical systems may be incomplete, the bridging of data gaps is a crucial challenge. This paper investigates data structures and missing data patterns in the cross-sectional time series framework, reviews missing value imputation techniques used for micro data in official statistics, and discusses their applicability to cross-sectional time series. It presents statistical methods and quality indicators that enable the (comparative) evaluation of imputation processes and completed datasets.

JEL Classification Numbers: C13, C52, C59, C80

Keywords: Missing or incomplete data, imputation quality, statistical matching

Acknowledgements: Special thanks go to our colleagues at the IMF Statistics Department, especially Mike Seiferling, as well as at the Social Protection & Labor Unit of the World Bank's Human Development Network for valuable feedback and discussions.

Authors' E-Mail Addresses: mdenk@imf.org, mweber1@worldbank.org

Contents Page

Figures

I. Introduction

Well-founded decisions and policy-making in international organizations is subject to the availability of high quality cross-country time series data as a basis for economic analysis. Without such data providing substantial evidence, there is ample room for speculation or merely theoretical solutions to political, societal, and economic questions. International organizations collect and aggregate relevant data from national authorities such as statistical offices, national banks, and governmental departments. However, data on particular topics, for example employment or government finance, are not collected regularly or in sufficient frequency in many developing countries due to the lack of a well-established statistical system with adequate resources and institutional capacity to set up and maintain costly data collection processes. Other developing countries collect data, but do not publish them in a timely manner because of under-staffed and -equipped data processing units. The recent financial crisis drew attention to the problem of allocating development assistance when relevant data are available only insufficiently or even not at all. Bridging data gaps is a crucial challenge in this regard, emphasizing the importance of (further) supporting developing countries in strengthening their institutional and methodological capabilities, the necessity of additional data collection initiatives, as well as the relevance of and need for sound statistical methods for data imputation in time series.

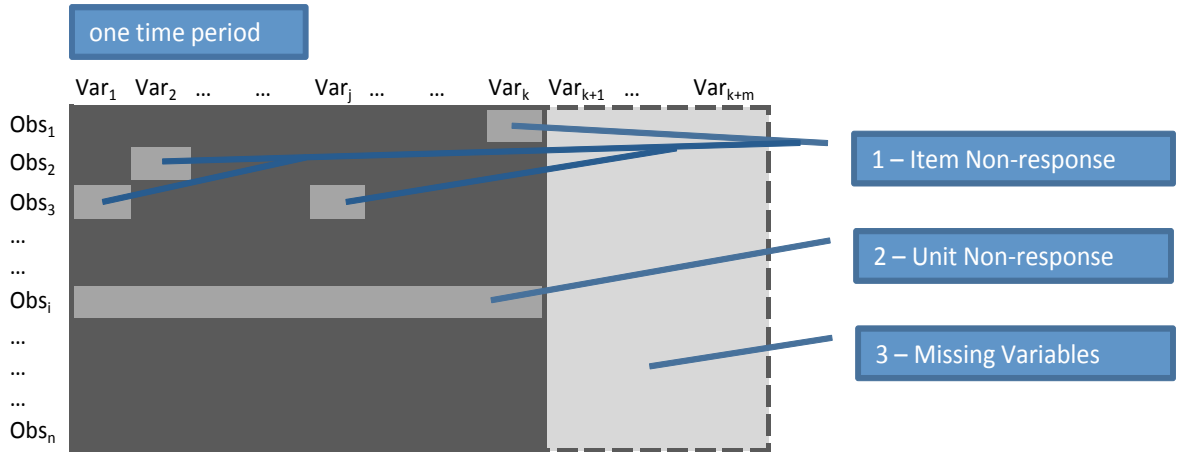
An ongoing project of the Social Protection and Labor Unit of the World Bank's Human Development Network deals with the imputation of labor market indicators in cross-country time series (Margolis, Newhouse, Weber, 2010ab). The aim of the project is to enable the assessment of the labor market situation during the recent financial crisis and in future projects. Missing (or low frequency) labor market indicators are imputed based on typically less fragmentary (and higher frequency) macro-economic indicators and models estimated for data-rich countries. This project gave reason to investigate existing statistical imputation methods and imputation quality measures as applied in official statistics. The present paper provides an overview of the findings of this methodological review with a focus on quality measures and evaluation routines for model comparison. It aims at introducing and promoting imputation techniques to data producers as well as analysts in international financial institutions.

The remainder of the paper is structured as follows. Section 2 investigates data structures and missing data patterns in the cross-sectional time series framework as compared to the traditional survey framework. Section 3 reviews missing data techniques used in official statistics that were originally developed for filling gaps in survey data. Section 4 discusses the relevance and applicability of these techniques in the context of cross-sectional time series. Statistical methods and quality indicators for the evaluation of the imputation process and the completed data as well as the comparison of different techniques are discussed in section 5. Section 6 summarizes the main ideas of the paper.

II. DATA STRUCTURES AND MISSING DATA PATTERNS

Missing data techniques commonly used in official statistics focus on filling gaps in survey data. Survey data are micro data and, thus, consist of multiple variables observed or measured for a sample of observation units from a population at one point in time. The gaps in the data can be classified as item non-response, unit non-response, or variables not included in the survey as illustrated in Figure 1 below. Item non-response refers to the situation of one or multiple variables missing for one or multiple observations. The variables (= items) missing may vary between observations. Item non-response can be dealt with by traditional or multiple imputation and statistical matching (see Section 3). Unit non-response means that all variables are missing for one or multiple observations. That is, no data are available at all for the respective observation units. Unit non-response is often accounted for by weighting algorithms (not discussed here, see e.g. Little, 1982; Holt, Smith, 1979; Cochran, 1977). Variables not included in the survey are missing for all observations. If available from other data sources, these variables can be added by statistical matching (see section 3) or record linkage (not discussed here, see e.g. Fellegi, Sunter, 1969; Winkler, 1995; Denk, 2008).

Figure 1. Missing data patterns for standard micro (=observation by variable) data



Time series data differ from micro survey data in the origination of the data, the data structure, the (interpretation of) missing data patterns, and the applicability of standard (survey) missing data techniques. The observation by variable data structure as used for survey data holds data for one time period. In contrast, time series contain data for multiple time periods for one or multiple aggregate observation units and for one or multiple observed variables (or aggregate statistical indicators). Time series in official statistics are usually macro data. This means that they often do not contain data for individual observation units, but rather for aggregate (or macro) units which are also called sections. In international statistics, these aggregate observation units or sections are countries most frequently. Variables are usually statistical indicators such as unemployment rate, current account balance, or GDP.

Overall, the following four different types of time series data structures exist.

1. single univariate time series: one variable/indicator observed for one observation unit/section over time

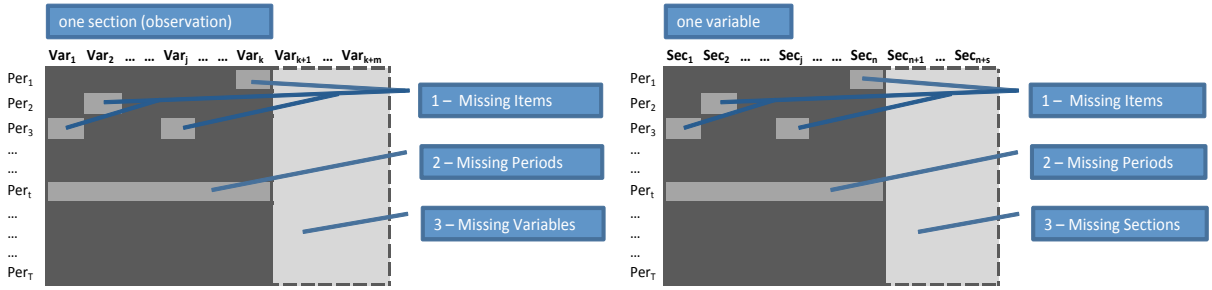
2. single multivariate time series: multiple variables/indicators observed for one observation unit/section over time

3. cross-sectional univariate time series: one variable/indicator observed for multiple sections over time

4. cross-sectional multivariate time series: multiple variables/indicators observed for multiple sections over time

Figure 2 shows the interpretation of different missing data patterns for single multivariate time series and univariate cross-sectional time series. There may be missing items, periods, and, depending on the type of time series data, missing variables or sections.

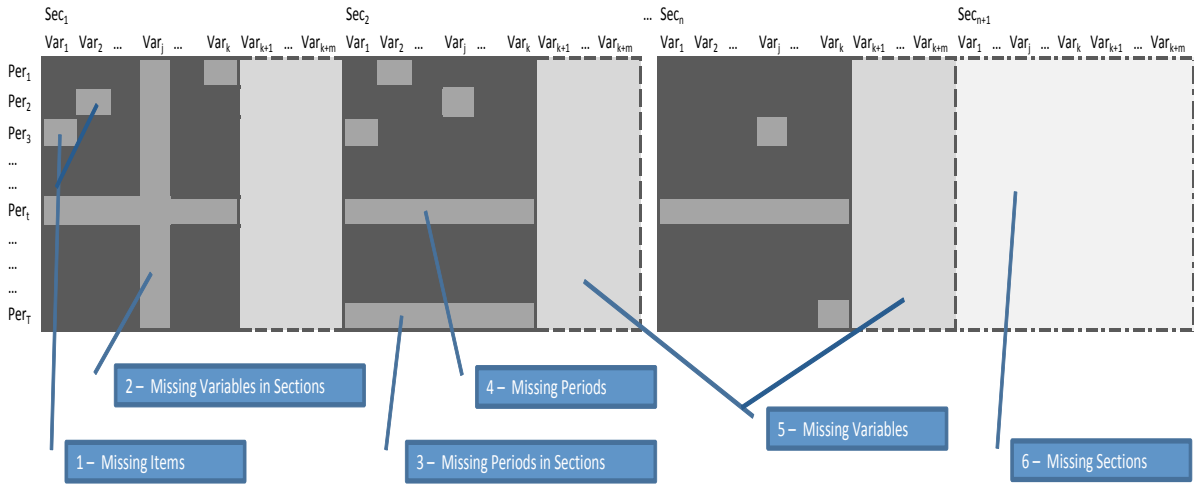
Figure 2. Missing data patterns for multivariate time series and univariate cross-sectional time series



For multivariate cross-sectional time series, the data structure gets more complex as none of the three dimensions (section, variable, time) is held constant. Consequently, all four missing data patterns described as well as their combinations may occur as depicted in Figure 3.

A special case of missing periods emerges in datasets containing data collected in different frequencies. For example, a time series dataset may contain some variables on a quarterly basis, but other variables only on an annual basis, or some sections may comprise annual data while other sections provide the same variables for every other year only. While the former situation results from some statistics being prepared at a higher frequency than other statistics, the latter is frequently observed in cross-country datasets that contain data from countries with well- and less-developed statistical systems.

Figure 3. Missing data patterns for multivariate cross-sectional time series



III. MISSING DATA TECHNIQUES

Imputation is a statistical technique to estimate missing or implausible values in a dataset based on collected values from the dataset or comparable data sources. The focus of this methodology lies on partially missing data due to item non-response. In cases with completely missing variables (e.g. for one or more sections/countries), statistical matching is more appropriate. In common practice, list-wise deletion (also termed "complete case analysis"), a procedure that simply excludes all observation units with missing values from further analysis, or similar approaches are used instead of proper imputation techniques. With these procedures, a large share of information gets lost and biased estimates are a frequent consequence. Researchers have recurrently demonstrated that estimates based on imputed datasets outperform estimates based on reduced datasets that ignore observation units and/or variables with missing values irrespective of the underlying imputation method (e.g. Colledge et al., 1978; Little, Rubin, 2002). Therefore, list-wise deletion cannot be considered a viable alternative to imputation unless data are missing completely at random. From the perspective of a producer of official statistics the removal of incomplete sections, variables, or time series from a dataset destined for dissemination is usually not acceptable. However, some statistical organizations may also refrain from publishing imputed statistics, adhering to a strict interpretation of their role as provider of "facts" that can only originate from reported data and should not be blended with mere estimates. This strategy transfers the challenge of imputation to users of the published data.

An important aspect to be considered when choosing a missing data technique is the underlying missing data mechanism (Rubin, 1987; Schafer, 1997; Little, Rubin, 2002). Random (ignorable) and non-random (systematic, informative) missing values can be distinguished. In case of data missing completely at random (MCAR), the missing data process is ignorable in imputation. MCAR means that the missingness of a variable neither

depends on the true (but missing) value of that variable nor on other (observed or non-observed) characteristics. A weaker assumption that many imputation techniques rely on is data being missing at random (MAR). In that case the missingness of a variable is independent of the true (but missing) value of that variable after controlling for other variables. In other words, the missingness only depends on other variables which can be taken into account in the imputation procedure. Exogenous shocks such as war or natural disasters are examples of factors that affect the missingness of (economic) time series. An imputation method not taking into account this additional information will typically be biased and over- or under-estimate the variable for the missing time period.

If values are missing in a non-random, systematic way (MNAR), the distributions of the variable among complete and missing observations cannot be expected to be the same. This effect is also known as selection bias. For MNAR data the missing data mechanism may not be ignored. It can be included in model-based imputation by simultaneously modeling the variable that "contains" missing values and the probability of that variable being missing. Another approach to dealing with MNAR data is the analysis of the effects of different missing data mechanisms on the imputation result. Multiple imputation (see below) is applied under different missing data scenarios and the results are compared and combined (e.g. Carpenter, Kenward, White, 2007).

In the following, traditional single imputation techniques, statistical matching, and multiple imputation are briefly outlined. A more detailed introduction to imputation and other types of missing data analysis is given in Little and Rubin (2002) or de Waal, Pannekoek, and Scholtus (2011).

A. Traditional approaches

Simple imputation approaches such as deterministic or mean imputation are common in data analytical practice due to their common availability and straightforward applicability in statistical software packages despite their unfavorable statistical characteristics. In official statistics, donor-based imputation is frequently used. The class of model-based approaches is very versatile, as special types of statistical models can be chosen for specific data constellations, though with the drawback that a particular method may not be applicable to other imputation problems. In practice, complex imputation problems are frequently dealt with by hybrid approaches that combine several different methods, e.g. model-based and donor-based imputation. A recent example of the possible implications of the imputation method chosen on the quality of the results is provided by Kaplan and Schulhofer-Wohl (2010).

Deterministic imputation is the simplest imputation approach (apart from deletion). It replaces missing values by values that are specified ad-hoc (Sande 1982). Each missing value may be treated differently in a manual procedure, or a few rules of thumb may be formulated

based on experience and/or by subject-matter experts. This approach does not correspond to any principle of statistical methodology and usually distorts the (marginal as well as joint) distributions of the imputed variables, but is applicable for any type of variable. An evaluation is hardly possible due to its ad-hoc character. Nevertheless, if only a few values are missing, the distortion may be negligible. For some types of data, for instance transaction data that can be viewed from two directions, such as external trade or international investment statistics, replacing missing values by their available mirror values seems reasonable. For example, country A's exports to country B (missing) may be substituted by country B's imports from country A. Still, estimating a model that uses mirror characteristics as explanatory variables to predict missing values would be a methodologically more sound choice from a statistical point of view.

Location-based imputation replaces missing values with a location parameter of the distribution, typically the mean (for metric variables), median (metric or ordinal variables) or mode (categorical variables). The distribution parameter is either based on all observed values for a variable or on all values within a subgroup (= stratum) defined by other variables. It can also be calculated from a comparable dataset. Using overall mean imputation on large parts of a dataset causes serious distortions in the distributions with high peaks at the imputed values and considerably reduced variability of the imputed variables. Imputation of different mean values for different subgroups of the data can reduce the distortion, if the variables defining the groups are correlated with the variables with missing values. This approach is also known as post-stratification (Holt, Smith, 1979). Further drawbacks of this method are (i) that mean or median, and thus the imputed data point, might take unobservable values and (ii) that, with respect to time series, exogenous shocks are not dealt with satisfactorily, as the missing values for a time period in which a shock occurred would be replaced by the average of time periods without a shock. As regards method classification, mean imputation can also be seen as particular kind of deterministic imputation.

Distribution-based imputation uses the entire empirical distribution of a variable for imputation instead of restricting it to one parameter of the distribution. The probabilities for the occurrence of observed values of a variable are estimated by means of the empirical distribution function (non-parametric) or a parametric distribution based on a distribution assumption and the parameters estimated from the observed values. The imputation value is drawn randomly from this probability distribution. As location-based imputation, distribution-based imputation can also be applied to more homogeneous subgroups of the dataset. One advantage as compared to mean imputation is that only observable values are imputed. Most often, only univariate distributions are taken into account, although a multivariate approach may better reflect reality in terms of reasonable value combinations per observation unit/section. More sophisticated distribution-based imputation techniques are iterative procedures that make use of the EM (expectation maximization) principle (Dempster, Laird, Rubin, 1977) (see section C. on multiple imputation below).

Model-based imputation uses correlations between available variables and variables with missing values to estimate a (linear) model which is then applied to predict the missing values. It is applicable to nominal, ordinal, and metric variables; the type of variable has to be accounted for in the choice of the model type. For example, (multinomial) logit or probit models may be used for categorical variables. In case of time series, auto-regression models are common. As the imputation model is predictive and not causal, all available variables that may improve the prediction should be used; especially variables involved in the survey design are of relevance. The danger is more in leaving out useful predictors than in including too many unimportant variables. The latter may lead to a loss in precision, but can be overcome by the usage of stepwise variable selection procedures. Still, the necessity of identifying the "best" model makes the application of model-based imputation comparably complex. Even though modeling tends to produce smoothed data, it better preserves the individual and joint distributions of imputed variables than other methods and usually reduces bias in the estimation of aggregates such as means and totals based on the completed dataset. Model-based imputation also allows taking into account external shocks.

Donor-based imputation takes imputation values from a so-called donor that is a complete observation with similar characteristics as the incomplete observation (the recipient). The similitude between donor and recipient is determined via matching variables to be selected based on their correlation with the variable to be imputed. Donor-based imputation is often used for imputation of categorical variables.

Hot-/cold-deck methods group the complete observations of a dataset into subsets that share the same values in the (usually categorical) matching variables. Hot-deck procedures select donors from the same dataset (the one with missing values); cold-deck procedures use other comparable data sources. To each observation with missing values, one of the donors of the matching subset is assigned. The selection of the donor can be carried out (i) sequentially, (ii) by means of a random process, (iii) based on distances with respect to other common variables, or (iv) based on ranks with respect to a common ordinal variable. Random selection procedures are the most common.

Nearest neighbor methods measure the distance between complete observations and observations with missing values usually based on metric matching variables. Tarsitano and Falcone (2010) show how to deal with mixed-type matching variables. Either the nearest neighbor or one of the k nearest neighbors that is selected randomly is used as a donor. A multi-donor approach can be pursued instead of choosing one particular donor from the set of potential donors. In this case, a set of donors is combined by calculating the imputed value as a (weighted) average or median of the donors' values, forfeiting the advantage of creating observable values at any rate. Various weighting schemes are conceivable, for example proportional to the similarity between donor and recipient, to the frequency of a donor already being used for other recipients, or to the frequency of the pattern of the matching variables of the donor occurring in the dataset.

If the number of matching variables is large, the number of potential donors may be very small. On the other hand, the usage of very few matching variables may result in a poor match. Attempts to impute for all missing variables in a single processing step usually lead to an excessive usage of the same complete observations as donors. In contrast to model-based and mean imputation, donor-based methods generally produce imputed datasets that appear more realistic, since they impute observed values (except for multi-donor approaches) and better reflect the distributional properties. However, donor-based imputation techniques involve a number of subjective decisions that critically affect the quality of the completed dataset, such as the selection of matching variables and the choice of distance measures, and may be criticized for their heuristic nature. Although donor-based methods can deal with recipient variables of any type, model-based approaches are often preferred in case of metric recipient variables, whereas the opposite holds true for categorical recipient variables.

B. Statistical Matching

Statistical matching can be regarded as a particular type of donor-based imputation. It enriches a recipient dataset with variables only available in a donor dataset by combining observations from the two datasets based on the similarity of matching variables that are available in both datasets. The matching process gives rise to completed observations with variables that were completely missing in the recipient set imputed from the donor set. In *constrained matching*, every recipient as well as every donor observation is included in the final dataset with a sample weight identical to its sample weight before matching in order to preserve the distributions of the two datasets. A precondition for constrained matching is the identity of the weighted population totals in both datasets. *Unconstrained matching* does not place such a restriction on the matches. A drawback of constraint matching is that, on average, the distances between matched observations may be larger than in unconstrained matching (e.g. Hollenbeck, Doyle, 1979).

Equivalence class matching subdivides the datasets into comparable subsets (= equivalence classes) of observations by means of agreement or similarity of matching variables or cluster analysis. To each recipient in a subset one or more donors from the same subset are assigned. Donors may be selected based on distance measures or randomly (cf. Okner, 1972). Multiple donors can be combined by some aggregation function, e.g. mean, median, or mode, depending on the type of variable (Van der Putten, 2000). Equivalence class matching corresponds to cold-deck donor-based imputation of completely missing variables.

Regression-based matching matches recipients and donors based on the agreement or similarity of additional variables estimated in both datasets (cf. Kadane, 1978; Moriarity, Scheuren, 2001; Raessler, 2002). Typically, these additional variables are estimated by means of regression models with the common matching variables as regressors and the (dis-) similarity quantified in terms of Mahalanobis distance. Regression-based matching is

comparable to model-based imputation with two datasets, but instead of the estimated value, the value of the nearest potential donor with respect to the estimated variables is used.

Propensity score matching (Rosenbaum, Rubin, 1983) has originally been designed in a different context but can be used for identifying donors in standard imputation or statistical matching situations. To this end, the propensity score is defined as the conditional probability of an observation being contained in the dataset of donors or the dataset of recipients given a set of variables available in both datasets. This propensity score is usually estimated via logistic regression on these common (matching) variables. The matching is carried out by assigning to each recipient the nearest donor in terms of the propensity score. Propensity score matching can be regarded as a special case of regression-based matching.

C. Multiple imputation

Multiple imputation (Rubin, 1987, 1996) is a simulation-based approach to the statistical analysis of incomplete data. The idea of multiple imputation is to extract relevant information from the observed portions of a dataset via a statistical model to impute multiple (usually about five) values for each missing cell and use these values to construct multiple completed datasets. The Bayesian interpretation of this approach is that multiple imputed values are drawn from an estimate of the posterior distribution (instead of using the expected value of this distribution as a single imputed value). These are then analyzed by standard complete data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing data uncertainty. In general, the benefit of such a procedure is that the imputation analysts can apply whatever statistical method they would have applied if there had been no missing values to each completed dataset and then use a simple procedure to combine the results. Standard single imputation procedures can be misleading by causing statistical analysis software to assume that the data has more observations than actually observed and to magnify the confidence by biasing standard errors and confidence intervals. Multiple imputation algorithms avoid this and provide an assessment of the uncertainty caused by the imputation process. The goal of multiple imputation is to provide a completed dataset that allows statistically valid inference, but not to recreate individual missing values by optimal point prediction. Despite some criticism of multiple imputation based on its reliance on simulation, there is evidence that multiple imputation (even with a very simplistic model) is preferable to standard (or even sophisticated) approaches with single imputation in terms of inferences from the completed dataset (e.g. Heitjan and Rubin, 1990). Some ideas on how many imputations are required in multiple imputation can be found in Graham, Olchowski, Gilreath (2007). The paper by Steele, Wang, and Raftery (2010) is an example of further methodological developments of multiple imputation. Rubin (1986) provides insight in using multiple imputation for statistical matching. Gelman et al. (2005) demonstrates how completed datasets obtained by multiple imputation can be used to improve model diagnostics.

A related approach consists in the generation of K jackknife or bootstrap samples (e.g. Burns, 1990; Efron, 1994) of the dataset with missing values, imputing the missing values in the samples, and aggregating the imputed values over the samples to obtain one imputed value for each missing value. A jackknife sample is a subset of the original dataset that is generated by omitting one observation. Usually, K is equal to the number of observations in the original dataset when jackknife samples are used. Bootstrap samples result from sampling with replacement from the original dataset. Multiple imputation is superior to this resampling approach, at least according to the inventor of multiple imputation (Rubin, 1996).

Iterative imputation procedures based on the EM (expectation maximization) algorithm (Dempster, Laird, Rubin, 1977) are also closely related to multiple imputation. The EM algorithm consists of (i) an expectation (E) step that replaces missing value by expected values of the distribution based on estimated distribution parameters and (ii) a maximization (M) step that estimates the parameters of the distribution by maximizing the data log-likelihood function. This means that first missing values are replaced by some initial estimates (may be taken from any other imputation method) and then distribution parameters are estimated based on the completed dataset. These parameter estimates are used to calculate expected values for the missing values and replace them again, creating a new completed dataset for parameter estimation. These steps are repeated until convergence. Data augmentation (Tanner, Wong, 1987) combines the ideas of multiple imputation and EM estimation. Gibbs' sampling (Metropolis et al., 1953; Hastings, 1970; Casella, George, 1992) is another Bayesian simulation method related to multiple imputation and the EM principle.

IV. Applicability of Missing Data Techniques to Time Series Data

The relevance and applicability of the discussed missing data techniques to time series data largely depends on the missing data pattern. In this context, relevant patterns are (i) missing items, (ii) missing periods, (iii) missing variables, and (iv) missing sections.

In case of missing items, all missing data techniques are applicable. Looking at variables, these techniques can be applied both, per variable (variable-wise) and per period (inter-variable). With respect to sections, the techniques can be used either per section (section-wise) or across sections. Variable-wise treatment of missing items means that the temporal pattern of the variable is taken into account. Inter-variable approaches focus on relationships between variables. Section-wise application of missing data techniques corresponds to dealing with missing values in one section at a time, whereas cross-sectional missing data treatment can make use of comparable sections to complete missing items. In any case, the specific time series characteristics of the data should be accounted for as discussed below.

Missing periods in single (uni- or multivariate) time series data can be dealt with variable-wise or in a combined variable-wise and inter-variable approach. Analogously, missing

periods in univariate cross-sectional time series can be filled by section-wise or combined section-wise and cross-sectional procedures. In multivariate cross-sectional time series, missing periods can be completed by variable-wise, section-wise, or combinations of variable-wise, section-wise, inter-variable, and cross-sectional approaches.

For imputing missing variables and/or sections in time series data, additional data sources have to be used irrespective of the type of time series. This is necessary since time is the only common dimension in these missing data patterns.

Missing Data Techniques with Time Series Data

Listwise deletion of time series data is only acceptable in special circumstances (i.e. if MCAR holds). For instance, if lower and higher frequency data are contained in a dataset, one may decide to use all data at the lower frequency only. This may result in a loss of seasonal effects, but the analysis of longer-term trends is still valid.

Deterministic imputation is applicable to time series data. For example, the "Carry Last Value Forward" strategy that replaces missing values by the most recent available value is easy to use. However, the before mentioned concerns related to methodological soundness, introduced bias, and quality measurability apply as for imputing survey data (cf. section 2).

Location-based imputation should be treated with caution in time series. Replacing a missing value with the mean of the same variable in the same section over all available periods will yield unfeasible results in most cases. This also applies to substituting a missing value with the mean of the same variable over all sections. Moving averages of the same variable in the same section over time may be considered instead. This may be regarded as a kind of location-based imputation, but also as model-based imputation. If the number of neighboring periods used in the calculation of the moving average is two, this kind of imputation corresponds to linear interpolation. Moving averages tend to produce rather smooth curves and are not able to predict exceptional peaks, which is especially critical in case of gaps larger than one period. Other types of interpolation, such as splines (Schoenberg, 1946), can also be used for imputing a single variable over time.

Distribution-based imputation for time series missing data requires the usage of special conditional distribution functions that account for lagged variables. Imputing with these more complex distribution functions that include temporal dependencies directly leads to model-based methods.

Model-based methods are most frequently adopted for incomplete time series. Models may either be cross-sectional and/or cross-variable or time series models or combinations of the two and seek to find relationships between sections and/or variables or time periods. Time series techniques applicable to imputation are auto regressive and/or moving average models

(e.g. Ferreio, 1987; Parzen, 1984), state space models (e.g. Durbin, Koopman, 2004), curve fitting or smoothing algorithms (e.g. de Jong, 1995; He, Yucel, Raghunathan, 2011), Kalman filters (Kalman, 1960; Harvey, 1989; Harvey, Pierse, 1984) or other types of dynamic Bayesian networks (e.g. Pearl, 1988, 2009). Apart from time series models, researchers in a broad range of application domains presented different classification algorithms that produce good results in the imputation of time series. Examples are neural networks (Alexiadis et al., 1998; Kihoro et al., 2007), k-means clustering (Hathaway, Bezdek, 2001), seasonal pattern recognition (Chiewchanwattana, Lursinsap, Chu, 2007), or genetic algorithms (Figueroa García, Kalenatic, Lopez Bello, 2008).

Neural networks are non-parametric models that imitate biological neural networks as in human brains to "learn" from data, for example to classify data or to estimate complex relations between input and output variables. One of the most important features of neural networks is their adaptivity. This means that starting from a (random) initial state of the network the interconnected artificial neurons adapt the network to obtain an optimum with respect to some optimization function. For an introduction see for instance Silipo (2003).

Cluster analysis aims at the creation of heterogeneous groups of homogeneous items (clusters). The similarity of items grouped into one cluster as well as the distance between clusters is maximized. K-means clustering is a cluster analysis algorithm that measures the heterogeneity of clusters by means of the distance between cluster centers and the similarity of items within a cluster by means of the distance between the items and the cluster center. An item is assigned to the closest cluster. For an introduction see for example Everitt et al. (2011).

Pattern recognition deals with the identification of certain structures (patterns) in the data by analyzing similarities of observations, observation groups, or observation sequences, for example to classify data. Important applications include image analysis, speech analysis, and person identification. Cluster analysis can be regarded as one class of methods used in pattern recognition. Seasonal pattern recognition refers to finding recurring structures in time series. For an introduction to pattern recognition see for instance Bishop (2006).

Genetic algorithms are heuristics that imitate the process of natural evolution to create useful solutions for optimization and search problems. The evolution starts from a randomly generated population of solutions to the problem and uses techniques such as inheritance, mutation, selection, and crossover to produce a new generation. Typical applications are scheduling problems, gene expression profiling, and linguistic analysis. For an introduction see for example Goldberg (1989).

Donor-based approaches and statistical matching help to identify comparable sections with respect to the same variable at all available time periods or with respect to other variables (at the same or at all available time periods). The former is also referred to as

variable-wise imputation, the latter as inter-variable imputation. In variable-wise donor imputation, the donor section is selected based on the similarity of the trajectories of donor and recipient. This donor is also called "time series donor". In inter-variable imputation, donor selection is based on the similarity of additional characteristics of donor and recipient. This donor is also called "cross-sectional donor". A combination of time series and cross-sectional perspective is feasible as well. Since most variables in time series are metric, nearest neighbor methods are most common. In case of non-metric variables, cold-/hot-deck procedures are feasible methodological options.

Multiple imputation can be used for the completion of time series irrespective of the missing data pattern. In recent years reports on the application of multiple imputation to patchy time series data became more frequent. Honaker and King (2009) describe a multiple imputation approach they developed specifically for cross-sectional time series that allows for smooth time trends, shifts across sections, and spatio-temporal correlations. Landrum and Becker (2001) propose a multiple imputation approach that pools information across geographic units (sections) as well as across different statistical models. He, Yucel, and Raghunathan (2011) present a multiple imputation approach for time series that makes use of non-parametric curve fitting. Palmer (2005) or Cano and Andreu (2010) are further examples for the application of multiple imputation to time series data.

Hybrid procedures are the common practice when imputing time series data. The underlying methods of hybrid procedures need to be carefully selected and combined, taking into account the applicability of the methods to the missing data pattern at hand.

V. Evaluation with Statistical Quality Measures

Thorough monitoring and evaluation of the imputation process is of high importance and vitally contributes to the quality of the resulting data. Imputation quality indicators typically measure

1. the extent to which imputation was required in a dataset,
2. the performance of the applied method,
3. the accuracy of the imputation results,
4. the variability of statistics based on the imputed dataset, and
5. the plausibility of imputed values.

These quality indicators should be reported in publications of statistical analyses that are based on imputed data and are also relevant when disseminating imputed data themselves. Overall indicators for the imputed dataset, variable-specific indicators, and indicators for

each imputed value are discerned. For time series, section- and period-specific indicators can be defined as well.

In order to quantify the plausibility of imputed values, editing procedures (Fellegi, Holt, 1976; de Waal, Pannekoek, Scholtus, 2011) may be applied to the completed dataset (i.e. after imputation). Editing aims at the identification of implausible or infeasible values, usually prior to imputation and any other type of (statistical) processing. Multiple rounds of imputation and plausibility checks may be required. A plausibility checking mechanism may also be integrated in the imputation process in order to avoid the imputation of implausible values. In that case, plausibility of imputed values does not need to be evaluated ex post. For most of the quality indicators, no generally valid thresholds discerning high- and low-quality imputation are available. A comparative usage is recommended to enable decisions between different approaches or different models, parameterizations, or scenarios within one approach. To this end, the relative advantage of one imputation procedure (or the completed dataset) over another procedure (or dataset) with respect to particular quality indicators is calculated. A simple example of this would be an evaluation of the statistical significance of the difference between performance indicators for compared methods which may be tested by means of a repeated measures analysis of (co-) variance with the imputation procedure as a factor. In case of a significant overall difference, pairwise comparisons of the quality measures for the imputation procedures can be carried out via t-tests (Student, 1908) to create a quality ranking of the imputation procedures.

The discussed imputation quality criteria are also applicable to imputation in time series. Adaptations or a different interpretation of the measures based on time series specific data structures are required in many cases, though, as described in the following subsections. In addition, the smoothness of an imputed variable (per section) over time may be analyzed to identify implausible peaks introduced to the trajectory by imputation, especially if imputation was carried out across sections or across variables instead of over time. This can be done by using confidence bands of moving averages or kernel estimates.

A. Degree of missingness

Measures of the degree of missingness in a dataset indicate the quality of the dataset in terms of completeness prior to the imputation process. Comparing these measures prior to and after imputation quantifies the reduction of the degree of missingness that was achieved and can be compared across different procedures. Moreover, several experimental studies, e.g. Kaiser (1986), show the impact of the degree of missingness on other imputation quality measures. An increase in missing values per observation, in the proportion of incomplete observations in the dataset, or in both severely affects (i) the quality of missing value estimates, (ii) the magnitude of the discrepancy in means, and (iii) the preservation of the covariance structure. An increase in sample size can reduce these adverse effects.

Indicators measuring the degree of missingness are mainly ratios of missing values, variables, observations, sections, or time periods with respect to corresponding totals. Examples are

6. the number of observations with one or multiple missing values as a ratio of the total number of observations. For cross-sectional time series, instead of observations sections with missing values are counted.

7. the frequency distribution of the number of variables to be imputed per observation (section) and appropriate distribution measures (e.g. min, max, median).

8. the frequency distribution of the number of periods to be imputed per variable (and/or section) and appropriate distribution measures. This measure only applies to time series.

9. the proportion of observations (or sections or periods) missing for specific variables (or sections or periods).

10. the total number of missing values as a ratio of the total number of cells (i.e. missing or non-missing data points) of the dataset.

For cross-sectional time series, weighted missingness rates can be of interest. Consider the case of sections being countries. Then, each section has a weight, for example in terms of population size, GDP, or purchasing power parity. These weights can be used to calculate weighted proportions of missing data which increases the comparability of missingness rates across countries. The cross-country comparability is relevant in data quality reports for cross-country datasets. In addition, contingency or correlation matrices of the missingness patterns of variables (or observations, sections, periods) may be used to assess the interdependency between the respective missingness patterns. This means that, instead of the variables themselves, their missingness patterns are statistically analyzed to identify relations between variables (or sections or periods) with respect to missingness patterns.

B. Performance of imputation method

Apart from the reduction of the degree of missingness as discussed above, performance criteria for imputation methods are typically method-specific. For some imputation approaches, no method-specific performance evaluation is possible, e.g. for deterministic, location-, or distribution-based imputation.

Overall method-specific performance indicators for model-based methods are standard quality criteria of the regression models involved, such as the coefficient of determination (R^2), various information criteria (e.g. Akaike or Bayes; e.g. Burnham, Anderson, 2002), or the p-values of the regression coefficients. In case of separate models for different variables/sections with missing values, these quality indicators are variable-/section-specific

but can be combined to overall performance measures by some aggregation function. This aggregation may be done by simple (weighted) averaging or by more sophisticated methods of deriving composite indicators from individual indicators (see for example the OECD's handbook on composite indicators (2008)). However, comparability of such composite performance indicators is limited to very specific situations, questioning its value added. An example for such a situation is the application of one and the same imputation method to different vintages of a dataset over time. For hybrid approaches, "one and the same" means that all individual methods used in that approach must stay the same.

For donor-based methods and equivalence class matching, distribution parameters of the usage frequency of individual donors or the values of the distance function between donor and recipient (if applicable) are computed. The less often donors are reused and the closer recipients and their donors are, the higher is the quality of the imputation. Typically, median and maximum are used as aggregation functions. The calculation of these measures is feasible at the overall dataset level as well as at the variable, section, or period level.

For regression-based and propensity score statistical matching, relevant performance indicators are standard regression quality measures (as used for model-based methods) as well as the distance between donor and recipient with respect to the estimated matching variables. As for donor-based methods, usually location and variability measures of the distance are used.

In multiple imputation, performance indicators of the model estimated for the posterior distribution from which the imputation values are drawn are of relevance. Again, standard quality measures for regression models such as R^2 or information criteria can be used. In addition, measures of the variability of the multiple imputed values are considered as imputation performance measures.

At the level of the individual imputed values, an imputation quality report should encompass the standard error or confidence interval for the estimated (=imputed) value in case of model-based imputation, multiple imputation, or regression-based statistical matching, and a reference to the donor used in case of donor-based methods. If distance-based hot-/cold-deck imputation, a nearest-neighbor approach, or equivalence class statistical matching is used, the quality report should additionally provide the value of the distance function between donor and recipient and the number of times the same donor was used in the whole dataset.

For hybrid approaches, performance measures of the component methods should be provided. The idea of aggregating these component performance measures to one “hybrid” performance indicator seems appealing; yet the results are usually hard to interpret.

C. Accuracy of imputation results

Commonly, the following four types of imputation accuracy are discerned (Chambers, 2000):

11. predictive accuracy or effectiveness: maximal preservation of true values,

12. ranking accuracy: maximal preservation of true ordering relationship in imputed values,

13. distributional accuracy: maximal preservation of (marginal and higher order) distributions of true values, and

14. estimation accuracy: maximal preservation of analytic results and conclusions.

This typology constitutes a hierarchy: fulfillment of predictive accuracy, which is the strongest type of accuracy, implies the other three types of accuracy. The relevance of predictive and ranking accuracy depends on the intended usage of the completed dataset. If the dataset is to be publicly released or used for the development of prediction models, these two criteria are crucial. If the objective is to produce and publish aggregated estimates, they are less important. Rubin (e.g. 1996) even claims that the aim of (multiple) imputation should rather be statistically valid inference based on the completed dataset (i.e. estimation accuracy) than the recreation of true values by optimal point prediction (i.e. predictive accuracy). However, it is not feasible to identify all possible analyses that could be carried out for the completed dataset. Therefore, a modified definition of estimation accuracy is usually considered measuring merely the reproduction of lower order moments (at least mean and variance) of the distributions of true values. According to this definition, distributional and estimation accuracy are equivalent for nominal and ordinal variables.

Since the true values of the missing data are unknown, the imputed values cannot be compared to their true counterparts. Hence, accuracy indicators are estimated by treating available values as missing, imputing these fictitiously missing values, and comparing the imputed values to the ignored true values. This technique of leaving out observations in an estimation procedure to validate estimation results is known as cross-validation. Repeated random sub-sampling validation, k-fold cross-validation, and leaving-one-out cross-validation are the most common types of cross-validation. For the validation of imputation results the leaving-one-out approach is typically used. One value is set to missing at a time, and, theoretically, the procedure is repeated for each value. In practice, the repetition is only carried out over a sample of the available values. Cross-validation is usually separately conducted for each variable with missing values. Overall measures can be derived from these variable-specific accuracy measures by aggregation. All accuracy measures depend on the variable type.

For nominal variables, a measure of how closely the imputed values estimate the true values is the proportion of off-diagonal entries for the square table obtained by cross-classifying imputed and true values. If the imputation method preserves individual values, this indicator is equal or close to zero. In case of dichotomous variables, a related measure is calculated as area under the receiver-operating characteristic (ROC) curve (Fawcett, 2006). This curve plots *sensitivity* vs. *1 - specificity*. *Sensitivity* is the proportion of correctly imputed "1"s (also called true positive rate). *1 - specificity* is the proportion of correctly imputed "0"s (also called true negative rate). The closer the value of the area under the ROC curve is to 1, the more accurate are the imputed values. Nominal variables with $p > 2$ categories can be transformed to a set of $p - 1$ dichotomous variables. Sensitivity, specificity, and area under the ROC curve can be calculated separately for each of these dummy variables and then aggregated to a measure for the original variable. The extent to which an imputation procedure preserves the marginal distribution of a nominal variable can be assessed by calculating the value of a Wald-type test statistic that compares imputed and true distributions of the variable across its categories (for details see Chambers, 2000).

For ordinal variables, imputation should satisfy ranking accuracy in addition to predictive and distributional accuracy. Distributional accuracy can be measured in the same way as for nominal variables. Predictive and ranking accuracy can be measured simultaneously. To this end, the magnitude of imputation errors is taken into account by means of the ordinal distance between imputed and true values in the assessment of predictive accuracy (for details see Chambers, 2000).

For metric variables, a measure of the closeness of imputed and true values is the weighted Bravais-Pearson correlation between imputed and true values. For data that are highly skewed or deviating from normality in some other way, this measure is not recommended due to its sensitivity to outliers and influential data values. Instead, it is preferable to focus on estimates of a (robust) regression model without intercept of true values on imputed values. The predictive accuracy assessment corresponds to testing whether the estimated regression parameter is equal to 1. The coefficient of determination of the model is a related regression-based measure for predictive accuracy, whereas the regression mean square error can be regarded as an inverse measure of predictive accuracy. To assess the preservation of the ordering relationship for continuous variables, imputed and true values are replaced by their ranks and the measures of predictive accuracy are calculated. A Kolmogorov-Smirnov test (Massey, 1951) of equality of probability distributions comparing the distributions of imputed and true values evaluates distributional accuracy of an imputation procedure. Another valid choice for assessing distributional accuracy is the Wilcoxon rank-sum or Mann-Whitney-U test (Wilcoxon, 1945; Mann, 1947). An appropriate test for estimation accuracy concerning the mean of the true distribution is a dependent t-test for paired samples (Student, 1908).

D. Variability of statistics based on the imputed dataset

In the evaluation of imputation results, the statistically most relevant measures are bias and variance of the estimates based on the completed data. The complexity of deriving closed-form solutions for variance and bias increases rapidly with the complexity of the missing data patterns and the imputation method. In general, the scope of theoretical work on the direct calculation of variance and bias is limited to rather simple constellations of missing data. This may be one reason for the neglect of the effect of imputation on the variance and bias by many analysts. Thereby, variances are underestimated and the validity of confidence statements is jeopardized (Cox, Folsom, 1978). To resolve this issue, simulation is recommended for the evaluation of imputation results of more complex imputation procedures. For this purpose, Rubin (1987) advocates the routine production of several sets of imputed values under different models or sets of assumptions as part of regular data processing. This leads to estimates of the imputation error and the effects of different models can be investigated as already discussed in the section on multiple imputation above.

Shao and Sitter (1996) propose a related methodology for measuring the imputation variance (for an exemplary application cf. Kaufman, Scheuren, 1997). Bootstrap samples (see, for instance, Efron, 1979; Efron, Gong, 1983) of complete and incomplete observations are generated, and the imputation procedure is applied to each bootstrap sample. The distribution of bootstrap estimates is then used for inference. This approach can also be regarded as a kind of cross-validation. While multiple imputation samples the imputed values from the posterior distribution of the incomplete variable (without replacement), the bootstrap approach draws samples from the original dataset (with replacement) and imputes the missing values for each sample.

VI. Conclusion

Starting from the need to fill data gaps in cross-country time series in order to analyze the effects of the recent financial crisis, this paper provides various methodological choices for imputation of missing data. It investigates data structures and missingness patterns of time series and offers evaluation procedures based on statistical quality criteria for assessing imputation outcomes. The methodological overview aims at raising awareness of data producers and analysts in international financial institutions regarding the challenges posed by missing data as well as techniques for handling them. In addition to promoting the usage of sound imputation procedures, the paper stresses the importance of accompanying any imputation process with reasonable quality indicators. A description of the applied missing data technique and quality assessment results should be published together with the data. This helps reducing ambiguity of incomplete data and facilitates data analysis and interpretation for advising policy makers.

ACKNOWLEDGEMENTS. Special thanks go to our colleagues at the IMF Statistics Department, in particular Mike Seiferling, as well as at the Social Protection & Labor Unit of the World Bank's Human Development Network for valuable feedback and discussions.

VII. REFERENCES

M.C. Alexiadis, P.S.H.S. Sahsamanoglou, I.M. Manousaridis (1998) Short-term forecasting of wind speed and related electrical power. Sol. Energy 63 (1), 61–68.

C. Bishop (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

K.P. Burnham, D.R. Anderson (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

E.M. Burns (1990) Multiple and replicate item imputation in a complex sample survey. Proceedings of the Bureau of the Census Annual Research Conference 1990.

S. Cano, J. Andreu (2010) Using multiple imputation to simulate time series. Proceedings of the 9th WSEAS International Conference on Applied Computer and Applied Computational Science (ACACOS'10), 117–122.

J.R. Carpenter, M.G. Kenward, I.R. White (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Statistical Methods in Medical Research 16(3), 259-275.

G. Casella, E.I. George (1992) Explaining the Gibbs Sampler. The American Statistician 46 (3), 167-174.

R. Chambers (2000) Evaluation criteria for statistical editing an imputation. National Statistics Methodological Series No. 28. ONS, UK.

S. Chiewchanwattana, C. Lursinsap, C.-H.H. Chu (2007) Imputing incomplete time-series data based on varied-window similarity measure of data sequences. Pattern Recognition Letters 28 (9), 1091-1103.

W.G. Cochran (1977) Sampling Techniques. Wiley, New York.

M.J. Colledge, J.H. Johnson, R. Paré, I.G. Sande (1978) Large scale imputation of survey data. ASA Proc. of the Section on Survey Research Methods, 431-436.

B.G. Cox, R.E. Folsom (1978) An empirical investigation of alternate item nonresponse adjustments. ASA Proc. of the Section on Survey Research Methods, 219-223.

A.P. Dempster, N.M. Laird, D.B. Rubin (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39 (1), 1–38.

M. Denk (2008) A Framework for Statistical Entity Identification to Enhance Data Quality. CENEX Workshop on Statistical Methodology, Area Integration of Surveys and Administrative Data, 29.-30.5.2008, Vienna, Austria.

J. Durbin, S.J. Koopman (2004) Time Series Analysis by State Space Methods. Oxford Univ. Press.

B. Efron (1994) Missing data, imputation, and the bootstrap (with discussion). Journal of the American Statistical Association 89, 463-478.

B. Efron, G. Gong (1983) A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation. The American Statistician 37 (1), 36-48.

B.S. Everitt, S. Landau, M. Leese, D. Stahl (2011) Cluster Analysis, 5th edition. Wiley, Chichester, UK.

T. Fawcett (2006) An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

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

I. P Fellegi, A. B Sunter (1969) A theory for record linkage. Journal of the American Statistical Association 64 (328), 1183–1210.

O. Ferreio (1987) Methodologies for the estimation of missing observations in time series. Statistical Probability Letters 5 (1), 65–69.

J. Figueroa García, D. Kalenatic, C. Lopez Bello (2008) Missing Data Imputation in Time Series by Evolutionary Algorithms. Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, 275–283.

A. Gelman, I. Van Mechelen, G. Verbeke, D.F. Heitjan, and M. Meulders (2005) Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. Biometrics 61, 74–85.

D.A. Goldberg (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Professional, Boston (MA).

J.W. Graham, A.E. Olchowski, T.D. Gilreath (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science 8(3), 206–213.

A.C. Harvey (1989) Forecasting, structural time series models and the Kalman filter. Cambridge Univ. Press.

A.C. Harvey, R.G. Pierse (1984) Estimating missing observations in economic time series. Journal of the American Statistical Association 79(385),125-131.

W.K. Hastings (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57 (1), 97–109.

R.J. Hathaway, J.C. Bezdek (2001) Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man. Cybernet. Part B 31 (5), 735–744.

Y. He, R. Yucel, T.E. Raghunathan (2011) A functional multiple imputation approach to incomplete longitudinal data. Statistics in Medicine Early View (Articles online in advance of print), Wiley Online Library http://wileyonlinelibrary.com/.

D.F. Heitjan, D.B. Rubin (1990) Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistical Association 85, 304-314.

K. Hollenbeck, P. Doyle (1979) Distributional characteristics of a merged microdata file. ASA Proc. of the Survey Research Methods Section, 418-420.

D. Holt, T.M.F. Smith (1979) Post-stratification. Journal of the Royal Statistical Society A 142, 33-36.

P. de Jong (1995) The simulation smoother for time series models. Biometrika 82(2), 339-350.

J.B. Kadane (1978) Some statistical problems in merging data files. 1978 Compendium of Tax Research, US Dept. of the Treasury, 159–171. (Reprinted in Journal of Official Statistics 17 (3), 423–433.)

R.E. Kalman (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82 (1), 35–45.

J. Kaiser (1986) Comparison of hot-deck variations in imputing missing values. ASA Proc. of the Section on Survey Research Methods, 653-658.

G. Kaplan, S. Schulhofer-Wohl (2010) Interstate migration has fallen less than you think: consequences of hot deck imputation in the Current Population Survey. NBER Working Paper Series 16536.

S. Kaufman, F. Scheuren (1997) Applying Mass Imputation Using the Schools and Staffing Survey Data. Proceedings of the Survey Research Methods Section, American Statistical Association, 129-134.

J.K. Kihoro, R.O. Otieno, C.Wafula (2007) Seasonal time series data imputation: Comparison between feed forward neural networks and parametric approaches. East African Journal of Statistics 1(1), 68-83.

M.B. Landrum, M.P. Becker (2001) A multiple imputation strategy for incomplete longitudinal data. Statistics in Medicine 20, 2741-2760.

R.J.A. Little (1982) Models for non-response in sample surveys. Journal of the American Statistical Association 77, 237-250.

R.J.A. Little, D.B. Rubin (2002) Statistical analysis with missing data, 2nd ed. Wiley, New York.

H.B. Mann (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 18(1), 50-60.

D. Margolis, D. Newhouse, M. Weber (2010a) What's happening now? Model-Based Imputation of Low-Frequency Variables in Macroeconomic Panel Data (ITSEM). World Bank Mimeo, January 2010.

D. Margolis, D. Newhouse, M. Weber (2010b) Employment changes and the crisis: What happened in data-poor countries? 5th IZA/World Bank Conference: Employment and Development, May 3-4, 2010, Cape Town, South Africa.

F.J. Massey (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 46(253), 68-78.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller (1953) Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics 21 (6), 1087–1092.

C. Moriarity, F. Scheuren (2001) Statistical matching: A paradigm for assessing the uncertainty in the procedure. Journal of Official Statistics 17 (3), 407–422.

OECD (2008) Handbook on Constructing Composite Indicators - Methodology and User Guide. Downloadable from the OECD's online bookshop at http://browse.oecdbookshop.org/oecd/pdfs/free/3008251e.pdf.

B.A. Okner (1972) Constructing a new database from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement 1, 325–342.

F.T. Palmer (2005) Multiple imputation of time series: an application to the construction of historical price indexes. BILTOKI No. 3/2005.

E. Parzen, ed. (1984) Time series analysis of irregularly observed data. Lecture Notes in Statistics 25. Springer, New York.

J. Pearl (1988) Probabilistic Reasoning in Intelligent Systems. Morgan-Kaufmann, San Francisco.

J. Pearl (2009) Causality: Models, Reasoning, and Inference, 2nd edition. Cambridge University Press, New York.

S. Raessler (2002) Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. Springer, New York.

P.R. Rosenbaum, D.B. Rubin (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41-55.

D.B. Rubin (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics 4, 87-94.

D.B. Rubin (1987) Multiple imputation for nonresponse in surveys. Wiley, New York.

D.B. Rubin (1996) Multiple imputation after 18+ years. Journal of the American Statistical Association, 91 (434), 473-489.

I.G. Sande (1982) Imputation in surveys: Coping with reality. The American Statistician 36 (3), 145-152.

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

I.J. Schoenberg (1946) Contributions to the problem of approximation of equidistant data by analytic functions. Quarterly Applied Mathematics 4, 45–99 & 112–141.

J. Shao, R. Sitter (1996) Bootstrap for Imputed Survey Data. Journal of the American Statistical Association 91 (435), 1278-1288.

R. Silipo (2003) Neural Networks. Intelligent Data Analysis, 2nd edition (M. Berthold, D.J. Hand, eds.), Springer, New York.

R.J. Steele, N. Wang, A.E. Raftery (2010) Inference from multiple imputation for missing data using mixtures of normals. Statistical Methodology 7(3), 351–365.

Student (1908) The probable error of a mean. Biometrika 6(1), 1–25.

M.A. Tanner, W.H. Wong (1987) The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82 (398), 528-540.

A. Tarsitano, M. Falcone (2010) Missing-Values Adjustment For Mixed-Type Data. *Working Paper WP15-2010,* Department of Economics and Statistics, University of Calabria. Download: http://www.ecostat.unical.it/RePEc/WorkingPapers/WP15_2010.pdf

P. Van der Putten (2000) Data fusion: A way to provide more data to mine in? Proc. 12th Belgian-Dutch Artificial Intelligence Conference (BNAIC'2000), De Efteling, Kaatsheuvel, The Netherlands.

T. de Waal, J. Pannekoek, S. Scholtus (2011) Handbook of Statistical Data Editing and Imputation. Wiley, New York.

F. Wilcoxon (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6), 80-83.

W.E. Winkler (1995) Matching and Record Linkage. In B.G. Cox et al. (eds.) Business Survey Methods, Wiley, New York, 355–384.