Environmental Earth Sciences (2018) 77:47
https://doi.org/10.1007/s12665-018-7228-6

ORIGINAL ARTICLE

A review on missing hydrological data processing

Yongbo Gao[1,2] · Christoph Merz[1,2] · Gunnar Lischeid[1,3] · Michael Schneider[2]

Received: 9 September 2016 / Accepted: 3 January 2018 / Published online: 19 January 2018

Abstract

Like almost all fields of science, hydrology has benefited to a large extent from the tremendous improvements in scientific instruments that are able to collect long-time data series and an increase in available computational power and storage capabilities over the last decades. Many model applications and statistical analyses (e.g., extreme value analysis) are based on these time series. Consequently, the quality and the completeness of these time series are essential. Preprocessing of raw data sets by filling data gaps is thus a necessary procedure. Several interpolation techniques with different complexity are available ranging from rather simple to extremely challenging approaches. In this paper, various imputation methods available to the hydrological researchers are reviewed with regard to their suitability for filling gaps in the context of solving hydrological questions. The methodological approaches include arithmetic mean imputation, principal component analysis, regression-based methods and multiple imputation methods. In particular, autoregressive conditional heteroscedasticity (ARCH) models which originate from finance and econometrics will be discussed regarding their applicability to data series characterized by non-constant volatility and heteroscedasticity in hydrological contexts. The review shows that methodological advances driven by other fields of research bear relevance for a more intensive use of these methods in hydrology. Up to now, the hydrological community has paid little attention to the imputation ability of time series models in general and ARCH models in particular.

Keywords Missing data · Imputation · Hydrological time series analysis · ARCH · ARIMA · Heteroscedasticity

Introduction

The phenomenon of missing data has been discussed extensively within and beyond statistics (Schafer and Graham 2002). It is a common problem in empirical studies in the social, medical or geographical sciences and occurs for a number of different reasons, including erroneous manual data entry, equipment errors during the collection of data or a loss of data due to defective storage technologies (Tannenbaum 2009).

It is a well-known fact, however, that numerous hydrological research databases contain missing values (Elshorbagy et al. 2002). There are many, often idiosyncratic, reasons for data to be missing. They include the failure of observation stations, incomparable measurements, manual data entry procedures that are prone to error and equipment error (Johnston 1999). Missing data generally reduces the power and the precision of statistical research methods (Roth et al. 1999). In addition to reducing the power of these methods, missing data can also lead to biased estimates of the relations between two or more variables (Pigott 2001). Both problems—reduction in power and bias of estimates—can lead to inaccurate conclusions in analyses of datasets that contain missing data (Graham 2009). Missing data is a relevant problem in deterministic hydrological modeling which relies on observed data including hydrometeorological input parameters like temperature and precipitation to model complex relations between variables relating to weather conditions and geographic surroundings (Gill et al. 2007; Kim and Ryu 2016; Henn et al. 2013). Therefore, gap-free time

✉ Yongbo Gao
cmerz@zalf.de

[1] Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Str. 84, 15374 Müncheberg, Germany

[2] Institute of Geological Sciences, Workgroup Hydrogeology, Freie Universität Berlin, Malteser Str. 74-100, 12249 Berlin, Germany

[3] Department of Earth and Environmental Science, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany

series are a necessary prerequisite for many statistical and deterministic model approaches in hydrology.

Common statistical approaches for hydrological analysis including the determination of the flow duration curve, the autocorrelation function, spectrum analysis and extreme value analysis, etc., based on complete time series without data gaps. Missing data create problems for all of the approaches listed above. When the available time series are long enough, researchers can use a subset of the data that contains a complete set of observations for a certain period. More often, due to personal and financial limitations, the available data have been collected over shorter observational periods. In these cases, the application of statistical methods that strictly require complete time series can be severely aggravated. To minimize the efforts required to tackle the missing data problem, imputation methods are favored using statistical models with different degrees of sophistication. The goal is to find a method that relays as much as possible on the measured data and as little as possible on theoretic assumptions (Aubin and Bertrand-Krajewski 2014).

These approaches are based on the assumption that the sample to be analyzed is a random sample from the entire database and hence contains a complete set of information (Farhangfar et al. 2008). Over recent decades, imputation methods which attempt to 'fix' datasets characterized by missing data by replacing them with inserting numerical values have improved dramatically (Peugh and Enders 2004). The rise of more sophisticated imputation methods has led many researchers to favor replacing missing values with imputed values over excluding them from the analysis entirely (Saunders et al. 2006). It is obvious that experts of local and regional water authorities try to minimize the gaps in time series of their monitoring programs when preparing data sets required for water resources management. In the last 20 years, statisticians have introduced imputation methods such as regression-based imputation, data imputation based on principal component analysis (PCA) or maximum likelihood techniques using the 'expectation–maximization' (EM) algorithm as well as 'multiple imputation' (MI). These methods offer promising solutions but their performance depend on the exact application and a knowledge of the theoretical background (Soley-Bori 2013).

In general, the choice of a specific imputation method is determined by the nature of the process generating the original data. Statistical models can be used to fix the data gap in the measured time series of different parameters in hydrology like rainfall, ground- and surface water level or temperature. Due to the different characteristics of rainfall and temperature behavior regarding autocorrelation and variance, different imputation methods have to be applied. For instance, data is often cross-sectional in nature and a familiar statistical tool such as PCA or linear regression approaches can be used for imputation purposes. In hydrological settings, however, the choice of an appropriate imputation method needs to take into account the most important features of hydrological data. Hydrological data are often time series data that are characterized by stable trends over time and a high autocorrelation of the observations. Moreover, hydrological time series often display random deviations from these trends and these deviations are not constant over time (Guzman et al. 2013). Given these features of the data-generating process underlying the hydrological data, the imputation of missing values should be based on statistical time series methods that take into account the time series nature of hydrological data. For instance, singular spectrum analysis (SSA) models or autoregressive moving average/autoregressive integrated moving average (ARMA/ARIMA) models have been applied in hydrological settings to predict medium and long-term hydrological runoff (Zhang et al. 2011). One feature of time series data that has received little attention in the hydrological literature so far is non-constant deviations around a trend, which is called heteroscedasticity. For this reason, autoregressive conditional heteroscedasticity (ARCH) time series models, which originate from finance and econometrics, will be discussed below. ARCH models may be used not only to explain and characterize observed hydrological time series but also to impute missing observations in existing datasets which are characterized by non-constant high variability.

The goal of this paper is to present an overview of different imputation methods that are available to time series analysis in hydrology. Imputation in hydrology has very often been done in an ad hoc manner, lacking a clear theoretical basis and a sound selection of methods depending on the statistical properties of the respective observable and the respective research question. This review paper aims at increasing awareness among the use of different imputation techniques in hydrologic context. In this attempt, particular emphasis is laid on the fact that hydrological data can be characterized as time series data in which statistical patterns such as autocorrelation or seasonality emerge over time and can be exploited for imputation purposes. A special focus will be laid on ARCH models and the discussion of the extent to which they might be applied to hydrological settings of missing data.

Patterns of missing data

To date, a variety of different statistical techniques are available to address the problems arising from missing data (Puma et al. 2009). An understanding of these methods is increasingly important as having complete and accurate databases is often the prerequisite for applying increasingly sophisticated statistical methods. Often, it is tempting to follow the simplest way of dealing with missing data, which consists of simply discarding (i.e., deleting) observations where information on one or more variables is missing. This approach is also one of the default options for statistical analysis in most software packages and is called 'listwise' or 'pairwise' deletion (Harrington 2008). Nevertheless, it must be pointed out that deleting of even a small share of all observations in a dataset will reduce the statistical power and the accuracy of the analyses undertaken (Rubin and Little 2002).

Before any missing data can be imputed, the most important question that researchers have to address relates to the underlying patterns of the 'missingness' or rather incompleteness of the data. In particular, they have to diagnose whether data in some observations is missing randomly or whether the observed incompleteness follows a particular pattern (Little and Rubin 1987). In this context, the classification system of missing data outlined by Rubin (1976) and colleagues remains in widespread use today. Following Rubin (1976), missing data can be seen as a probabilistic process and allows three so-called missing data mechanisms to be identified which describe the relationship between the measured variables and the probability of a missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Assume that our data contains one variable of primary interest Y and a number of additional variables, referred to as a vector X. Following this notation, and with m being an indicator variable for missing observations in Y, i.e., $m = 1$ if a data point is missing and $m = 0$ if a data point has been observed, the probability that a value in Y is missing can be expressed as a function of Y and X with

$$\Pr\,(m\;=\;1|\,X,\,Y). \tag{1}$$

First, suppose that the probability of a missing observation in Y is completely independent of any observed or unobserved measurements of this variable or other variables X and also independent of the other observations in the dataset. If this is the case, the absence of a value in a given observation is called missing completely at random (MCAR) (Allison 2012). This mechanism is what researchers consider to be purely random missingness. The case of MCAR missing data causes the fewest problems for statistical analyses. In a dataset including missing values that are MCAR, the subset of all observations containing the missing data can be deleted. The remaining subset then contains all observations with complete information. This approach often is called listwise/pairwise deletion (McKnight et al. 2007). As the resulting dataset containing only the observations with complete data is a random sample from the initial data, it can easily be shown that results based on its statistical analysis will be unbiased (Rubin 1976). Mathematically, MCAR implies that

$$\Pr\,(m\;=\;1|X,\,Y)\;=\;\Pr\,(m\;=\;1). \tag{2}$$

Another pattern of missing values is called missing at random (MAR). MAR is a less restrictive assumption regarding the pattern of missing values compared to MCAR. When data are missing at random, the probability of missing data in a variable for a given observation is only related to any other observed variable rather than to Y itself. This implies that

$$\Pr\,(m = 1|X,\ Y) = \Pr\,(m\;=\;1|X). \tag{3}$$

Data that contain information MAR require more attention than data is MCAR: all simple imputation methods for missing data, i.e., listwise and pairwise deletion or arithmetic mean imputation, will give biased results in analyses of the relations between variables in the dataset (Pigott 2001). Nevertheless, unbiased results can be obtained in the case of data MAR. This requires the application of more sophisticated imputation methods, however, including single and multiple imputations (Donders et al. 2006).

In cases where neither the MCAR nor the MAR assumption holds, data are said to be 'Missing Not At Random' (MNAR) (McKnight et al. 2007). If cases are MNAR, there is a relationship between the variables that include missing data and those for which the values are present and hence the following equation is valid

$$\Pr\,(m\;=\;1|X,\,Y). \tag{4}$$

When missing data are MNAR, results from statistical analyses will be biased and there is little what imputation techniques can do to ease the problem (Donders et al. 2006). It is thus important to investigate whether the missing pattern is random or not before any statistical test is conducted. For a full discussion, see Rubin and Little (2002).

An overview of traditional techniques for handling missing data

Dozens of techniques to deal with the missing data problem have been used over the decades (Baraldi and Enders 2010). The more common traditional approaches to deal with missing data include removing the values with incomplete data/deletion, or so-called single-imputation methods where missing values are replaced (Peugh and Enders 2004). Whereas deletion methods reduce the sample size, the purpose of single-imputation methods is to retain the sample size and statistical power in subsequent analyses (Cool 2000). However, single-imputation methods have drawbacks which are addressed by more complicated multiple imputation methods which are often based on Monte-Carlo-type simulations and require more computational sophistication

Table 1 Summary of common imputation methods used to handle missing data in hydrological data sets

Imputation method	Problems/limitations	Literature
Listwise deletion	Reduces sample size, produces bias when MCAR is violated	Mcdonald et al. (2000)
Pairwise deletion	Reduces sample size, only practical for small portions of missing data	Wothke (2000)
Single imputation	Generates biased parameter estimation	Enders (2010)
Arithmetic mean and median imputation	Decreases variance	Roth (1994), McKnight et al. (2007)
Regression-based imputation	Danger of overestimation of correlation, decreases variance	Greenland and Finkle (1995)
Principal component based imputation	Complex identification of dimensions needed for processing	Pandey et al. (2011)
Multiple imputation	Requires statistical sophistication and expertise	Graham and Hofer (2000)

than simple imputation methods.[1] This section reviews the most important imputation methods (Table 1). Despite their widespread use, these methods still have shortcomings in their procedures which will be also illustrated in this section.

Listwise and pairwise deletion

The elimination of all observations which have missing data in one or more variables is called listwise deletion (Mcdonald et al. 2000). The primary benefit of listwise deletion is convenience (King et al. 1998). This approach has several drawbacks, as addressing incomplete data by deleting observations will inevitably reduce the sample size (Rubin et al. 2007). It is a well-established fact in statistics that smaller sample sizes reduce the statistical power and precision of standard statistical procedures (Rubin and Little 2002). A reduction in the precision of tests and estimates renders inference (such as hypothesis testing) conservative. A more severe effect could be that it can introduce a systematic bias. If the data is MCAR, a sample excluding observations with missing values will be a random draw from the complete sample and estimates remain unbiased. If, however, the relatively strong assumption of MCAR is violated, the deletion of observations with missing data will bias the value of the estimates of interest. A simulation by Raaijmakers (1999) demonstrated that the statistical power is reduced between 35% (with 10% missing data) and 98% (with 30% missing data) using listwise deletion.

The elimination of observations on a case-by-case basis depending on which variables are used in a specific analysis is called pairwise deletion. It is different to listwise deletion as an observation is deleted only if a variable used in the analysis contains a missing value (Wothke 2000). For example, if a respondent does not provide information on variable *A*, the respondent's data could be used to calculate other correlations, such as the one between *B* and *C*. Pairwise deletion is often an improvement in listwise deletion because it preserves much more information by minimizing the number of cases discarded compared to listwise deletion (Roth 1994). Among the most important problems of pairwise deletion is the limited comparability of different analyses as the number of observations varies between different pairwise comparisons (Croninger and Douglas 2005). Moreover, estimates of covariances and correlations might be biased when using pairwise deletion since different parts of the sample are used for each analysis (Kim and Curry 1977). Despite its convenience, this method is practical only when the data contains a relatively small portion of observations with missing data. If a negligible share of the observations contains missing data, the analysis of the remaining observations will not lead to serious inference problems (Tsikriktsis 2005). Deletion techniques are the default options for missing data techniques in most statistical software packages, and these techniques are probably the most basic methods of handling missing data (Marsh 1998).

Single-imputation methods

Single-imputation approaches generate a single replacement for each missing value with suitable values prior to the actual analysis of the data (Enders 2010). A variety of different missing data imputation methods have been developed over the years and are readily available in most standard statistical packages. As has been discussed above, all imputation methods produce biased results if the relatively strong MCAR assumption is violated. In particular, imputation is advantageous compared to listwise or pairwise deletion as it generates a complete dataset. Hence, it also makes use of the data that deletion techniques would discard. Nevertheless, as will be discussed below, these methods have potentially drawbacks and even in an ideal MCAR situation most of these approaches generate biased parameter estimation.

[1] As in "Multiple imputation" section below, multiple imputation generates multiple datasets containing imputed values which are enhanced by a random error term. The desired statistical analyses are then carried out multiple times on these different datasets and their results aggregated. This approach allows getting more appropriated standard errors on the estimates of the desired parameters.

Many different single-imputation methods have been introduced and applied: arithmetic mean imputation, principal component analysis (PCA) and regression-based imputation are the most commonly known and will be briefly introduced below.

Arithmetic mean and median imputation

Arithmetic mean imputation replaces missing values in a variable with the arithmetic mean of the observed values of the same variable (Roth 1994). Median imputation replaces missing values with the median value of the observed values of the same variable (McKnight et al. 2007). Both approaches are very convenient since they generate a complete dataset easily (Hawkins and Merriam 1991). Median imputation is preferable when the distribution of the underlying variable is not symmetric but instead skewed (McKnight et al. 2007).

However, even in situations where the strong MCAR assumption holds, these approaches distort the resulting parameter estimates (Enders 2010). For instance, they attenuate the standard deviation and the variance of estimates obtained from analyses of mean-imputed variables (Baraldi and Enders 2010). The reason for the reduction in the standard deviation of estimates is that the imputed values are identical and at the center of the distribution, which reduces the variability of the data (Little 1988). This fact also attenuates the magnitude of estimated covariances and correlations between mean-imputed variables and other variables in a dataset (Malhotra 1987).

Regression-based imputation

Regression-based imputation replaces missing data with predicted values from a regression estimation (Greenland and Finkle 1995). The basic idea behind this method is to use information from all observations with complete values in the variables of interest to fill in the incomplete values which is intuitively appealingly (Frane 1976). Different variables tend to be correlated in many applications (Allison 2001). Exploiting information from all observations with complete information is a strategy which regression-based imputation methods share with multiple imputation and maximum likelihood imputation methods, although the former approach does so in a less sophisticated way (Raghunathan 2004). Note that maximum likelihood imputations refer not to the estimation method used by the regression-based imputation methods but rather to the technique of selecting among different values that might be chosen to assess a missing value.

The first step in the imputation process is to estimate regression equations that relates the variable that contains missing data (the dependent variable of the regression) to a set of variables featuring complete information across all observations in the dataset (independent variables of the regression). This regression is estimated only for the subset of the data that contains all observations that have complete information both for the dependent variable and the independent variables. The results of the regression are estimates due to the relation of independent to dependent variables.

The second step exploits this information. Using the regression results from the first step, missing values for the observations that could not have been included in the regression are replaced by predictions obtained by combining the observed values of the independent variables with the estimates, from the first step, of how they are related to the dependent variables. These predicted values fill in the missing values and produce a complete dataset (Frane 1976). In the case of k variables with $n - r$ missing values in the k-th variable (n being the total number of observations and r being the number of complete observations), a linear regression can be estimated based on all r complete observations. The regression yields the estimated regression coefficient. Based on these estimates, missing values in the k-th variable can be predicted, i.e., imputed with

$$\widehat{y_{t,k}} = \widehat{\beta}_0 + \sum_{j=1}^{k-1} \widehat{\beta}_j y_{i,j} \quad \forall i \in [r, n]. \tag{5}$$

While regression-based imputations most frequently rely on simple linear regressions, it is worth noting that more flexible regression approaches can equally be used and might even be more advantageous depending on the application. In "Introduction to time series analysis and its application to imputation of missing values" section, we will discuss to what extent time series regression approaches can be used in regression-based imputations of hydrological data.

From a methodological viewpoint, regression imputation is superior to mean imputation, but it can lead to predictable biases (van der Heijden et al. 2006). In particular, regression-based imputation methods lead to the opposite problem as mean imputation as missing data is replaced with values that are highly correlated to other variables in the data. Consequently, the application of regression-based imputation methods will lead to overestimated correlations and R^2 statistics in subsequent data analysis.

Imputation based on principal component analysis (PCA)

Principal component analysis (PCA) was originally conceived as a multivariate exploratory data analysis technique. It is used to extract patterns from datasets by transforming the data into a new coordinate system such that the greatest variance, by some projection of the data, comes to lie on the first coordinate (called the first principal component). The second greatest variance lies on the second coordinate, and so on. Moreover, PCA can be used to compress high-dimensional vectors into lower-dimensional ones (Pandey et al. 2011). The principal idea behind PCA is to find a smaller dimensional linear representation of data vectors so that the original data can be approximated from the lower-dimensional representation with minimal mean square error. Graphically, PCA can then be used to interpret a projection of the original data points on a lower-dimensional space, which minimizes the reconstruction error (Jolliffe 1993).

Formally, PCA can be expressed as follows. Assume that observed data points $x_1, x_2, \ldots, x_n \in R^p$ are p-dimensional vectors. PCA defines a projection of these data on a q-dimensional space (with $q \leq p$) as

$$f(\lambda) = \mu + \upsilon_q \lambda. \tag{6}$$

In this q-dimensional model, μ is a vector of the mean values of length p, υ_q is a $p \times q$ matrix with q orthogonal unit vectors and λ is the q-dimensional projection of each original data vector x. A projection of the original data can be found by maximizing the variance of the projection of the original data along the new (reduced) dimensions of the projection space

$$\min_{\mu,\lambda_{1\ldots N},\upsilon_q} \sum_{n=1}^{N} x_n - \mu - \upsilon_q \lambda_n. \tag{7}$$

Here, μ can be interpreted as the intercept of the projection space in the original space, $\lambda_1,\ldots,\lambda_n$ are the projection coordinates of the original observations $x_1,\ldots,x_n$. Note that PCA can be also be derived from a maximization of the variance of the projected data points along the new dimensions. The results are computationally equivalent.

While originally not devised as an imputation method, PCA can be used to replace missing values in a dataset and hence also as an imputation tool. For this purpose, an iterative PCA algorithm was proposed by Kiers (1997). The algorithm can be summarized as follows:

1. Missing values are initially replaced by the sample mean.
2. PCA is conducted on the now complete dataset by minimizing the reconstruction error as described above to derive μ, λ_n and υ_q.
3. Initially missing values are replaced by imputed values based on the results from step (2) with $x_n = \mu + \upsilon_q \times \lambda_n$.
4. Steps (2) and (3) are repeated until the imputed values of initially missing values converge.

It can be shown that the iterative PCA corresponds to an expectation–maximization (EM) algorithm and is thus often named an EM-PCA algorithm (de Leeuw 1986; Dempster et al. 1977). This approach is computationally more efficient as it does not require the computation of the full covariance matrix. It needs to be stressed that one the biggest disadvantages of PCA is the number of dimensions q in PCA needs to be chosen by the analyst and is not a result of the analysis. This has been identified as a core issue and is a very difficult task that has been extensively discussed in the literature. For a treatment of this issue see, for instance, Jolliffe (2002).

Multiple imputation

In order to ease the negative impact of regression imputation mentioned above, more sophisticated approaches have been developed. The principal idea here is to replace each missing item with two or more plausible values, representing a distribution of possibilities. Therefore, these are known as multiple imputation approaches (MI) (Graham and Hofer 2000). Recent advances in computational power have made multiple imputation available as relevant procedures are included in standard statistical software packages more frequently. The biggest advantage of multiple imputation is that inference regarding statistics such as correlations error obtained from multiple imputation are not overestimated because they incorporate uncertainty due to missing data (Lee and Carlin 2010). However, there are some disadvantages in MI. The biggest disadvantage of MI is that it requires more computational effort, since both the imputation and the subsequent analyses have to be carried out multiple times (Rubin 2004). It should be noted, however, that given the advances in computing hardware and software this is not a burden in practice and most statistical software packages nowadays contain routines for MI.

While there are different approaches to MI imputation, the underlying sequence of computations steps is similar (Allison 2000): First, the missing data are imputed by an appropriate model M times to produce M complete datasets (Fig. 1). Most often, regression-based imputation techniques are used in this step. In each of the M steps, the predicted values from the regression analysis are varied by a random term of zero mean and a specified standard deviation. After this step, the desired statistical analysis can be carried out on each of the M datasets using standard complete data analysis methods. This yields a set of M results, of which average values and standard errors can thereafter be computed (Allison 2000). This approach avoids the underestimation of standard errors and hence is often preferable to single-imputation methods.

Despite its desirable properties, multiple imputation requires statistical and computational sophistication. For this reason, the remainder of the paper focuses on single-imputation methods, which still seem to be more frequently used in hydrological settings.

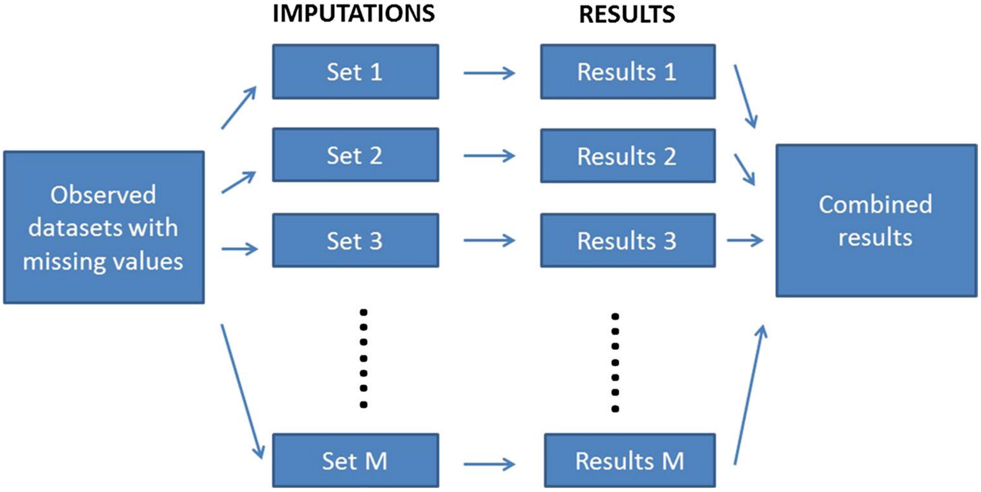

Fig. 1 Scheme of multiple imputation

Introduction to time series analysis and its application to imputation of missing values

Overview

A time series, in our context, is a discrete time series defined as a series of observations of Y where are observations over several consecutive time periods $t = 1,\ldots,T$. y_t might be the amount of discharge from a given measurement station that is measured on a daily basis. The hydrologist might be interested in analyzing runoff over time and how it depends on different types of boundary conditions. The assumption that observations in the dataset are independent thus seems far fetched. In hydrology, it is reasonable to assume that there are periods characterized by high runoff, in which today's runoff will be related to the amount of runoff the day before and hence past values are correlated with today's runoff value.

Data imputation approaches can make efficient use of dependencies between different observations in a time series, defined as data resulting from the observation of subjects which are repeatedly measured over a series of time points (Hedeker and Gibbons 1997). In contrast to conventional approaches, time series techniques allow for the assumption that y_t is not independent of preceding observations of y. This is called autocorrelation, or serial correlation, where y_t is a function of a previous value of y. Adapted approaches exploit autocorrelation to a model if a given phenomenon is not only based on conditions in t but also on its own history (for instance Y_{t-1}). In the following, the adaptation of PCA to time series data which is often called singular spectrum analysis (SSA) is discussed before we move on to a more comprehensive discussion of time series regression techniques.

Singular spectrum analysis (SSA)

The aim of singular spectrum analysis (SSA) is to decompose a time series into regular oscillatory components and random noise, applying the principles of PCA to time series data (Hassani 2007). For this reason, SSA can be considered a time series version of PCA. SSA, on the other hand, can be applied to a univariate time series y_t with $t = 1,\ldots,T$ in order to separate a signal in a time series (trends or oscillatory movements) from a noise component that is random. To that end, a so-called trajectory matrix is formed from the original data. Taking the time series $\mathrm{Y} = (y_1, y_2,\ldots, y_n)$ of length n and choosing a window length L (with $1 < L < n$), $K = n - L + 1$ lagged vectors x_j of the original time series can be generated with $x_j = (y_i, y_{i+1},\ldots,y_{j+L-1})$ for $j = 1,2,\ldots,K$. These vectors form the trajectory matrix X with

$$X = \left[X_1,\ldots,X_K\right]' = \begin{bmatrix} y_1 & y_2 & \ldots & y_L \\ y_2 & y_3 & & y_{L+1} \\ \vdots & & \ddots & \vdots \\ y_K & y_{K+1} & \cdots & y_n \end{bmatrix}. \tag{8}$$

In a second step, and similarly to PCA, the trajectory matrix is then subjected to a single value decomposition yielding a set of so-called eigentriples which contain the principal components of Y (Wall et al. 2003).

By projecting the principal components back onto the eigenvectors, a time series (referred to as the 'reconstructed components') can be recovered in the original time units, each one corresponding to one of the PCs.

This third step of SSA splits the elementary matrices X_i into several groups and sums the matrices within each group. Finally, diagonal averaging transfers each of these matrices into a time series, which is an additive component of the initial series y_t.

It should be noted here that the window length L must be chosen by the researcher. The choice of L is important, as it defines the maximum length of the oscillations that can be detected employing SSA. While the literature provides some guidance by providing rules of thumb for the choice of L, ultimately any ex-ante choice of L remains arbitrary ,and there are no tests available that would allow statistical inference to be conducted regarding the choice of L. In the context of imputation, Kondrashov and Ghil (2006) propose an iterative approach to determine a suitable choice of L. In particular, they iteratively produce estimates of missing data points, which are then used to compute a self-consistent lag-covariance matrix and its empirical orthogonal functions. This approach allows the window length L to be optimized by cross-validation.

Time series regression

Autoregressive and moving average models (ARMA, ARIMA)

Similarly, to linear regression frameworks, for instance, time series regressions can easily be used for regression-based imputations methods. Imputed values are then derived from a prediction based on time series regression instead of regression to an external variable. In particular, one can treat time series prediction as a problem of missing data where the missing data located in the future are predicted based on regression to preceding data (Sorjamaa et al. 2007).

Different time series regression methods can be distinguished depending on the assumptions they make regarding the autocorrelation between different observations of Y. The most crucial assumptions related to the number of previous observations of Y that are considered in computing the contemporary value of Y (the order of the autocorrelation) and whether the correlation between the actual value of Y and preceding values is constant or changes over time. It is beyond the scope of this paper to provide a detailed overview of these methods. Stock and Watson give a thorough treatment of time series methods (Stock et al. 2007).

Formally, there are different ways of specifying a stochastic process that generates time series where y_t and y_{t-j} are correlated over time, i.e., autocorrelation exists between different measures of y. One possible specification is an autoregressive process AR(p) of pth order with

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t. \tag{9}$$

In (9) epsilon is a random error term that follows a standard normal distribution and is independent over time with $E(\varepsilon_t, \varepsilon_{t-i}) = 0$ for all $i \neq t$. p denotes the number of lagged values of y_t that are considered. ε_t is an independent and identically distributed error term with zero mean and constant variance. Commonly used autoregressive (AR) models make the assumption that autocorrelation is constant over time and depends only on the intervals j between the y_t and $y_{t-j.}$

An alternative specification of a stochastic process that generates autocorrelation in a time series is moving average (MA) processes, in which the contemporary value of y_t is a function of its mean μ and a sequence of random innovations with

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-q}. \tag{10}$$

While y_t is not directly a function of previous values y_{t-q} in MA processes, autocorrelation between y_t and y_{t-q} is a consequence of the same random innovations ε_{t-q} entering the computation of different y_t.

In time series modeling, there is often an explicit recognition that time series models are merely intended to act as an approximation characterizing the dynamic behavior of the underlying series, the intention being to approximate autocorrelation structures over time (Adhikari and Agrawal 2013). Only in rare circumstances is it intended to provide a 'true' model of a time series. Instead, the focus is often to determine whether a time series model provides an approximation to observed behavior. While a 'true' model may take a large number of lagged terms to provide a proper fit with the specification in (9), it is often possible to fit an observed autoregressive (AR) time series more economically by combining it with a moving average (MA) component consisting of a sum of weighted lags of the error term ε_t (Box and Jenkins 1976). The resulting ARMA model is written as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q.} \tag{11}$$

Equation (11) is often referred to as an ARMA(p, q) model as it contains a pth-order autoregressive component in the observable time series, y_t, and a qth-order moving average component of the unobservable random shocks ε_t. It is generally assumed that ε_t follows a so-called white noise process with zero mean $E(\varepsilon_t)$ and constant variance $E(\varepsilon_t^2) = \sigma^2$. Moreover, it must to be noted that for Eq. (11) to be a tractable model that can be fitted to data, weak stationarity of the underlying time series' y_t is required. Weak stationarity is given if at least a time series' mean, variance and autocovariances are independent of t—whereas higher moments of the distribution of y_t over time might well depend on t. If $E(y_t) = \mu$, $E(y_t - \mu) = \sigma^2$ and $E[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j$, then a time series of y_t is said to be weakly stationary. Strict stationarity, on the other hand, would imply that a time series' distribution does not depend on t at all and hence $E(y_t) = \mu$ and $E(y_t - \mu) = \sigma^2$ and all higher moments are independent of t.

If a time series y_t is not stationary, stationarity can often be achieved by differencing the time series once or more (Box and Jenkins 1976). If differencing is required the ARMA (p, q) model (Autoregressive Moving Average) becomes an ARIMA (p, d, q) model (Autoregressive Integrated Moving Average), where d denotes the order of differencing, i.e., the number of times y_t is differenced to achieve stationarity.

When fitting ARIMA models, the choice of p, d and q can be guided by examining the autocorrelation and partial autocorrelation of y_t and ε_t over time (the latter measuring the correlation between y_t and y_{t-j} after accounting for the correlation between y_t and $y_{t-1}, y_{t-2},\ldots,y_{t-j+1}$). Stationarity is achieved and hence d is determined if autocorrelations between y_t and y_{t-j} become insignificant for increasing j. Moreover, an examination of the partial autocorrelation between y_t and y_{t-j} provides information about whether the order of the AR process p: p should be chosen as the number of lags for which the partial autocorrelation between y_t and y_{t-j} is still significant. In a similar way, the parameter q can be obtained by examining the full or partial autocorrelation of the error terms. A comprehensive procedure to choose the right parameters can be found in Box and Jenkins (1976).

ARMA and ARIMA models can easily be generalized to incorporate the influence of past, current or future values of exogenous factors (x variables) on the observed time series y_t. These approaches can be extended to ARMAX/ARIMAX by including exogenous variables (Feinberg and Genethliou 2005). Formally, they can be expressed as

$$\begin{aligned} y_t =& \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \beta_1 x_{t,1} + \beta_2 x_{t,2} \\ &+ \cdots + \beta_k x_{t,k} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \end{aligned} \tag{12}$$

where β_k denotes the effect of the exogenous variable x_k on the outcome variable y_t. Both ARMA/ARIMA and ARMAX/ARIMAX models can be readily estimated using common statistical software packages. The estimates obtained from fitted models can then be used as the basis for predictions employed to impute missing values, as described for the linear OLS (ordinary least squares) regression above.

Autoregressive conditional heteroscedasticity (ARCH) models

While ARMA/ARIMA models prove to be valid in many applications, the assumption of constant variance of the error terms $E(\varepsilon_t^2) = \sigma^2$ over time might, however, be too restrictive. In finance, for instance, periods of relatively stable stock markets might be followed by periods of crisis and turmoil (Baur and Lucey 2009), inducing a time-dependent autocorrelation of the error terms with $\mathrm{E}(\varepsilon_t^2) = \sigma_t^2$. In stable markets, autocorrelation might be relatively high (i.e., prices today will be similar to prices yesterday) and stock price movements are predictable (Fama and French 1988). In phases of turmoil, however, price movements might be bigger and autocorrelation is lower (Eom et al. 2004). In hydrology, the local climate might be characterized by a period of stable conditions followed by change in weather that drastically alters relevant outcomes (Hughes et al. 2011). In both examples, the assumption of constant autocorrelation might be too narrow. More realistic would be an assumption of changing variance and hence changing autocorrelation of the observed outcomes over time (heteroscedasticity).

The ARCH model is an extension of more restrictive AR models with constant autocorrelation of the outcome of interest (Zhu and Wang 2008). It is a non-linear regression model that captures not only past values of y_t but also time-varying volatility within the structure of standard time series models described above. While it is beyond this article to detail the mathematical underpinnings of Engle's work, it should be stressed that ARCH models are based on the assumption—while holding the unconditional variance of ε_t constant with $E(\varepsilon_t^2) = \sigma^2$—that its conditional variance could follow an AR process of its own with

$$\varepsilon^2 = \zeta + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \upsilon_t. \tag{13}$$

where υ_t is a white noise process. Based on this specification the ARCH model extends the standard ARMA/ARIMA model to incorporate time-varying volatility. While they require more additional assumptions (see Engle 1982 for technical details), ARCH models and their generalizations have proved useful for modeling flexible time series characterized by non-constant volatility.

Discussion

Regression-based imputation methods used in practical work are mainly based on linear regression approaches as they are well understood and easy to implement. In a hydrological setting, however, the assumption of the linear regression is well established but seemed to be too restrictive (Astel et al. 2004; Machiwal and Jha 2008). Imputation in hydrology often lacks a clear theoretical basis and a sound selection of methods depending on the statistical properties of the respective observable and the respective research question. The time series nature of hydrological data requires more flexible non-linear models such as the ARIMA and ARCH models, as shown above. It should be noted that there are multitude of alternative imputation methods based on non-linear regression approaches as well as non-probabilistic algorithms and machine learning approaches such as artificial neural networks (ANN), support vector machines (SVM) and classification and regression trees (CART) in hydrology. A good introduction to machine learning approaches can be found in Flach (2012). These approaches require very large training data sets as well as a sound expertise, e.g., in order to minimize the risk of overfitting. Thus, machine learning approaches cannot be recommended for unexperienced users. In addition, they are very likely to outperform the methods given in our review only for fairly long data gaps. Here we highlighted the potential role of time series methods as they explicitly model the particular statistical properties of hydrological time series (autocorrelation and heteroscedasticity) which are mostly neglected in algorithmic machine learning approaches.

Similarly, to ARMA/ARIMA models, ARCH models can easily be generalized and also allow the influence of past, current and future values of exogenous variables x_t to be modeled on the time series of interest. The estimation of ARCH is again possible relying on standard statistical software packages and predictions can be used to impute missing values in a time series. For the reasons stated above, when modeling the outcome variable of interest (y_t), time series models should focus on its variance and changes in variance over time. The increased importance of risk and uncertainty considerations in water resources management and hydrological modeling calls for new time series techniques that allow time-varying variances to be modeled, beyond the purpose of data imputation. For hydrologic nonstationary time series modeling e.g., rainfall in arid regions and hydrological time series through climate change conditions the testing of ARCH models with larger data bases is strongly recommended in the future (Modarres and Ouarda 2013).

Conclusion

Missing data is a common problem in hydrological data and poses a serious problem for many statistical and modeling approaches in hydrology. Therefore, researchers need to resort to imputation methods in order to replace missing values with approximations as these approaches require gap-free dataset. For reasons of convenience, researchers often resort to simple solutions to deal with missing data such as simply discarding observations characterized by missing data or by replacing missing data with a 'naïve' guess (such as the mean of all other observations). Despite their convenience, we have argued that these solutions have severe statistical shortcomings. Hydrological time series data are typically characterized by pronounced autocorrelation and seasonality. Making efficient use of these features could improve the performance of imputation methods considerably compared to widespread methods like mean-value imputation, etc. Even a relatively simple imputation algorithm that exploits the time series nature of the data—the preceding value approach—performs significantly better. More sophisticated approaches, like imputation methods based on principal component analysis (PCA) or regression, can improve the accuracy of missing value imputation and reduce statistical problems induced by naïve imputation approaches.

Autoregressive conditional heteroscedasticity (ARCH) models, which originate from finance and econometrics, propose an even better solution to the problem outlined above. Moreover, they can generate more accurate forecasts of future volatility and perform better than models that ignore heteroscedasticity. So ARCH models may not only be used for the imputation of missing observation in existing datasets but also to explain and characterize observed hydrological time series like precipitation, discharge and groundwater head fluctuations which are characterized by non-constant high variability. For this reason, they could be valuable for hydrological time series modeling in water resources management and flood control applications.

It must be stressed that there have been few studies concerning the imputation of missing data in the time series context in hydrology in general (e.g., Wang et al. 2005; Chen et al. 2008). Despite its particular focus on selected methods, our review shows that there are methodological advances that bear relevance for a more intensive use of these methods in hydrology.

References

Adhikari R, Agrawal R (2013) An introductory study on time series modeling and forecasting. arXiv:13026613

Allison PD (2000) Multiple imputation for missing data: a cautionary tale. Sociological methods research, vol 28. Sage Publications, pp 301–309

Allison PD (2001) Missing data, vol 136. Sage Publications, Philadelphia

Allison PD (2012) Handling missing data by maximum likelihood. SAS Global Forum Proceedings, pp 1–21

Astel A, Mazerski J, Polkowska Z, Namieśnik J (2004) Application of PCA and time series analysis in studies of precipitation in Tricity (Poland). Adv Environ Res 8:337–349

Aubin J, Bertrand-Krajewski J (2014) Analysis of continuous time series in urban hydrology: filling gaps and data reconstitution. Proceedings of the METMA VII and GRASPA14 conference. Torino (IT)

Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48:5–37

Baur DG, Lucey BM (2009) Flights and contagion—an empirical analysis of stock–bond correlations. J Financ Stab 5:339–352

Box GE, Jenkins GM (1976) Time series analysis, control, and forecasting, vol 3226. Holden Day, San Francisco, p 10

Chen CH, Liu CH, Su HC (2008) A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting. Hydrol Process 22:3697–3711

Cool AL (2000) A review of methods for dealing with missing data. Texas A&M University, College Station

Croninger RG, Douglas KM (2005) Missing data and institutional research. New Dir Inst Res 2005:33–49 de Leeuw J (1986) In: Proceedings of a workshop on multidimensional data analysis, Pembroke College, Cambridge University, England, 30 June–2 July 1985, vol 7. DSWO Press

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Donders ART, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59:1087–1091

Elshorbagy A, Simonovic S, Panu U (2002) Estimation of missing streamflow data using principles of chaos theory. J Hydrol 255:123–133

Enders CK (2010) Applied missing data analysis. Guilford Press, New York

Engle RF (1982) Autoregressive conditional heteroscedastisity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1008

Eom KS, Hahn SB, Joo S (2004) Partial price adjustment and autocorrelation in foreign exchange markets. University of California, Berkeley

Fama EF, French KR (1988) Permanent and temporary components of stock prices. J Polit Econ 96:246–273

Farhangfar A, Kurgan L, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Patt Recogn 41:3692–3705

Feinberg EA, Genethliou D (2005) Load forecasting. In: Chow JH, Wu FF, Momoh JA (eds) Applied mathematics for restructured electric power systems. Springer, US, pp 269–285

Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge

Frane JW (1976) Some simple procedures for handling missing data in multivariate analysis. Psychometrika 41:409–415

Gill MK, Asefa T, Kaheil Y, McKee M (2007) Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. Water Resour Res. https://doi.org/10.1029/2006WR005298

Graham JW (2009) Missing data analysis: making it work in the real world. Annu Rev Psychol 60:549–576

Graham JW, Hofer SM (2000) Multiple imputation in multivariate research. In: Little TD, Schnabel KU, Baumert J (eds) Modeling longitudinal and multiple group data: practical issues, applied approaches, and specific examples. Lawrence Erlbaum Associates, Mahwah, NJ, pp 201–218

Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142:1255–1264

Guzman JA, Moriasi D, Chu M, Starks P, Steiner J, Gowda P (2013) A tool for mapping and spatio-temporal analysis of hydrological data. Environ Model Softw 48:163–170

Harrington D (2008) Confirmatory factor analysis. Oxford University Press, USA

Hassani H (2007) Singular spectrum analysis: methodology and comparison. J Data Sci 5(2):239–257

Hawkins M, Merriam V (1991) An overmodeled world. Direct Mark, pp 21–24

Hedeker D, Gibbons RD (1997) Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychol Methods 2:64

Henn B, Raleigh MS, Fisher A, Lundquist JD (2013) A comparison of methods for filling gaps in hourly near-surface air temperature data. Gloss Meteorol AMS. https://doi.org/10.1175/JHM-D-12-027.1

Hughes CE, Cendón DI, Johansen MP, Meredith KT (2011) Climate change and groundwater. In: Anthony J, Jones A (eds) Sustaining groundwater resources. Springer, pp 97–117

Johnston CA (1999) Development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data. Virginia Tech, Master thesis [17479]

Jolliffe IT (1993) Principal component analysis: a beginner's guide—II. Pitfalls Myths Ext Weather 48:246–253

Jolliffe I (2002) Principal component analysis. Wiley Online Library, New York

Kiers HA (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 62:251–266

Kim J-O, Curry J (1977) The treatment of missing data in multivariate analysis. Sociol Methods Res 6:215–240

Kim J, Ryu JH (2016) A heuristic gap filling method for daily precipitation series. Water Resour Manage 30:2275–2294

King G, Honaker J, Joseph A, Scheve K (1998) List-wise deletion is evil: what to do about missing data in political science. In: Annual meeting of the American political science association, Boston

Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Process Geophys 13:151–159

Lee KJ, Carlin JB (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 171:624–632

Little RJA (1988) Missing-data adjustments in large surveys. J Bus Econ Stat 6:287–296

Little R, Rubin D (1987) Analysis with missing data. Wiley, New York

Machiwal D, Jha M (2008) Comparative evaluation of statistical tests for time series analysis: application to hydrological time series. Hydrol Sci J 53:353–366

Malhotra NK (1987) Analyzing marketing research data with incomplete information on the dependent variable. J Mark Res 24:74–84

Marsh HW (1998) Pairwise deletion for missing data in structural equation models: nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Struct Equ Model Multidiscip J 5:22–36

Mcdonald RA, Thurston PW, Nelson MR (2000) A Monte Carlo study of missing item methods. Organ Res Methods 3:71–92

McKnight PE, McKnight KM, Sidani S, Figueredo AJ (2007) Missing data: a gentle introduction. Guilford Press, New York

Modarres R, Ouarda T (2013) Generalized autoregressive conditional heteroscedasticity modelling of hydrologic time series. Hydrol Process 27(22):3174–3191

Pandey PK, Singh Y, Tripathi S (2011) Image processing using principle component analysis. Int J Comput Appl 15(4):37–40

Peugh JL, Enders CK (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res 74:525–556

Pigott TD (2001) A review of methods for missing data. Educ Res Eval 7:353–383

Puma MJ, Olsen RB, Bell SH, Price C (2009) What to do when data are missing in group randomized controlled trials. NCEE 2009-0049. National Center for Education Evaluation and Regional Assistance

Raaijmakers QA (1999) Effectiveness of different missing data treatments in surveys with Likert-type data: introducing the relative mean substitution approach. Educ Psychol Measur 59:725–748

Raghunathan TE (2004) What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 25:99–117

Roth PL (1994) Missing data: a conceptual review for applied psychologists. Pers Psychol 47:537–560

Roth PL, Switzer FS, Switzer DM (1999) Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques. Organ Res Methods 2:211–232

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Rubin DB (2004) Multiple imputation for nonresponse in surveys, vol 81. Wiley, New York

Rubin DB, Little RJ (2002) Statistical analysis with missing data. Wiley, Hoboken

Rubin LH, Witkiewitz K, St Andre J, Reilly S (2007) Methods for handling missing data in the behavioral neurosciences: do not throw the baby rat out with the bath water. J Undergrad Neurosci Educ 5:71–77

Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R (2006) Imputing missing data: a comparison of methods for social work researchers. Soc Work Res 30:19–31

Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7:147

Soley-Bori M (2013) Dealing with missing data: key assumptions and methods for applied analysis (No. 4). Technical report, Boston University

Sorjamaa A, Hao J, Reyhani N, Ji Y, Lendasse A (2007) Methodology for long-term prediction of time series. Neurocomputing 70:2861–2869

Stock JH, Watson MW, Addison-Wesley P (2007) Introduction to econometrics. Addison and Wesley, Boston

Tannenbaum CE (2009) The empirical nature and statistical treatment of missing data. University of Pennsylvania, ProQuest Dissertations Publishing, Philadelphia

Tsikriktsis N (2005) A review of techniques for treating missing data in OM survey research. J Oper Manag 24:53–62

van der Heijden GJ, Donders ART, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59:1102–1109

Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M (eds) A practical approach to microarray data analysis. Springer, pp 91–109

Wang W, Vrijling JK, Van Gelder PH, Ma J (2005) Testing and modeling autoregressive conditional heteroskedasticity of streamflow processes. Nonlinear Process Geophys 12:55–66

Wothke W (2000) Longitudinal and multigroup modeling with missing data. In: Little TD, Schnabel KU, Baumert J (eds) Modeling longitudinal and multilevel data. Erlbaum, Mahwah, pp 219–240

Zhang Q, Wang B-D, He B, Peng Y, Ren M-L (2011) Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. Water Resour Manage 25:2683–2703

Zhu F, Wang D (2008) Local estimation in AR models with nonparametric ARCH errors. Commun Stat Theory Methods 37:1591–1609