Stat Methods Appl (2019) 28:359–383
https://doi.org/10.1007/s10260-018-00435-9

ORIGINAL PAPER

Reconstructing missing data sequences in multivariate time series: an application to environmental data

Maria Lucia Parrella[1] · **Giuseppina Albano**[1] ·
Michele La Rocca[1] · **Cira Perna**[1]

Accepted: 5 August 2018 / Published online: 18 August 2018

Abstract Missing data arise in many statistical analyses, due to faults in data acquisition, and can have a significant effect on the conclusions that can be drawn from the data. In environmental data, for example, a standard approach usually adopted by the Environmental Protection Agencies to handle missing values is by deleting those observations with incomplete information from the study, obtaining a massive underestimation of many indexes usually used for evaluating air quality. In multivariate time series, moreover, it may happen that not only isolated values but also long sequences of some of the time series' components may miss. In such cases, it is quite impossible to reconstruct the missing sequences basing on the serial dependence structure alone. In this work, we propose a new procedure that aims to reconstruct the missing sequences by exploiting the spatial correlation and the serial correlation of the multivariate time series, simultaneously. The proposed procedure is based on a spatial-dynamic model and imputes the missing values in the time series basing on a linear combination of the *neighbor* contemporary observations and their lagged values. It is specifically oriented to spatio-temporal data, although it is general enough to be applied to generic stationary multivariate time-series. In this paper, the procedure has been applied to the pollution data, where the problem of missing sequences is of serious concern, with remarkably satisfactory performance.

Keywords Spatial correlation · Missing values · PM_{10} data · Time series

✉ Giuseppina Albano
pialbano@unisa.it

[1] Dip. di Scienze Economiche e Statistiche, Università of Salerno, Fisciano (Salerno), Italy

1 Introduction

In the last few decades, a considerable number of epidemiological studies establishes the link between air pollution and health (see, for example, Biggeri et al. 2002; Raaschou-Nielsen et al. 2013), recognising air quality to be a fundamental issue for human health. Among the pollutants, PM_{10} refers to the particles with diameters up to $10\,\mu\text{m}$ due to the emission produced by motor vehicles, industrial activities and other natural sources. PM_{10} particles are so small that they can get into the lungs, potentially causing severe health problems (see Aga et al. 2003), confirming the acute, adverse effects of PM_{10} on mortality. In 2013 the International Agency for Research on Cancer and the World Health Organization designated airborne particulates a Group 1 carcinogen. In response, the European Union has developed an extensive body of legislation which establishes health-based standards and objectives for some pollutants in the air. For example, in the air quality directive (2008/EC/50), the EU has set two limit values for particulate matter (PM_{10}) for the protection of human health: the PM_{10} daily mean value may not exceed 50 micrograms per cubic metre ($\mu\text{g}/\text{m}^3$) more than 35 times in a year and the PM_{10} annual mean value may not exceed 40 micrograms per cubic metre ($\mu\text{g}/\text{m}^3$). In this context, missing data, usually due to equipment failures or to errors in measurements, represent the main problem to count the number of excesses of the fixed limit and monitoring the air quality. The most common approach usually adopted by the Environmental Protection Agencies to handle missing values is by deleting those observations from the study, obtaining a massive underestimation of the number of excesses of the admitted limit for PM_{10}. In the literature, various techniques have been proposed to impute missing values in environmental data (see, for example, Norazian et al. 2008; Junninen et al. 2004). However, these methods perform well only when the number of missing values is small (see, for example, Fitri et al. 2010). Moreover, many models have been proposed in literature taking into account the spatio-temporal dependence of PM_{10} located in near places (see, for example, Cameletti et al. 2011). In particular, Cameletti et al. (2011) compare different hierarchical spatio-temporal models differing from each other in how they model the residual detrended process and in how the spatio-temporal correlation is treated. Regardless, the authors explicitly observe that in the dataset they used for their comparisons the percentage of missing data is less than 20% and they are not sequential.

In the more general contexts of multivariate time series, characterised by both serial and cross-correlation, ignoring the missing values can lead to bias and error during data mining. Nevertheless, it happens very frequently that the only procedures implemented in the multivariate time series analysis to face missing values are a list-wise deletion, single imputation (mean imputation, k-Nearest Neighbour imputation) or multiple imputations (multivariate normal, multivariate imputation by chained equations). Although all of them do not account for the fact that variables are correlated and that those relationships exist across time. Recently, some new approaches have been proposed to fill this gap with clear evidence of substantial advantages compared with the standard methods described above (see, in particular, Pollice and Lasinio 2009; Liu and Molenaar 2014; Oehmcke et al. 2016 and Sect. 1.1 for a description of such proposals). Mainly, the new approaches combine two different imputation methods in separate stages (for example, k-NN and Fourier transform), so that the first one accounts for cross-correlation among variables and the second one deals with serial correlation in univariate time series. Recently, in Calculli et al. (2015) the authors proposed a multivariate hidden dynamic geostatistical model and maximum likelihood parameter estimate using EM algorithm, able to handle with multiple variables sampled at different monitoring networks and missing data, were provided.

In this paper we propose a new procedure for estimating (even long) missing sequences in time series, focusing on an application to PM_{10} data. Our method uses an approach based on the *generalized spatial-dynamic autoregressive model*. This model was first proposed in Dou et al. (2016) and belongs to the family of *spatial econometric models* (see Lee and Yu 2010b for an introduction and a survey of such kinds of models). These models include, in the form of a weighted multivariate autoregression, the distances among the considered locations (i.e., among the monitoring stations). In this way, we can take into account spatial correlation in the data and estimate missing sequences in PM_{10} for a station by looking at the values of PM_{10} in the near stations, but also by looking at the previous lags of the same station and the neighbour stations.

The advantages of our imputation procedure compared with the alternative approaches described in Sect. 1.1 are several. First of all, it takes into account both the serial correlation and the spatial correlation simultaneously, in a single stage. Secondly, it does not depend on any tuning parameter (like, for example, the value k of the Nearest Neighbor method) or user choice (like, for instance, the selection of the method for the combination of results from multiple imputation methods). Moreover, our approach is computationally feasible and scales up to high dimension of the multivariate time series, so that it can also be applied in those cases where the length of the time series (= number of observations) is smaller than the dimension (= number of variables/locations), contrary to what happens, for example, in Liu and Molenaar (2014). The simulation study, presented in Sect. 4.2, shows that our imputation procedure has a good performance also when the percentage of missing values (or the length of missing sequences) is very high, even for moderate sample sizes. Last but not least, our imputation procedure also works in those cases where the whole line (i.e., all observations at time t) or the entire column (i.e., all observations of location i) of the dataset are missing. This latter is a very interesting feature almost never assured by other imputation methods.

The paper is organised as follows. In the following of this section, we briefly describe the imputation techniques proposed in the literature for multivariate time series to introduce our competitor. Section 2 describes the proposed imputation method, its underlying model and the estimation procedure. Moreover, the theoretical justifications for our proposal are presented in Sect. 2.4 and stated in Theorem 1, whose proof is reported in the Appendix. In Sect. 3 an application to environmental data is presented and discussed. In Sect. 4 a simulation experiment is performed to validate the method. Some conclusions close the paper.

1.1 Literature review for multivariate time series imputation

There are several methods proposed in the literature for the imputation of missing values, and many of them are also implemented in R software packages. However, very few approaches are suitable for multivariate time series.

In the context of PM_{10} concentration, in Pollice and Lasinio (2009) an imputation technique based on linear spatial regression is used, where no spatial correlation structure is assumed for the imputed data, all the stations being considered equivalent in the linear predictor, presenting some problems in the case of massive datasets or high dimensionality.

In Oehmcke et al. (2016) an algorithm was proposed to create an ensemble of models based on weighted k-Nearest Neighbours imputation with Dynamic Time Warping as distance measurement with a linear interpolation preprocessing step. This ensemble creates training problems with multivariate data that the authors solve with a preprocessing step, normalising distances, applying correlation weighting and penalising gaps, penalising this preprocessing whenever values are consecutively missing. The ensemble ensures diversity among individual models with diversity methods such as different penalty strength. Apparently, in this way, the serial and spatial correlations are considered in two different steps, dealing with some inefficiencies with a not computationally feasible algorithm.

Finally, in Liu and Molenaar (2014) the authors propose an imputation method based on vector autoregressive models, called iVAR. They show through a simulation study that their method produces better estimates compared to the standard approaches, thanks to the autocorrelation structure captured by VAR. However, they also highlight several limitations of the procedure. Above all, the difficulty to handle large datasets, since VAR models are affected by the curse of dimensionality problem.

The most popular R packages available on the CRAN are, among others, `AmeliaII` (Honaker et al. 2011), `mice` (Buuren and Groothuis-Oudshoorn 2011), `VIM` (Kowarik and Templ 2016), `missMDA` (Josse and Husson 2016) and `imputeTS` (Moritz et al. 2017). Among these, the only ones that give direct support for longitudinal data are the packages `imputeTS` (Moritz et al. 2017) and `AmeliaII` (Honaker et al. 2011).

The `imputeTS` package can handle with univariate time series imputation and include multiple imputation algorithms. There are also other packages that deal with univariate time series imputation (see Moritz et al. 2017 for a list). However, the common feature of all these packages is that they only rely on univariate time dependencies to impute the missing values. As a consequence, they cannot reconstruct long sequences of missing values and, therefore, cannot be considered as our competitors.

The `AmeliaII` package is designed to impute cross-sectional data and it also considers the case of longitudinal data. The imputation model in Amelia assumes that the complete data (that is, both observed and unobserved) are multivariate normal. It draws imputations of the missing values using a bootstrapping approach, the EMB (expectation-maximization with bootstrapping) algorithm. To deal with time series, `AmeliaII` builds a general model of patterns within variables across time by creating a sequence of polynomials of the time index, up to the user-defined k-th order, ($k \leq 3$). If cross-sectional units are specified, these polynomials can be interacted with the cross-section unit to allow the patterns over time to vary between cross-sectional units. Moreover, to improve multivariate time-series imputation, AmeliaII can also include lags and leads of specific variables into the imputation model. So, it can also reconstruct sequences of missing values and, therefore, it can be considered as a competitor of our procedure. We will use the AmeliaII method in the application and the simulation study for a comparison with our approach.

2 The proposed method for the imputation of missing values and missing sequences

2.1 The model

Our imputation method is model based. We use a spatio-temporal model for PM_{10} time series, intended to manage a network of near monitoring stations.

Let us consider a multivariate stationary process $\{\mathbf{y}_t\}$ of order p, where the vector $\mathbf{y}_t$ collects the observations at time t from p different locations (=stations). The *generalized Spatial Dynamic Panel Data* model (*G-SDPD*) first introduced in Dou et al. (2016) is

$$\mathbf{y}_t = \boldsymbol{\mu} + D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t + D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1} + D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad (1)$$

where $\boldsymbol{\mu}$ is a constant vector (connected to the mean value of the process) and the errors $\boldsymbol{\varepsilon}_t$ are serially uncorrelated, have zero mean value and may show cross-sectional correlation and heteroscedasticity, so they generally have full variance/covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, for all t. The matrices $D(\boldsymbol{\lambda}_j)$ are diagonal, for $j = 0, 1, 2$, and the vectors $\boldsymbol{\lambda}_j$ collects the parameters λ_{ji} for $i = 1, \ldots, p$ and $j = 0, 1, 2$.

Model (1) belongs to the category of *econometric spatial models* born by the seminal idea of Anselin (1988). They are based on the presence of a *spatial matrix* $\mathbf{W}$ which measures the spatial correlation among the p components of the multivariate time series. This matrix has zero main diagonal and it is generally assumed known. In the fundamental idea of econometric spatial modelling, the matrix $\mathbf{W}$ was related to "proximity in space" between the single components of the time series, and therefore it was derived by the inverse of some physical distance among locations. However, the concept of space may also be intended in a wider sense and it can be measured by some association or correlation distance among the components of the multivariate time series. In our real data application, we will consider both the approaches.

The *G-SDPD* model in (1) is characterized by the sum of three terms (overlooking the intercept for brevity) : (i) the *spatial component*, driven by matrix $\mathbf{W}$ and the vector parameter $\boldsymbol{\lambda}_0$, which captures spatial effects in the data; (ii) the *dynamic component*, driven by the vector parameter $\boldsymbol{\lambda}_1$, which takes into account the dependence on past observations; and (iii) the *spatial–dynamic component*, driven by $\mathbf{W}$ and $\boldsymbol{\lambda}_2$, which considers spatial effects from past observations. For the estimation of the parameters $\boldsymbol{\lambda}_j$, in Dou et al. (2016) the authors propose an innovative procedure based on the combination of the least squares' method and the Yule–Walker approach. The appeal of this procedure is that the estimators of the $\boldsymbol{\lambda}_j$ coefficients can be written in closed form, contrary to most of the alternative estimators proposed in the spatial econometrics literature (see Lee and Yu 2010b and references therein). This novel approach allows the estimation procedure to be fast and easy to implement. The details will be shown in Sect. 2.2.

2.2 Estimation of model parameters

In the sequel, we assume that $\mathbf{y}_1, \ldots, \mathbf{y}_T$ are realizations from the stationary process defined by (1). A natural estimator for the mean is $\bar{\mathbf{y}} = T^{-1} \sum_{1 \le t \le T} \mathbf{y}_t$. At this stage we assume $\boldsymbol{\mu} = 0$ in (1). In practice, we should centre the data first, i.e. replace $\mathbf{y}_t$ by $\mathbf{y}_t - \bar{\mathbf{y}}$.

Having a process with mean zero, we denote with $\boldsymbol{\Sigma}_j = Cov(\mathbf{y}_t, \mathbf{y}_{t-j}) = E(\mathbf{y}_t \mathbf{y}'_{t-j})$ the autocovariance matrix of the process at lag j, where the prime subscript denotes the transpose operator.

The parameters of model (1) can be estimated following Dou et al. (2016). In particular, assuming that the process is stationary, from (1) we derive the Yule–Walker equation system

$$(\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W})\boldsymbol{\Sigma}_1 = (D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W})\boldsymbol{\Sigma}_0,$$

where $\mathbf{I}$ is the identity matrix of order p. The i-th row of the above multivariate equation system is

$$(\mathbf{e}'_i - \lambda_{0i}\mathbf{w}'_i)\boldsymbol{\Sigma}_1 = (\lambda_{1i}\mathbf{e}'_i + \lambda_{2i}\mathbf{w}'_i)\boldsymbol{\Sigma}_0, \quad i = 1, \ldots, p, \tag{2}$$

where $\mathbf{w}_i$ is the i-th row vector of $\mathbf{W}$ and $\mathbf{e}_i$ is the unit vector with the i-th element equal to 1. Note that (2), for a given i, is a system of p linear equations with three unknown parameters, λ_{0i}, λ_{1i} and λ_{2i}. Replacing $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_0$ by the sample (auto)covariance matrices

$$\widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{T}\sum_{t=1}^{T-1} \mathbf{y}_{t+1}\mathbf{y}'_t \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t\mathbf{y}'_t,$$

the vector $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})'$ is estimated by the least squares method, i.e. to solve the minimization problem

$$\min_{\lambda_{0i}, \lambda_{1i}, \lambda_{2i}} \|\widehat{\boldsymbol{\Sigma}}'_1(\mathbf{e}_i - \lambda_{0i}\mathbf{w}_i) - \widehat{\boldsymbol{\Sigma}}_0(\lambda_{1i}\mathbf{e}_i + \lambda_{2i}\mathbf{w}_i)\|_2^2,$$

where $\| \cdot \|_2$ denotes the Euclidean norm. The resulting generalized Yule–Walker estimator can be written in closed form:

$$(\widehat{\lambda}_{0i}, \widehat{\lambda}_{1i}, \widehat{\lambda}_{2i})' = (\widehat{\mathbf{X}}'_i\widehat{\mathbf{X}}_i)^{-1}\widehat{\mathbf{X}}'_i\widehat{\mathbf{Y}}_i, \quad i = 1, 2, \ldots, p, \tag{3}$$

where $\widehat{\mathbf{X}}_i = \left(\widehat{\boldsymbol{\Sigma}}'_1\mathbf{w}_i, \widehat{\boldsymbol{\Sigma}}_0\mathbf{e}_i, \widehat{\boldsymbol{\Sigma}}_0\mathbf{w}_i\right)$ and $\widehat{\mathbf{Y}}_i = \widehat{\boldsymbol{\Sigma}}'_1\mathbf{e}_i$.

2.3 The novel iterative imputation procedure

We assume that the time series has (sequences of) missing values at some locations and the goal of this section is to propose an imputation procedure for such values. We also assume that missing values happen at random, in the sense that they occur independently on their (unobserved) levels and independently on the other values in the multivariate time series.

The proposed method consists of the prediction of the missing values basing on a linear interpolation of the *neighbor* contemporary observations and their lagged values. In such a way, we simultaneously take into account the cross-correlation and the serial correlation among the observations. The weights of the interpolation are given by the strength of the spatial correlation among locations. Then, the procedure runs in recursive steps of substitution/estimation of missing values, until some form of convergence is reached.

In this section, we assume that the observed multivariate time series is a realisation of a process as in (1), with possibly non-zero intercept $\boldsymbol{\mu}$.

Let $\boldsymbol{\delta}_t = (\delta_{t1}, \ldots, \delta_{tp})$ be a vector of zeroes/ones that identifies all the missing values in the observed vector $\mathbf{y}_t$, so that $\delta_{ti} = 0$ if the observation y_{ti} is missing, otherwise it is $\delta_{ti} = 1$.

The procedure starts, at iteration 0, by initializing the mean centered vector $\widetilde{\mathbf{y}}_t^{(0)}$ as

$$\widetilde{\mathbf{y}}_t^{(0)} = \boldsymbol{\delta}_t \circ \left(\mathbf{y}_t - \bar{\mathbf{y}}^{(0)}\right), \qquad t = 1, \ldots, T, \text{ with } \bar{\mathbf{y}}^{(0)} = \frac{\sum_{t=1}^{T} \boldsymbol{\delta}_t \circ \mathbf{y}_t}{\sum_{t=1}^{T} \boldsymbol{\delta}_t}, \tag{4}$$

where the operator $\circ$ denotes the Hadamard product (which substantially implies replacing the missing values with zero) and the ratio between the two vectors in the formula of $\bar{\mathbf{y}}^{(0)}$ is made componentwise.

Then, the generic iteration s of the procedure, with $s \geq 1$, requires that:

(a) we estimate $(\widehat{\boldsymbol{\lambda}}_0^{(s-1)}, \widehat{\boldsymbol{\lambda}}_1^{(s-1)}, \widehat{\boldsymbol{\lambda}}_2^{(s-1)})$ as explained in Sect. 2.2, but using the centered data $\{\widetilde{\mathbf{y}}_1^{(s-1)}, \ldots, \widetilde{\mathbf{y}}_T^{(s-1)}\}$;

(b) we compute, for $t = 1, \ldots, T$,

$$\begin{aligned} \widehat{\mathbf{y}}_t^{(s)} &= D\left(\widehat{\boldsymbol{\lambda}}_0^{(s-1)}\right) \mathbf{W}\, \widetilde{\mathbf{y}}_t^{(s-1)} + D\left(\widehat{\boldsymbol{\lambda}}_1^{(s-1)}\right) \widetilde{\mathbf{y}}_{t-1}^{(s-1)} + D\left(\widehat{\boldsymbol{\lambda}}_2^{(s-1)}\right) \mathbf{W}\, \widetilde{\mathbf{y}}_{t-1}^{(s-1)} \\ \bar{\mathbf{y}}^{(s)} &= \frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{\delta}_t \circ \mathbf{y}_t + (\mathbf{1} - \boldsymbol{\delta}_t) \circ (\widehat{\mathbf{y}}_t^{(s)} + \bar{\mathbf{y}}^{(s-1)})\right) \\ \widetilde{\mathbf{y}}_t^{(s)} &= \boldsymbol{\delta}_t \circ (\mathbf{y}_t - \bar{\mathbf{y}}^{(s)}) + (\mathbf{1} - \boldsymbol{\delta}_t) \circ \widehat{\mathbf{y}}_t^{(s)}, \end{aligned} \tag{5}$$

where $\mathbf{1}$ is a vector of ones. We iterate steps (a) and (b) with increasing $s = 1, 2, \ldots$, until

$$\|\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)}\|_2^2 \leq \gamma \tag{6}$$

with γ sufficiently small. At the end of the procedure, the reconstructed multivariate time series is given by $\widetilde{\mathbf{y}}_t^{(s)} + \bar{\mathbf{y}}^{(s)}$, $t = 1, 2, \ldots, T$, where the original missing data have been replaced by the estimated values.

Remark 1 Note that the mean vector used to center the data is initially computed with the observed values alone, but then it is updated at each iteration and computed using the whole series, including the imputed values [compare the formula of $\bar{\mathbf{y}}^{(0)}$ in (4) with the formula in (5)]. Similarly and simultaneously, the sample covariance matrices $\widehat{\Sigma}_0$ and $\widehat{\Sigma}_1$ are updated at each iteration and used to derive new estimations of the spatial coefficients and more precise estimations of the missing values.

2.4 Theoretical foundations

Let denote with α the proportion of missing values, defined as

$$\alpha = \frac{1}{Tp}\sum_{t=1}^{T} \|\mathbf{1} - \boldsymbol{\delta}_t\|_2^2. \tag{7}$$

The following assumptions are required for the consistency of the imputation procedure:

A1 The spatial weight matrix $\mathbf{W}$ has zero main diagonal elements; moreover, matrix $S(\boldsymbol{\lambda}_0) = (\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W})$ is invertible.

A2 The disturbance $\boldsymbol{\varepsilon}_t$ satisfies $Cov(\mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t) = 0$. Moreover, the process $\mathbf{y}_t$ in model (1) is strictly stationary and α-mixing, with the mixing coefficients satisfying condition A2(c) of Dou et al. (2016) (there you can also find a definition for the α-mixing coefficients).

A3 The rank of matrix $(\boldsymbol{\Sigma}_1'\mathbf{w}_i, \boldsymbol{\Sigma}_0\mathbf{e}_i, \boldsymbol{\Sigma}_0\mathbf{w}_i)$ is equal to 3, for all i.

These assumptions guarantee the validity of Theorem 1 of Dou et al. (2016), which assures the consistency of the estimator (3), so justifying step (*a*) of the iterative procedure in Sect. 2.3. However, we must also prove that in presence of missing values the difference in (6) tends to zero as the number of iterations goes to infinity, $s \to \infty$, so that the algorithm reaches a stable solution. The latter is assured by the following assumption and next Theorem 1 (shown in the Appendix).

A4 Missing values occur at random, independently of the process level. Moreover, the spatial coefficients are such that $\|D(\boldsymbol{\lambda}_0)\mathbf{W}\|_2 + \|D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}\|_2 < 1$.

Theorem 1 *Under the assumptions A1–A4, for $s \to \infty$ the difference in (6) goes to zero and the iterative imputation procedure converges to a unique solution, for any α and T. Such a solution converges in probability to the true limit $\mathbf{y}_t^*$ as long as $T \to \infty$.*

Remark 2 As clear from the proof of Theorem 1 in the Appendix, the last part of Assumption A4 is only a sufficient condition necessary to guarantee the convergence of the iterative algorithm to a stable solution under any fixed T. It could be relaxed if one considers the exact convergence rates of all the stochastic limits as a function of the percentage of missing values, α, for $T \to \infty$. However, the evaluation of such rates goes beyond the aims of this paper.

By assumption A1, we can formulate model (1) in the following reduced form

$$\mathbf{y}_t = \boldsymbol{\mu}^* + \mathbf{A}^*\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t^*, \tag{8}$$

where $\boldsymbol{\mu}^* = [\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}]^{-1}\boldsymbol{\mu}$, the coefficient matrix is $\mathbf{A}^* = [\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}]^{-1}[D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}]$ and the error process is $\boldsymbol{\varepsilon}_t^* = [\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}]^{-1}\boldsymbol{\varepsilon}_t$. Given the (8), one can verify the stationarity assumptions by checking the eigenvalues of matrix $\mathbf{A}^*$. Moreover, since the model can be seen as a particular VAR(1) process, it is easy to derive the link between the autocovariance matrix $\boldsymbol{\Sigma}_j$ and the variance-covariace matrix of the error process $\boldsymbol{\Sigma}_\varepsilon$. It is given by

$$\boldsymbol{\Sigma}_j = \sum_{i=0}^{\infty}(\mathbf{A}^*)^{j+i}[\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}]^{-1}\boldsymbol{\Sigma}_\varepsilon[\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}]^{-1\prime}(\mathbf{A}^{*\prime})^i, \qquad j = 0, 1.$$

Of course, the above expression becomes simpler if one assumes a scalar variance-covariance matrix for the error process, i.e. $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon\mathbf{I}$. However, note that a full variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ allows for other sources of correlation in the data that are not captured and explained by the spatial matrix $\mathbf{W}$, without the necessity of including other regressors in the model (as instead is made in the other SDPD models that assumes a scalar error variance, as in Lee and Yu 2010a, b).

3 Application to environmental data

In our analysis, we consider daily PM_{10} data (in μg/m^3) from 1 January 2015 to 19 October 2016 (658 days) by gravimetric instruments at 24 sites in Piemonte. Piemonte is an Italian region in the western part of the Po valley and surrounded on three sides by the Alps. Data were provided from the website of Agenzia Regionale per la Protezione Ambientale (ARPA) Piemonte http://www.arpa.piemonte.gov.it. Data are shown in Fig. 1. The red points indicate the missing values showing the presence of a lot of missing sequences in the time series. For this dataset, we assume that missing values occur at random, meaning that they occur independently on their (unobserved) values. The presence of a spatial correlation between the considered time series is clear, in the sense that high/low values of PM_{10} concentration are in common between the stations and the same for high/low variability.

The names of the sites are reported in Table 1, together with the number and the corresponding percentage of missing values for each station, ranging from 0% to 34%.

In order to impute the missing values and missing sequences by means of our procedure, we use two different spatial matrices, denoted as $\mathbf{W}_1$ and $\mathbf{W}_2$. For the first one, $\mathbf{W}_1$, we take a normalized sample correlation matrix of $\mathbf{y}_t$, i.e. we let w_{ij} be the sample correlation between the i-th and j-th locations for $i \neq j$, and $w_{ii} = 0$ for $i = 1, \ldots, p$, and then replace w_{ij} by $w_{ij}/\sum_{k=1}^{p}|w_{kj}|$. For the second spatial matrix, $\mathbf{W}_2$, we consider $w_{ij} = 1/(1 + d_{ij})$ where d_{ij} is the geographical distance between the i-th and j-th stations for $i \neq j$, and $w_{ii} = 0$ for $i = 1, \ldots, p$, and then we row-normalize the matrix as before.

In all the runs of our procedure, here and in the simulation study, we set the maximum number of iterations to $N_{iter} = 100$ which determines a convergence value of $\delta = O(10^{-10})$.

Our missing data imputation was compared regarding performance with the one implemented in the `AmeliaII` imputation procedure as described in Sect. 1.1. For it,

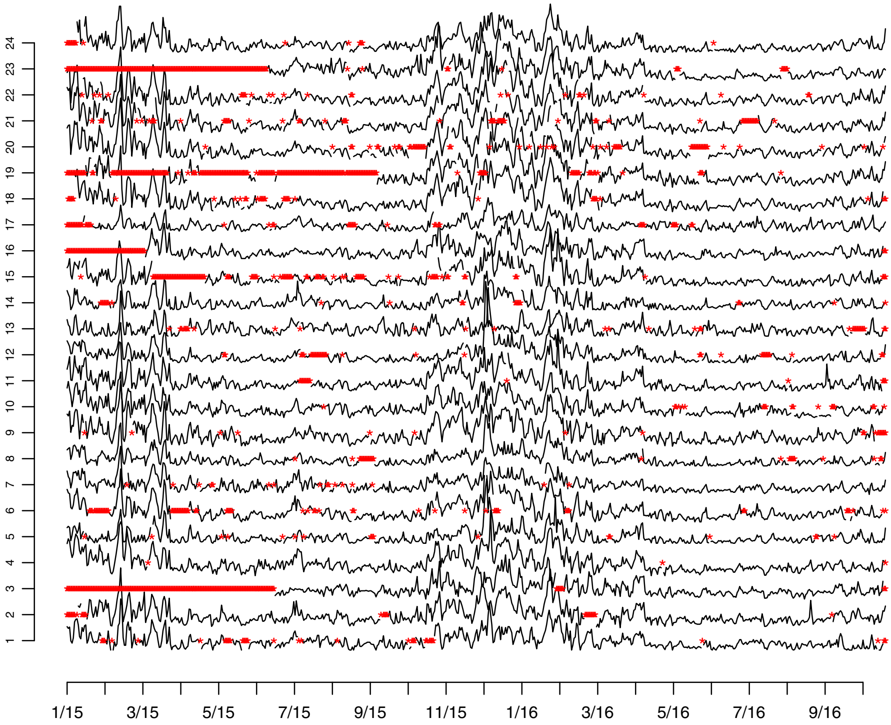

Fig. 1 Plots of the 24 time series for PM_{10} data, observed daily from January 2015 to October 2016 at the Italian stations listed in Table 3. Red points indicate missing values (color figure online)

we always consider a polynomial path of order $k = 3$ and we also add the lagged values and leads in the imputation model. In a view to do this, in the command `amelia()` we use the arguments `ts`="time", `cs`="location", `lags`="serie", `leads`="serie" and `polytime`=3, where "time" is the name of the covariate representing the time index, "location" is the name of the covariate representing the cross-section index, while "serie" contains the data.

In Fig. 2, as an illustrative example of how our procedure works, we consider the case of Torino Grassi ($i = 19$), in which the highest percentage of missing values is observed (34.04%), at the first step of the iterative procedure. On the top, the scatter plot of $y_{19\,t}$ against the spatial regressor $\mathbf{w}_i^T \mathbf{y}_t$, the dynamic regressor $y_{i,t-1}$ and the spatial-dynamic regressor $\mathbf{w}_i^T \mathbf{y}_{t-1}$ are reported on the left, center and right, respectively. All the values have been mean-centered and the missing values have been replaced by the mean value, therefore they are concentrated on the x-axis. The three plots on the bottom show the time-plots of the estimated spatial signal (dotted red line, on the left panel), the estimated dynamic signal (dotted blue line, in the center) and the estimated spatial-dynamic signal (dotted green line, on the right panel), which are superimposed on the observed time series (black line). The dashed horizontal line shows the EU threshold. Note how the estimated dynamic signal (blue line) is not able to reconstruct the missing sequences at the first step. In Fig. 3, the case of Torino Grassi is shown at

Table 1 Number of missing values in the considered stations and percentage of missing values

	Station	Missings	%		Station	Missings	%
1	Alba Tanaro	34	5.17	13	Cuneo Alpini	32	4.86
2	Alessandria-D’Annunzio	27	4.10	14	Druento La Mandria	21	3.19
3	Arquata Scrivia Minzoni	175	26.60	15	Novara Verdi	97	14.74
4	Asti Baussano	3	0.46	16	Novi Ligure Gobetti	65	9.88
5	Biella Sturzo	17	2.58	17	Pinerolo Alpini	43	6.53
6	Borgaro Torinese Caduti	62	9.42	18	Torino Consolata	32	4.86
7	Borgomanero Molli	16	2.43	19	Torino Grassi	224	34.04
8	Borgosesia-Tonella	25	3.80	20	Torino Lingotto	59	8.97
9	Carmagnola I Maggio	18	2.74	21	Torino Rebadeugo	53	8.05
10	Casale Monferrato Castello	18	2.74	22	Torino Rubino	25	3.80
11	Cerano Bagno	14	2.13	23	Tortona Carbone	174	26.44
12	Cossato Pace	34	5.17	24	Vercelli CONI	15	2.28

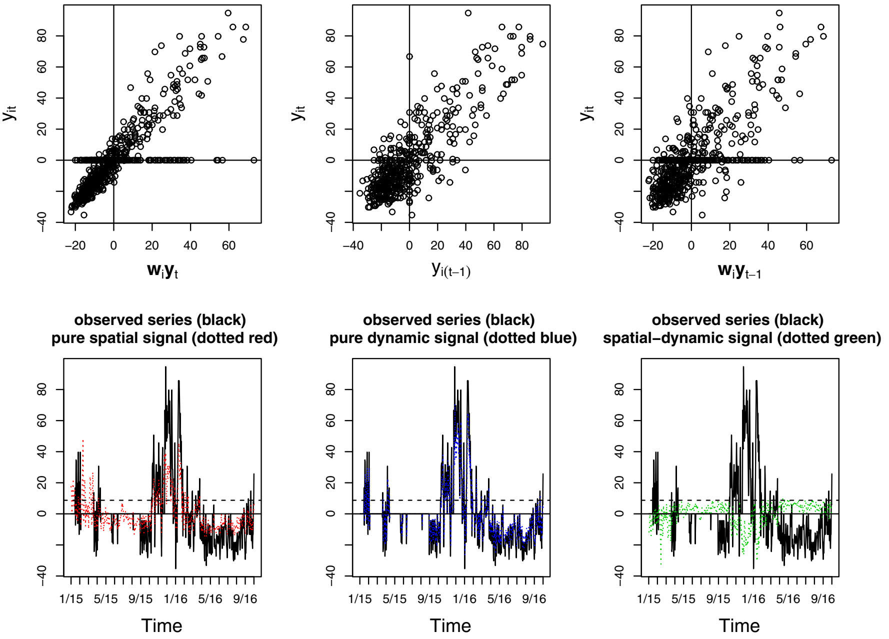

Fig. 2 Results for Torino-Grassi station (i.e., $i = 19$), at the first step of the procedure, using $\mathbf{W}_1$. The three plots on the top show the scatter-plots of the PM_{10} data y_{it} (after centering around the mean) against the spatial regressor $\mathbf{w}_i^T\mathbf{y}_t$, the dynamic regressor $y_{i,t-1}$ and the spatial-dynamic regressor $\mathbf{w}_i^T\mathbf{y}_{t-1}$ (on the left, center and right, respectively). The missing values have been replaced by the mean. On the bottom, the time-plots of the estimated spatial signal (dotted red line, on the left panel), the estimated dynamic signal (dotted blue line, in the center) and the estimated spatial-dynamic signal (dotted green line, on the right panel), are superimposed on the observed time series (black line). The dashed horizontal line shows the EU threshold (color figure online)

the last step of our imputation procedure. At this step, the missing values have been replaced by the estimated values. Therefore, also the estimated dynamic signal (blue dotted line) is entirely defined. Finally, in Fig. 4 the reconstructed time series of Torino Grassi is plotted. It is obtained simply as the sum of all the estimated components of Fig. 3, i.e. the pure spatial signal (red curve), the pure dynamic signal (blue curve), the spatial-dynamic signal (green curve) and the mean. The solid horizontal line denotes the mean value of the time series while the dashed red line is the upper threshold for the PM_{10} set by the European Commission ($50\,\mu g/m^3$).

Moreover, the 2008/50/EC directive states that the threshold $50\,\mu g/m^3$ can be exceeded no more than 35 days a year. So it becomes interesting to estimate the mean number of days per year of exceedances the legal limit. In Table 2 this mean number is calculated *(i)* based on the original data in which missing values are present; *(ii)* based on the reconstructed time series in which the proposed method is implemented by using the spatial matrix W_1 and *(iii)* basing on the reconstructed series using the spatial matrix W_2. The number of surplus with respect to the legal limit of 35 times per year is also indicated in all the cases. Clearly, by using the original data, the number of exceedances is often heavily underestimated (see Torino Grassi and Tortona Carbone

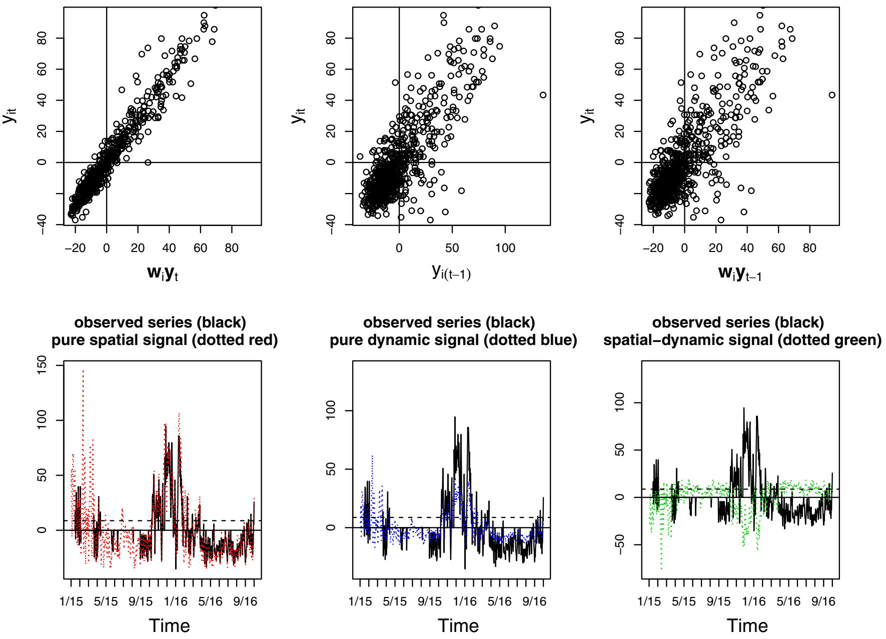

Fig. 3 As in Fig. 2, for the last step of the iterative procedure (color figure online)

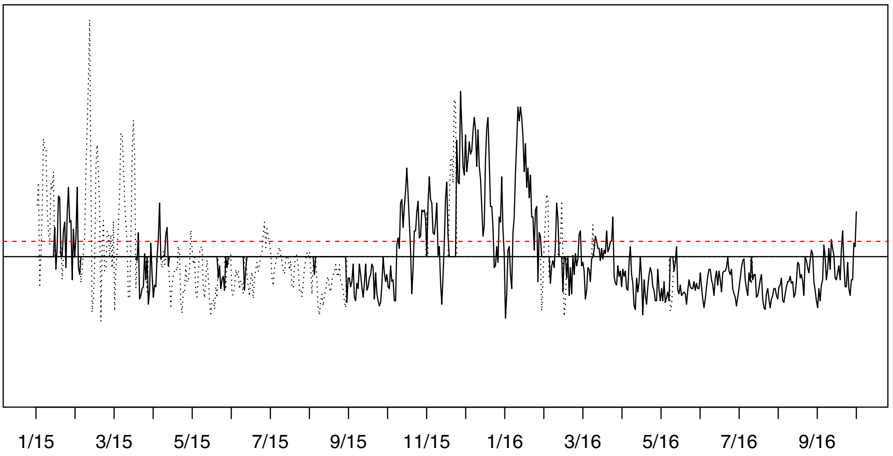

Fig. 4 Final output of our iterative procedure for the Torino-Grassi station. The dotted segments are the missing sequences that have been reconstructed by our imputation procedure. The solid horizontal line denotes the mean value of the time series while the dashed red line is the upper threshold for the PM_{10} set by the European Commission (color figure online)

Table 2 Number of days when the PM_{10} level exceeds 50 µg/m^3, for each station, during the year

Stations		Number of days exceeding the threshold					
		Original data		Fitted with W_1		Fitted with W_2	
		Days	Surplus	Days	Surplus	Days	Surplus
1	Alba-Tanaro	34.95		37.17	(+2)	37.17	(+2)
2	Alessandria-DAnnunzio	62.13	(+27)	67.67	(+33)	67.67	(+33)
3	ArquataScrivia-Minzoni	30.51		46.6	(+12)	46.6	(+12)
4	Asti-Baussano	65.46	(+30)	65.46	(+30)	65.46	(+30)
5	Biella-Sturzo	15.53		16.64		16.64	
6	BorgaroTorinese-Caduti	51.59	(+17)	55.47	(+20)	55.47	(+20)
7	Borgomanero-Molli	19.41		21.08		21.08	
8	Borgosesia-Tonella	19.41		21.08		21.08	
9	Carmagnola-IMaggio	76	(+41)	80.99	(+46)	80.99	(+46)
10	CasaleMonferrato-Castello	47.71	(+13)	48.26	(+13)	48.26	(+13)
11	Cerano-Bagno	51.59	(+17)	52.14	(+17)	52.14	(+17)
12	Cossato-Pace	26.07		26.63		26.63	
13	Cuneo-Alpini	13.31		13.87		13.87	
14	Druento-LaMandria	18.31		18.31		18.31	
15	Novara-Verdi	37.17	(+2)	43.82	(+9)	43.82	(+9)
16	NoviLigure-Gobetti	33.28		41.05	(+6)	39.94	(+5)
17	Pinerolo-Alpini	8.88		10.54		11.09	
18	Torino-Consolata	68.23	(+33)	72.11	(+37)	72.11	(+37)
19	Torino-Grassi	62.13	(+27)	91.53	(+57)	91.53	(+57)
20	Torino-Lingotto	59.91	(+25)	65.46	(+30)	65.46	(+30)
21	Torino-Rebaudengo	72.11	(+37)	82.65	(+48)	82.65	(+48)
22	Torino-Rubino	59.91	(+25)	63.79	(+29)	63.79	(+29)
23	Tortona-Carbone	39.94	(+5)	59.35	(+24)	58.8	(+24)
24	Vercelli-CONI	39.38	(+4)	41.6	(+7)	41.6	(+7)

Such a number should be less than 35 in a year. In brackets, the number of additional days exceeding the limit of 35, calculated (i) basing on the original data (i.e. ignoring the missing values), (ii) basing on the reconstructed series using our iterative procedure and the spatial matrix W_1 and (iii) basing on the reconstructed series using the spatial matrix W_2

in which the number of missing values is higher). Moreover, the number of surpluses estimated using the matrix W_1 and W_2 are in accordance in almost all the cases, showing that the covariance matrix is in accordance with the geographical distances.

4 Validating the procedure

To validate our procedure we evaluate the performance of our imputation technique against Amelia package on different sets of data: (1) PM_{10} concentrations, in which we randomly remove some observed data, as described in the next subsection; (2) some simulated data, both with zero and with non-zero mean value, as described in the last two subsections.

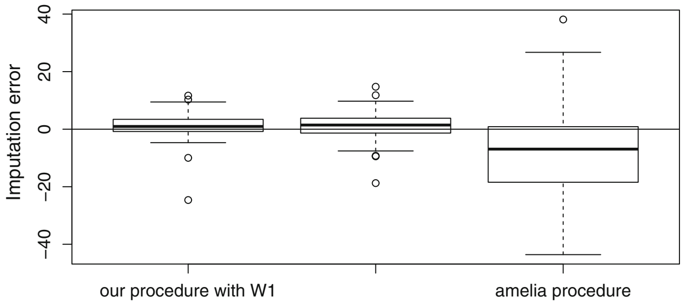

Fig. 5 From PM_{10} data, 50 observed values were removed and considered as missing (chosen randomly in the dataset, among which 30 as a single missing sequence). Then, all the missing values (original + simulated) have been imputed by using our procedure and by using the `AmeliaII` imputation procedure. The boxplots show the distribution of the imputation error for the 50 simulated missing values. The first one on the left shows our results using the spatial matrix $\mathbf{W}_1$, while the second one shows our results using the spatial matrix $\mathbf{W}_2$. The boxplot on the right summarises the results obtained with the `AmeliaII` imputation procedure, by using a local polynomial of order $k = 3$ and adding the (first order) lagged values of the series

4.1 Some performance indicators with PM_{10} data

From PM_{10} data, 50 observed values were removed and considered as missing (chosen randomly in the dataset, among which 30 as a single missing sequence). Then, all the missing values (original and simulated) have been imputed by using our procedure and by using the `AmeliaII` imputation procedure. The imputation error have been calculated, only for the simulated missing values, as

$$e_{it} = y_{it} - \tilde{y}_{it}, \tag{9}$$

where y_{it} is the true value and $\tilde{y}_{it}$ is the estimated value. The boxplots in Fig. 5 shows the distribution of the imputation error for the 50 simulated missing values. In the first case on the left, our procedure has been implemented using the spatial matrix $\mathbf{W}_1$ while the second boxplot in the centre is based on the spatial matrix $\mathbf{W}_2$. In the third boxplot on the right, the `AmeliaII` imputation procedure was applied considering a local polynomial of order $k = 3$ and adding the lags and leads of the series. Our imputation method produces a distribution of the imputation error with a lower variability with respect to AmeliaII, and the performance is substantially the same with the two spatial matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. Note that such a lower variability of our imputation procedure is also confirmed by the simulation study reported in Sects. 4.2 and 4.3. We explicitly observe that the boxplot for the Amelia procedure shows a bias probably due to the difficulty of correctly estimating the true mean because of the many missing sequences in the dataset (this difficulty is also confirmed by the simulation results in Sect. 4.3.

4.2 Performance with synthetic data with zero mean value

To further test the new imputation procedure, in this section we present the results of a Monte Carlo simulation study. We consider a case with $p = 30$ locations and

$T = (50, 100, 500, 1000)$ observations. The spatial matrix $\mathbf{W}_1$ has been randomly generated as a full rank symmetric matrix and has been row-normalized. The parameters of model (1) have been randomly generated in the interval $[-0.6, 0.6]$, assuring that the stationarity condition of the model is guaranteed. Assumption A4 is not satisfied for the simulated model, since $\|D(\boldsymbol{\lambda}_0)\mathbf{W}_1\|_2 + \|D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}_1)\|_2 \approx 5$. In such a way, we show how robust the imputation procedure is against violation of this assumption (remember that A4 is only a sufficient condition). The error component $\boldsymbol{\varepsilon}_t$ has been generated from a multivariate normal distribution, with mean vector zero and diagonal variance-covariance matrix, with heteroscedastic variances $(\sigma_1^2, \ldots, \sigma_p^2)$, where the standard deviations have been generated randomly from a Uniform distribution $U(0.5; 1.5)$.

In the first part of the simulation study, we assume $\boldsymbol{\mu} = 0$ in the model (1).

We simulated $N = 400$ replications of the model and, for each one, we removed 50 values and considered them as missing values. Of course, we kept a record of the true values. In particular, we simulated a missing sequence of length 30 for location 2 (i.e., the first 30 values of this location have been removed and considered as missing). The other 20 missing values have been generated randomly at other locations. Let M denote the index set of all the missing values, so that $\{y_{it}, \forall i, t \in M\}$ are all considered missing.

For each Monte Carlo replication, indexed by $r = 1, \ldots, N$, the imputation error have been calculated as in (9)

$$e_{it}^{(r)} = y_{it} - \tilde{y}_{it}^{(r)}, \qquad \forall i, t \in M,$$

where y_{it} is the true value (that has been removed from data and not used for the estimations) and $\tilde{y}_{it}^{(r)}$ is the estimated value. In the same way, we also derive the imputation error for the `AmeliaII` procedure. Note that $e_{it}^{(r)}$ also represents an estimation of the error component ε_{it} of model (1). Remembering this, we derive the average estimation error and the average squared error as follows,

$$AE_{it} = \frac{1}{N}\sum_{r=1}^{N}\left(y_{it} - \tilde{y}_{it}^{(r)}\right), \tag{10}$$

$$ASE_{it} = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(y_{it} - \tilde{y}_{it}^{(r)}\right)^2}, \qquad \forall i, t \in M \tag{11}$$

and compare them with the true values $E(\varepsilon_{it}) = 0$ and $sd(\varepsilon_{it}) = \sigma_{\varepsilon_i}$, respectively. Figure 6 shows the estimation error for the missing data for increasing time series length ($T = 50$ on the left side, $T = 100$ on the center and $T = 500$ on the right side). The x-axis of the plots summarises the 50 missing values, and the x-labels show the index of the location where the missing values have been simulated. As explained above, the first 30 missing values are sequentially generated from location 2, so they represent a missing sequence, whereas the last 20 are isolated missing values occurring at other locations.

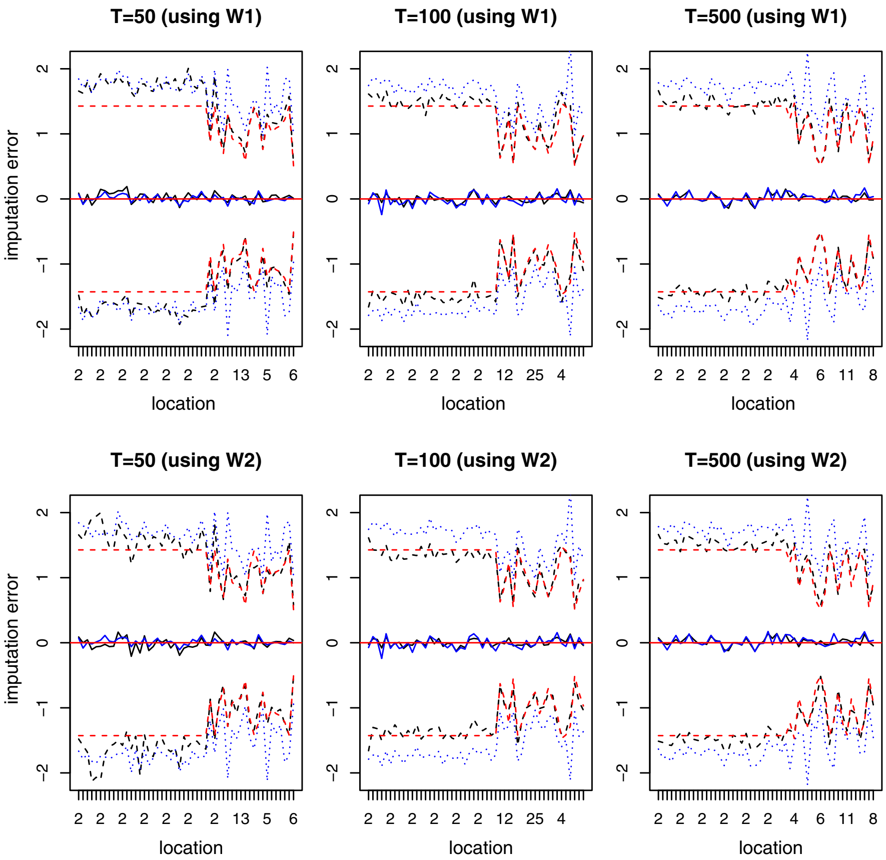

Fig. 6 Using the synthetic data of Sect. 4.2, the plots evaluate the estimation error for the missing data when the length of the time series is $T = 50, 100, 500$, respectively. The x-axis summarizes the 50 missing values and the x-labels show the number of the location where the missing values have been generated. The first 30 missing values are sequentially generated from location 2, so they represent a missing sequence, whereas the last 20 are isolated missing values occurring at other locations. The red dashed bands represent the interval $0 \pm \sigma_{\varepsilon_i}$, showing the true variability of the error component for those locations. The solid black line is $mean(y_{it} - \tilde{y}_{it})$, while the dashed black bands represent the interval $mean(y_{it} - \tilde{y}_{it}) \pm sd(y_{it} - \tilde{y}_{it})$, where the *mean* and *sd* operators are calculated using the formula (10) and (11), respectively, and $\tilde{y}_{it}$ are estimated using our approach. The solid blue line is $mean(y_{it} - \hat{y}_{it})$, while the dotted blue bands represent the interval $mean(y_{it} - \hat{y}_{it}) \pm sd(y_{it} - \hat{y}_{it})$, where $\hat{y}_{it}$ are estimated using the package `AmeliaII`. When T increases, the blue and black bands should tend to be equal to the red ones. The three plots on the top consider the case when the spatial matrix is W_1 whereas the three plots on the bottom consider the case when the spatial matrix is W_2 (color figure online)

The red dashed bands in Fig. 6 represent the interval $0 \pm \sigma_{\varepsilon_i}$, for different values of i, showing the true variability of the heteroscedastic error component ε_{it} (note that for the missing sequence, i.e. the first 30 points, the true variability is constant because the location is the same). The solid black line is $mean(y_{it} - \tilde{y}_{it})$, while the dashed black bands represent the interval $mean(y_{it} - \tilde{y}_{it}) \pm sd(y_{it} - \tilde{y}_{it})$, where the *mean* and *sd* operators are calculated using the formula (10) and (11), respectively, and $\tilde{y}_{it}$ are estimated using our approach. Finally, the solid blue line is $mean(y_{it} - \hat{y}_{it})$, while the dotted blue bands represent the interval $mean(y_{it} - \hat{y}_{it}) \pm sd(y_{it} - \hat{y}_{it})$, where $\hat{y}_{it}$ are estimated using the package `AmeliaII`. Note that, when T increases from 50 to 500 (from the left to the right side of the figure), the blue and black bands should tend to be equal to the red ones, as expected for the consistency of the procedures. This latter is true for our procedure (black dashed bands) since the variance is very soon close to the true value. On the other side, for the AmeliaII procedure (blue dotted bands), the variance of the imputation error is higher, and the convergence to the true value appears to be much slower. Concerning bias, both the methods seem to be substantially unbiased (as shown by the lines along the x-axis).

Moreover, to show the robustness of our imputation procedure when the spatial matrix $\mathbf{W}_1$ is not given, we derive the imputation errors using the spatial matrix $\mathbf{W}_2$, the row-normalized sample correlation matrix, as in Sect. 4. Note that data have still been simulated using matrix $\mathbf{W}_1$, the true spatial matrix, but then the estimations have been derived using matrix $\mathbf{W}_2$, which is the sample correlation matrix. In such a way, we analyse the robustness of the results when the true spatial matrix is unknown and, therefore, replaced with the estimated sample correlation matrix. The results are shown in the three plots on the bottom of Fig. 6. The performance is equivalent to the case where the spatial matrix is known.

Finally, we analyse the performance of the imputation procedures for different percentages of missing values in the dataset. Tables 3 and 4 show the results. The first one reports the mean (and standard deviations in brackets) of the standardized root mean square error for the estimated missing values, i.e. the mean value (and standard deviations) of $ASE_{it}/\sigma_{\varepsilon_i}$ calculated over $i, t \in M$, using our imputation procedure and the spatial matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ (true and estimated spatial matrices, respectively) and using the Amelia procedure. The data are simulated as before, but with different percentages of missing values. In particular, over the total number of observations of the multivariate time series, $T \times p$, the 2%, 5%, ..., 50% are randomly chosen and considered as missing. For example, there are 750 missing values when $T = 50$, $p = 30$ and the percentage is 50%. Among these, we always simulate a sequence of missing values which is long the 2%, 5%, ..., 50% of the time series length T, respectively. For example, when the time series length is $T = 100$ and the percentage is 50%, the missing sequence includes 50 sequential values.

The results in Table 3 confirm the better performance of our imputation procedure compared with Amelia. In fact, our procedure is consistent since the averages, and the standard deviations of the standardised root mean square error always decrease for increasing values of the time series length. On the other side, we can note that the Amelia procedure always produces worse results, and it does not improve when the time series length increases.

In order to better investigate on the bias of the two imputation procedures, Table 4 is derived similarly to Table 3, but now we focus on the average estimation error given by formula (10). The new results show that the two procedures are substantially equivalent in terms of bias, though our method shows slightly better results.

Table 3 Mean values of $ASE_{it}/\sigma_{\varepsilon_i}$ over $i, t \in M$ (with standard deviations in brackets) for the Monte Carlo replications of the synthetic data simulated in Sect. 4.2, with different percentages of missing values, using our imputation with $\mathbf{W}_1$ and $\mathbf{W}_2$ (true and estimated, respectively) and using the Amelia procedure

	$T = 50$	$T = 100$	$T = 500$	$T = 1000$
2% of missings				
Our procedure with $\mathbf{W}_1$	1.084 (0.071)	1.051 (0.099)	1.028 (0.086)	1.026 (0.083)
Our procedure with $\mathbf{W}_2$	1.053 (0.119)	1.043 (0.131)	1.064 (0.111)	1.073 (0.105)
Amelia	1.502 (0.268)	1.469 (0.284)	1.445 (0.271)	1.474 (0.274)
5% of missings				
Our procedure with $\mathbf{W}_1$	1.112 (0.137)	1.059 (0.065)	1.034 (0.079)	1.032 (0.085)
Our procedure with $\mathbf{W}_2$	1.062 (0.172)	1.054 (0.12)	1.083 (0.11)	1.083 (0.108)
Amelia procedure	1.454 (0.274)	1.472 (0.283)	1.492 (0.28)	1.47 (0.27)
10% of missings				
Our procedure with $\mathbf{W}_1$	1.146 (0.119)	1.088 (0.101)	1.049 (0.086)	1.043 (0.087)
Our procedure with $\mathbf{W}_2$	1.094 (0.153)	1.075 (0.132)	1.088 (0.115)	1.089 (0.113)
Amelia procedure	1.531 (0.272)	1.481 (0.282)	1.471 (0.271)	1.479 (0.275)
30% of missings				
Our procedure with $\mathbf{W}_1$	1.232 (0.133)	1.152 (0.112)	1.095 (0.109)	1.09 (0.108)
Our procedure with $\mathbf{W}_2$	1.168 (0.156)	1.137 (0.145)	1.124 (0.133)	1.126 (0.129)
Amelia procedure	1.477 (0.27)	1.476 (0.26)	1.484 (0.27)	1.485 (0.265)
50% of missings				
Our procedure with $\mathbf{W}_1$	1.358 (0.197)	1.225 (0.14)	1.141 (0.13)	1.134 (0.13)
Our procedure with $\mathbf{W}_2$	1.356 (0.271)	1.246 (0.222)	1.165 (0.152)	1.162 (0.147)
Amelia procedure	1.498 (0.257)	1.497 (0.261)	1.494 (0.266)	1.498 (0.269)

Table 4 Mean values of AE_{it} over $i, t \in M$ (with standard deviations in brackets) for the Monte Carlo replications of the synthetic data simulated in Sect. 4.2, with different percentages of missing values, using our imputation with $\mathbf{W}_1$ and $\mathbf{W}_2$ (true and estimated, respectively) and using the Amelia procedure

	50	100	500	1000
2% of missings				
Our procedure with $\mathbf{W}$	0.003 (0.046)	0.001 (0.06)	0.001 (0.06)	0.001 (0.052)
Our procedure with $\hat{\mathbf{W}}$	0.008 (0.046)	0.002 (0.057)	0 (0.063)	0 (0.052)
Amelia procedure	0.012 (0.079)	0.016 (0.075)	0.005 (0.075)	0.002 (0.07)
5% of missings				
Our procedure with $\mathbf{W}$	−0.005 (0.057)	−0.003 (0.054)	−0.001 (0.053)	−0.001 (0.055)
Our procedure with $\hat{\mathbf{W}}$	−0.007 (0.048)	0.002 (0.051)	−0.002 (0.053)	−0.001 (0.057)
Amelia procedure	−0.008 (0.067)	−0.007 (0.069)	0.001 (0.074)	0 (0.072)
10% of missings				
Our procedure with $\mathbf{W}$	0.006 (0.056)	0.003 (0.057)	0.003 (0.055)	−0.001 (0.055)
Our procedure with $\hat{\mathbf{W}}$	0 (0.053)	0 (0.052)	0.002 (0.056)	−0.001 (0.056)
Amelia procedure	0.003 (0.073)	0.004 (0.075)	0.004 (0.074)	0 (0.073)
30% of missings				
Our procedure with $\mathbf{W}$	0.002 (0.062)	0.004 (0.063)	0.002 (0.057)	0 (0.057)
Our procedure with $\hat{\mathbf{W}}$	−0.001 (0.057)	0.003 (0.059)	0.001 (0.058)	0 (0.058)
Amelia procedure	−0.002 (0.078)	0.002 (0.076)	−0.001 (0.074)	0.001 (0.074)
50% of missings				
Our procedure with $\mathbf{W}$	0.002 (0.073)	−0.001 (0.063)	0.001 (0.059)	0 (0.058)
Our procedure with $\hat{\mathbf{W}}$	0.007 (0.069)	−0.001 (0.067)	0.001 (0.06)	0 (0.059)
Amelia procedure	−0.001 (0.072)	0.003 (0.074)	0 (0.074)	0.001 (0.073)

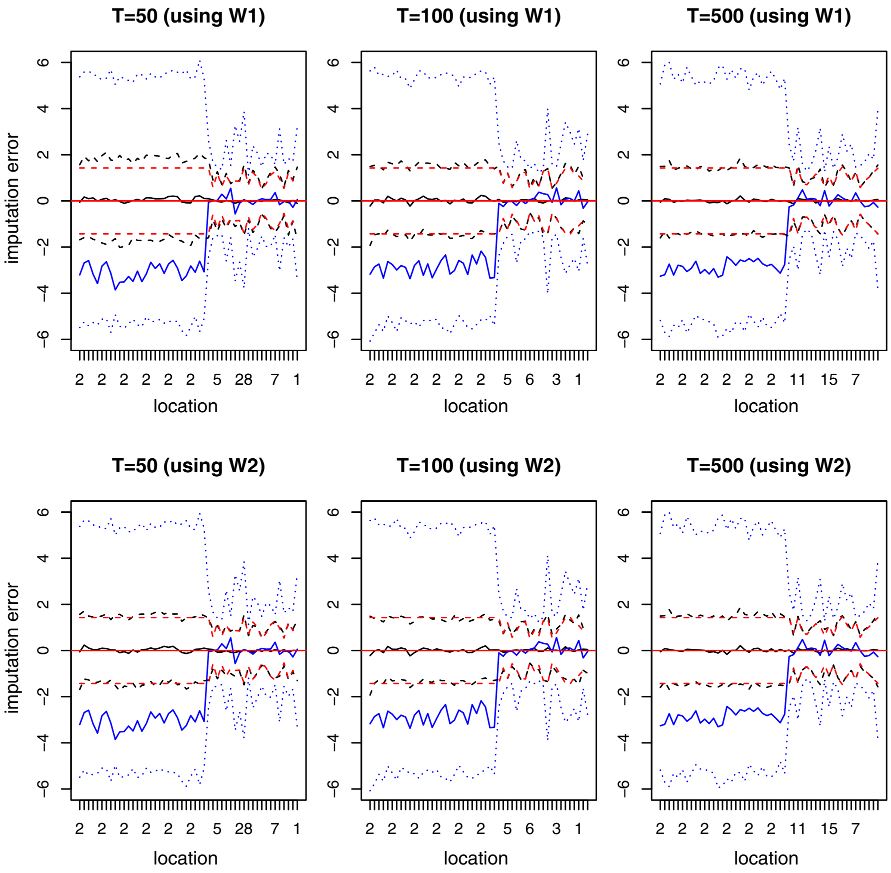

Fig. 7 Using the synthetic data of Sect. 4.2, the plots evaluate the estimation error for the missing data when the length of the time series is $T = 50, 100, 500$, respectively. The x-axis summarizes the 50 missing values and the x-labels show the number of the location where the missing values have been generated. The first 30 missing values are sequentially generated from location 2, so they represent a missing sequence, whereas the last 20 are isolated missing values occurring at other locations. The red dashed bands represent the interval $0 \pm \sigma_{\varepsilon_i}$, showing the true variability of the error component for those locations. The solid black line is $mean(y_{it} - \tilde{y}_{it})$, while the dashed black bands represent the interval $mean(y_{it} - \tilde{y}_{it}) \pm sd(y_{it} - \tilde{y}_{it})$, where the mean and sd are calculated using the formula (10) and (11), respectively, and $\tilde{y}_{it}$ are estimated using our approach. The solid blue line is $mean(y_{it} - \hat{y}_{it})$, while the dotted blue bands represent the interval $mean(y_{it} - \hat{y}_{it}) \pm sd(y_{it} - \hat{y}_{it})$, where $\hat{y}_{it}$ are estimated using the package `AmeliaII`. When T increases, the black and blue bands should tend to be equal to the red ones (but this happens only for black bands). The three plots on the top consider the case when the spatial matrix is W_1 whereas the three plots on the bottom consider the case when the spatial matrix is W_2 (color figure online)

4.3 Performance with synthetic data with non-zero mean value

Here we use the same setup as in the previous section apart from the mean value of the process, that is now assumed different from zero. In particular, we simulate the multivariate time series with intercept vector $\boldsymbol{\mu} = (10, \ldots, 10)'$. Figure 7 is organized similarly to Fig. 6, but now the plots show a substantial bias for the Amelia

Table 5 Ratio between the computational times of Amelia and our procedures.

Percentage of missing values	$T = 50$	$T = 100$	$T = 500$	$T = 1000$
2%	1.52	2.30	5.74	7.80
50%	1.46	2.26	5.58	6.74

procedure when it estimates a sequence of missing values, whereas our procedure remains unbiased, in both the cases base on $\mathbf{W}_1$ and $\mathbf{W}_2$.

4.4 Computational times

From the computational point of view, it is worth to stress that all the estimations involved in our imputation procedure are expressed in closed form, as they substantially represent weighted sums. As a consequence, the computational procedure is very fast and can be used efficiently also for very high dimensional datasets. Table 5 reports the ratio between the computational times of Amelia (in the numerator) and of our procedure (in the denominator), for the simulated data with different time series lengths and different missing percentages. In order to make comparisons fair, bootstrap is absent in both the methods.

5 Conclusions

This paper deals with missing values, sparse or sequences, in multivariate time series. We propose an iterative imputation technique able to handle with also high order time series. It is essentially based on the spatio-temporal model first introduced in Dou et al. (2016) and whose estimation can be easily implemented since the involved estimators are obtained in closed form. The model belongs to the category of *econometric spatial dynamic models*, driven by the presence of the spatial matrix $\mathbf{W}$ which measures the spatial correlation among the p components of the multivariate time series. We prove the consistency of our procedure in Theorem 1. Moreover, the simulation experiment, in which suitable missing sequences and values were randomly generated shows that our procedure produces estimates very close to the true values. Also, the estimated errors seem to have variance less than the ones produced by looking at the method implemented in AmeliaII package. Further, when the time series length increases the variance of the imputation error converges to zero, as expected by the consistency of the procedure. Finally, the proposed imputation procedure seems to be robust to different percentages of missing values, showing a good performance even when half of the dataset is missing. This latter feature is due to the high capability of the imputation model to capture different sources of correlation from the observed part of the data (in time and/or space), and to use them in order to better forecast the missing part of the time series.

Appendix A: Proof of Theorem 1

For the sake of simplicity, we assume here a process with zero mean value. First, we consider the case in which the parameters $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ are known. For $s \geq 1$ we have

$$\widetilde{\mathbf{y}}_t^{(s+1)} - \widetilde{\mathbf{y}}_t^{(s)} = (\mathbf{1} - \boldsymbol{\delta}_t) \circ \left(\widehat{\mathbf{y}}_t^{(s+1)} - \widehat{\mathbf{y}}_t^{(s)} \right)$$

and $\|\widetilde{\mathbf{y}}_t^{(s+1)} - \widetilde{\mathbf{y}}_t^{(s)}\|_2 \leq \|\widehat{\mathbf{y}}_t^{(s+1)} - \widehat{\mathbf{y}}_t^{(s)}\|_2$, so we focus on $\widehat{\mathbf{y}}_t^{(s+1)} - \widehat{\mathbf{y}}_t^{(s)}$. We can write

$$\widehat{\mathbf{y}}_t^{(s+1)} - \widehat{\mathbf{y}}_t^{(s)} = D(\boldsymbol{\lambda}_0)\mathbf{W} \left(\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)} \right) + [D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}] \left(\widetilde{\mathbf{y}}_{t-1}^{(s)} - \widetilde{\mathbf{y}}_{t-1}^{(s-1)} \right),$$

and

$$\begin{aligned} \|\widehat{\mathbf{y}}_t^{(s+1)} - \widehat{\mathbf{y}}_t^{(s)}\|_2 &\leq \|D(\boldsymbol{\lambda}_0)\mathbf{W}(\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)})\|_2 \\ &\quad + \|\, [D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}]\, (\widetilde{\mathbf{y}}_{t-1}^{(s)} - \widetilde{\mathbf{y}}_{t-1}^{(s-1)})\|_2 \\ &\leq \|D(\boldsymbol{\lambda}_0)\mathbf{W}\|_2 \|\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)}\|_2 \\ &\quad + \|D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}\|_2 \|\widetilde{\mathbf{y}}_{t-1}^{(s)} - \widetilde{\mathbf{y}}_{t-1}^{(s-1)}\|_2. \end{aligned} \tag{12}$$

Defining the vector operator $\Delta^j(\mathbf{x}_t) = (1 - \boldsymbol{\delta}_{t-j}) \circ \mathbf{x}_{t-j}$ and iterating the inequality in (12), we obtain

$$\begin{aligned} &\leq \sum_{j=0}^{s-1} \binom{s-1}{j} \|D(\boldsymbol{\lambda}_0)\mathbf{W}\|_2^{s-1-j} \|D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}\|_2^j \left\| \Delta^j \left(\widehat{\mathbf{y}}_t^{(2)} - \widehat{\mathbf{y}}_t^{(1)} \right) \right\|_2 \\ &\leq (\|D(\boldsymbol{\lambda}_0)\mathbf{W}\|_2 + \|D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}\|_2)^{s-1} \max_j \left\| \Delta^j \left(\widehat{\mathbf{y}}_t^{(2)} - \widehat{\mathbf{y}}_t^{(1)} \right) \right\|_2. \end{aligned}$$

In the extreme case when $\alpha = 0$, we have $\Delta^j(\cdot) \equiv \mathbf{0}$ for all t, j, because all the vectors $(\mathbf{1} - \boldsymbol{\delta}_t)$ are zero, and the convergence of the iterative procedure is trivially proved. In the opposite case when $\alpha = 1$ (only theoretically, of course), we have $\Delta^j \equiv \mathbf{0}$ for all j because all $\widehat{\mathbf{y}}_t^{(s)}$ are zero vectors, so again the convergence of the iterative algorithm is trivially proved. In the remaining more realistic cases when $\alpha \in (0, 1)$, the convergence is guaranteed by assumption A4. All this implies that the iterative procedure always converges to a limit $\widetilde{\mathbf{y}}_t^{(\infty)}$, for any T and α, under the assumption of known parameters $\boldsymbol{\lambda}_j$.

Note that A4 is a sufficient condition to get the convergence of (12) for any T, and could be relaxed if one considers that $\max_j \left\| \Delta^j \left(\widehat{\mathbf{y}}_t^{(2)} - \widehat{\mathbf{y}}_t^{(1)} \right) \right\|_2$ also converges to zero in probability as $T \to \infty$, with a rate depending on the percentage of missing values, α. However, the analysis of the exact rate goes beyond the aim of this paper.

In order to deal with the estimated parameters, we consider the following Taylor expansion of $\widehat{\mathbf{y}}_t^{(s)}$

$$\begin{aligned}\widehat{\mathbf{y}}_t^{(s)} &= D(\boldsymbol{\lambda}_0)\mathbf{W}\widetilde{\mathbf{y}}_t^{(s-1)} + D(\boldsymbol{\lambda}_1)\widetilde{\mathbf{y}}_{t-1}^{(s-1)} + D(\boldsymbol{\lambda}_2)\mathbf{W}\widetilde{\mathbf{y}}_{t-1}^{(s-1)} \\ &+D(\widehat{\boldsymbol{\lambda}}_0^{(s-1)} - \boldsymbol{\lambda}_0)\mathbf{W}\widetilde{\mathbf{y}}_t^{(s-1)} + D\left(\widehat{\boldsymbol{\lambda}}_1^{(s-1)} - \boldsymbol{\lambda}_1\right)\widetilde{\mathbf{y}}_{t-1}^{(s-1)} \\ &+D(\widehat{\boldsymbol{\lambda}}_2^{(s-1)} - \boldsymbol{\lambda}_2)\mathbf{W}\widetilde{\mathbf{y}}_{t-1}^{(s-1)}\end{aligned}$$

where the first row of the equality is exactly the quantity analysed in the (12) whereas the other rows depend on the differences $\widehat{\boldsymbol{\lambda}}_j^{(s)} - \boldsymbol{\lambda}_j$, $j = 0, 1, 2$. Now, remembering the (4), $\widehat{\boldsymbol{\lambda}}_j^{(0)} - \boldsymbol{\lambda}_j$ converges to zero in probability for $T \to \infty$ by Theorem 1 in Dou et al. (2016), for all j. So, assuming $\widehat{\boldsymbol{\lambda}}_j^{(s-1)}$ and $\widetilde{\mathbf{y}}_t^{(s-1)}$ consistent and following the iterative algorithm, by induction, we can conclude that also $\widehat{\boldsymbol{\lambda}}_j^{(s)} - \boldsymbol{\lambda}_j$ converges to zero in probability for $T \to \infty$. Combining this with the previous result in (12), we finally have that $\widehat{\mathbf{y}}_t^{(s)}$ and $\widetilde{\mathbf{y}}_t^{(s)}$ converge in probability to the limit $\mathbf{y}_t^*$, for $T \to \infty$ and $s \to \infty$.

Of course, the convergence rates of all the previous stochastic limits are expected to depend on the percentage of missing values, α. Again, the evaluation of the exact rates goes beyond the aims of this paper. Here, instead, we analyse heuristically the "quality" of the imputation output, i.e. how near the final imputed values are to the true latent ones, as a function of the proportion of missing values in the time series.

By simple algebra, we can write

$$\begin{aligned}\mathbf{y}_t - \widehat{\mathbf{y}}_t^{(s+1)} &= D(\boldsymbol{\lambda}_0)\mathbf{W}\left(\mathbf{y}_t - \widetilde{\mathbf{y}}_t^{(s)}\right) + [D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}]\left(\mathbf{y}_{t-1} - \widetilde{\mathbf{y}}_{t-1}^{(s)}\right) \\ &+\ D\left(\widehat{\boldsymbol{\lambda}}_0^{(s)} - \boldsymbol{\lambda}_0\right)\mathbf{W}\widetilde{\mathbf{y}}_t^{(s)} + \left[D\left(\widehat{\boldsymbol{\lambda}}_1^{(s)} - \boldsymbol{\lambda}_1\right) + D\left(\widehat{\boldsymbol{\lambda}}_2^{(s)} - \boldsymbol{\lambda}_2\right)\mathbf{W}\right]\widetilde{\mathbf{y}}_{t-1}^{(s)} + \boldsymbol{\varepsilon}_t \\ &= L_1(\alpha) + L_2(\alpha) + \boldsymbol{\varepsilon}_t.\end{aligned}$$

In the extreme case when all data are missing (only theoretically, of course), $\alpha = 1$ and the algorithm imputes zero to all data since $\widetilde{\mathbf{y}}_t^{(s)} \equiv \mathbf{0}$ for all s and t, as obvious. In such a case, $L_2(\alpha) \equiv \mathbf{0}$ and the imputation error is

$$\mathbf{y}_t - \widehat{\mathbf{y}}_t^{(s)} = D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t + [D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W}]\,\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \qquad \text{for } \alpha = 1, \forall s, \forall T,$$

which is something very different from desired (*worst imputation quality*), still centered around zero but with higher variability. In the other cases, assuming $\alpha \in [0, 1)$ approximately fixed for $T \to \infty$, we have $(\widehat{\boldsymbol{\lambda}}_j^{(s)} - \boldsymbol{\lambda}_j) \xrightarrow{p} 0$ by Theorem 1 of Dou et al. (2016) and, then, $(\mathbf{y}_t - \widetilde{\mathbf{y}}_t^{(s)}) \xrightarrow{p} 0$ by (12). Therefore, $L_1(\alpha) \xrightarrow{p} \mathbf{0}$ and $L_2(\alpha) \xrightarrow{p} \mathbf{0}$ for $T \to \infty$ and the imputation error converges to

$$\mathbf{y}_t - \widehat{\mathbf{y}}_t^{(s)} \xrightarrow{p} \boldsymbol{\varepsilon}_t \qquad \text{for } \alpha \in [0, 1), s \to \infty \text{ and } T \to \infty, \tag{13}$$

as desired (*best imputation quality*), but the convergence rate of the (13) is expected to be faster as long as the proportion α approaches to zero. The fastest convergence rate is derived in Dou et al. (2016) and is reached when $\alpha = 0$. •

References

Aga E, Samoli E, Touloumi G, Anderson HR, Cadum E, Forsberg B (2003) Short-term effects of ambient particles on mortality in the elderly: results from 28 cities in the APHEA2 project. Eur Resp J Suppl 40:28s33s

Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht

Biggeri A, Baccini M, Accetta G, Lagazio C (2002) Estimates of short-term effects of air pollutants in Italy. Epidemiologia e Prevenzione 26:203205

Calculli C, Fassò A, Finazzi F, Pollice A, Turnone A (2015) Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. Environmetrics 26:406–417

Cameletti M, Ignaccolo R, Bande S (2011) Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics 22:985996

Dou B, Parrella ML, Yao Q (2016) Generalized Yule–Walker estimation for spatio-temporal models with unknown diagonal coefficients. J Econom 194:369–382

Fitri MDNF, Ramli NA, Yahaya AS, Sansuddin N, Ghazali NA, Al Madhoun W (2010) Monsoonal differences and probability distribution of PM_{10} concentration. Environ Monit Assess 163:655–667

Honaker J, King G, Blackwell M (2011) Amelia II: a program for missing data. J Stat Softw 45(7):1–47

Josse J, Husson F (2016) missMDA: a package for handling missing values in multivariate data analysis. J Stat Softw 70(1):1–31

Junninen H, Niska H, Tuppurrainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907

Kowarik A, Templ M (2016) Imputation with the R package VIM. J Stat Softw 74(7):1–16

Lee LF, Yu J (2010a) Estimation of spatial autoregressive panel data models with fixed effects. J Econom 154:165–185

Lee LF, Yu J (2010b) Some recent developments in spatial panel data models. Reg Sci Urban Econ 40:255–271

Liu S, Molenaar PC (2014) iVAR: a program for imputing missing data in multivariate time series using vector autoregressive models. Behav Res Method 46(4):1138–1148

Moritz S, Bartz-Beielstein T (2017) imputeTS: time series missing value imputation in R. R J 9:207–218

Norazian MN, Shukri YA, Azam RN, Al Bakri AMM (2008) Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34:341–345

Oehmcke S, Zielinski O, Kramer O (2016) kNN ensembles with penalized DTW for multivariate time series imputation. In: International joint conference on neural networks (IJCNN), IEEE

Pollice A, Lasinio GJ (2009) Two approaches to imputation and adjustment of air quality data from a composite monitoring network. J Data Sci 7:43–59

Raaschou-Nielsen O, Andersen ZJ, Beelen R, Samoli E, Stafoggia M, Weinmayr G (2013) ir pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). Lancet Oncol 14(9):813–822

van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67