Atmospheric Environment 102 (2015) 96–104

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Imputation of missing data in time series for air pollutants

W.L. Junger*, [1], A. Ponce de Leon

Rio de Janeiro State University, Department of Epidemiology, Brazil

HIGHLIGHTS

- We propose a method for imputation of missing values in times series.
- Simulations showed adequate goodness-of-fit.
- The findings also suggest good accuracy and precision.
- We implemented the method as an open source R library.

ARTICLE INFO

Article history:
Received 10 June 2014
Received in revised form
13 November 2014
Accepted 21 November 2014
Available online 24 November 2014

Keywords:
Air pollution
Data imputation
EM algorithm
Environmental epidemiology
Missing data
Particulate matter
Time series

ABSTRACT

Missing data are major concerns in epidemiological studies of the health effects of environmental air pollutants. This article presents an imputation-based method that is suitable for multivariate time series data, which uses the EM algorithm under the assumption of normal distribution. Different approaches are considered for filtering the temporal component. A simulation study was performed to assess validity and performance of proposed method in comparison with some frequently used methods. Simulations showed that when the amount of missing data was as low as 5%, the complete data analysis yielded satisfactory results regardless of the generating mechanism of the missing data, whereas the validity began to degenerate when the proportion of missing values exceeded 10%. The proposed imputation method exhibited good accuracy and precision in different settings with respect to the patterns of missing observations. Most of the imputations obtained valid results, even under missing not at random. The methods proposed in this study are implemented as a package called *mtsdi* for the statistical software system *R*.

1. Introduction

Missing data are a major concern in epidemiological studies (Eekhout et al., 2012), especially of the health effects of environmental air pollutants, which are often caused by equipment failure or data corruption. Under the Bayesian framework, missing data are extra parameters to be estimated (Gelman et al., 2014) but it is not as trivial otherwise. The analysis of incomplete data has been studied widely and many methods have been developed (Schafer, 1997; Little and Rubin, 1989; Little, 1992; Dempster et al., 1977; Rubin, 1976; Beale and Little, 1975; Hartley and Hocking, 1971), but it has received little attention in epidemiological contexts (Miettinen, 1985; Rothman et al., 2008). However, several studies have addressed the impact of incomplete data analysis on epidemiological measures using stochastic simulations (Gorelick, 2006; Plaia and Bondì, 2006; Junninen et al., 2004; Engels and Diehr, 2003) and new methods have been proposed. Simple methods are likely to yield biased estimates and the most sophisticated depend on strong assumptions about the sources of the missing data, while they also involve complex computations (Little and Rubin, 1989; Schafer, 1997).

Rubin (1976) classified incomplete data according to their generating mechanisms. Data can be missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR). The MCAR condition is too restrictive because it assumes that the missing data comprise a random sample taken from the observed values. In epidemiological research, the distributions of missing data are often related to the disease status or the exposure. Thus, the MAR assumption may be more realistic (Greenland and Finkle, 1995).

* Corresponding author.
E-mail addresses: wjunger@ims.uerj.br (W.L. Junger), ponce@ims.uerj.br (A. Ponce de Leon).
[1] This author was sponsored by the Brazilian Research Council – CNPq.

http://dx.doi.org/10.1016/j.atmosenv.2014.11.049
1352-2310/

The default method used by most statistical software is complete case analysis, i.e., the exclusion of incomplete observations. Under the MCAR assumption, this may yield unbiased estimates. However, a higher proportion of incomplete observations may result in a loss of precision (Rothman et al., 2008; Greenland and Finkle, 1995). Under MAR, the complete case analysis no longer relies on a random sample of the source population and selection bias is likely to occur (Donders et al., 2006). In time series analysis, this problem can be exacerbated because excluding incomplete observations may corrupt temporal structures such as autocorrelation, trends, and seasonality (Box et al., 1994).

The approaches used to estimate parameters in a missing data situation can be classified into two major groups: likelihood-based and imputation-based (Little and Rubin, 1989). Likelihood-based methods are flexible, they do not require *ad hoc* methods, and they yield an adequate estimate of the variance, but it may be necessary to solve highly complex likelihood equations (McLachlan and Krishnan, 1997; Little and Rubin, 1989; Hartley and Hocking, 1971). These methods generally use computational routines that are tailored for specific analyses; thus it is difficult to make them readily available in general purpose statistical software systems. By contrast, imputation-based methods are usually simpler, and they can be implemented in most commercial statistical software systems. Some methods may be computationally intensive, e.g., multiple imputation (Schafer, 1997).

Despite the simplicity, imputation-based methods do have some drawbacks. The data analysis takes place post-imputation; thus, the extra variability due to imputation is usually not considered, so the variance of the estimated association is usually underestimated. Another important characteristic of imputation-based methods is that the simplest types often yield biased estimates of the association (Donders et al., 2006). Multiple imputation estimation may consider the extra variability, thereby obtaining more precise confidence intervals (Schafer, 1997) but these methods are not addressed in the current study.

The simplest and the often misused method is replacing the missing values with the unconditional mean (UM) of the variable. Under MAR, this yields inconsistent estimates of the variance of the regression coefficients. If the MCAR assumption holds, the variance estimates are consistent but underestimated. Thus, hypothesis testing and the estimates of the confidence intervals will be distorted by both the bias and the overestimated precision (Little and Rubin, 1989; Little, 1992). Imputation using the median (MD) may yield better results for skewed distributions (Miettinen, 1985). Single imputation based on unconditional or conditional means tends to distort the marginal distribution of the data due to the higher concentration of observations around the mean. This may be a major concern if one is interested in the tails of the distribution, e.g., hypothesis testing (Little and Rubin, 1989).

The latter method can be improved by using the information from measured covariates of the same study unit to impute the missing value based on the prediction from a linear regression model. The regression coefficients are estimated using the complete case analysis. Under the MCAR assumption, this yields consistent estimates of the association (Little, 1992).

Hartley and Hocking (1971) proposed a set of iterative equations for likelihood estimation of the mean vector and the covariance matrix of a multivariate normal distribution with missing data. Later, this method was extended to accommodate any distribution from the exponential family, which is now referred to as the expectation–maximization (EM) algorithm (Dempster et al., 1977). Under the assumptions of a multivariate normal distribution, the EM algorithm is an iterative version of Buck's calculator (McLachlan and Krishnan, 1997; Buck, 1960).

In this article, we present an imputation-based method that is suitable for multivariate time series data, using the EM algorithm for estimation of the mean vector and covariance matrix of the normal distribution for the underlying framework. In addition to the correlations among covariates, the algorithm also considers the temporal components of the time series. Different approaches are implemented for filtering the temporal components. Simulations were performed to assess the procedure's validity and comparisons were made with some frequently used methods. Our method has been implemented as a package called *mtsdi* (multivariate time series data imputation) for the statistical software *R* (R Core Team, 2013), which is available at R repositories.

This article is organized as follows: in Section 2 the statistical concepts of the imputation, filtering models, and simulation details are presented; in Section 3 results from the validity and performance analyses as well as the penalization are showed; the discussion on the findings is presented in Section 4.

2. Methods

2.1. Imputation procedure

Let $\mathbf{x}_t$, $(t = 1, \ldots, n)$, be the t-th realization of the p-variate normal random vector $\mathbf{X}$ with m unobserved components. The vector $\mathbf{x}_t$ can be rearranged such that the m missing elements will be in the first positions, i.e., $\mathbf{x}_t = (x_{t1}, \ldots, x_{tm}, x_{t(m+1)}, \ldots, x_{tp})^T$, which are denoted by $\mathbf{x}_t = (\mathbf{x}_{t1}, \mathbf{x}_{t2})^T$. Furthermore, we consider that the observed period can be spanned over B time windows, with indices b, $(b = 1, \ldots, B)$, and each time window has a different underlying regime of covariance over time. Thus, the mean vector at time t and the covariance matrix at window b can be partitioned by following the same configuration of $\mathbf{x}_t$, i.e.,

$$\tilde{\mu}_t = \begin{bmatrix} \tilde{\mu}_{t1} \\ \tilde{\mu}_{t2} \end{bmatrix} \text{ and } \tilde{\mathbf{\Sigma}}_b = \begin{bmatrix} \tilde{\mathbf{\Sigma}}_{b11} & \tilde{\mathbf{\Sigma}}_{b12} \\ \tilde{\mathbf{\Sigma}}_{b21} & \tilde{\mathbf{\Sigma}}_{b22} \end{bmatrix}.$$

Our proposed imputation method is a modification of the EM algorithm for estimating the mean vector and the covariance matrix of a multivariate normal distribution with missing data (Dempster et al., 1977). The algorithm comprises the following steps: (i) replace the missing values by estimates; (ii) estimate the parameters μ and $\mathbf{\Sigma}$; (iii) estimate the level for each of the univariate time series; (iv) re-estimate the missing values using updated estimates of the parameters and the level of the time series. These steps are iterated until some convergence criterion is reached.

In general, the initial estimates $\tilde{\mu}_0$ and $\tilde{\mathbf{\Sigma}}_0$ are the mean vector and the covariance matrix estimated from the observed incomplete data. At the $(k+1)$-th iteration of the E (estimation) step of the EM algorithm, the missing values are imputed with conditional means given the observed values and the previous estimates of the parameters given by

$$\tilde{\mathbf{x}}_{t\mathbf{1}}^{(k+1)} = E\left[\mathbf{X}_{t1} \middle| \mathbf{x}_{t\mathbf{2}},\ \tilde{\mu}_t^{(k)},\ \tilde{\mathbf{\Sigma}}_b^{(k)}\right] = \tilde{\mu}_{t\mathbf{1}}^{(k)} + \tilde{\mathbf{\Sigma}}_{b\mathbf{12}}^{(k)} \tilde{\mathbf{\Sigma}}_{b\mathbf{22}}^{(k)-1} \left(\mathbf{x}_{t\mathbf{2}} - \tilde{\mu}_{t\mathbf{2}}^{(k)}\right).$$

The contributions to the covariance matrix are given by

$$\begin{aligned} \widetilde{\mathbf{x}_{t1}\mathbf{x}_{t1}^T}^{(k+1)} &= E\left[\mathbf{X}_{t1}\mathbf{X}_{t1}^T \middle| \mathbf{x}_{t2},\ \tilde{\mu}_t^{(k)},\ \tilde{\mathbf{\Sigma}}_b^{(k)}\right] \\ &= \tilde{\mathbf{\Sigma}}_{b11}^{(k)} - \tilde{\mathbf{\Sigma}}_{b12}^{(k)} \tilde{\mathbf{\Sigma}}_{b22}^{(k)-1} \tilde{\mathbf{\Sigma}}_{b21}^{(k)} + \tilde{\mathbf{x}}_{t1}\tilde{\mathbf{x}}_{t1}^T \end{aligned}$$