Computational Statistics (2021) 36:2917–2938
https://doi.org/10.1007/s00180-021-01099-y

ORIGINAL PAPER

Bootstrap joint prediction regions for sequences of missing values in spatio-temporal datasets

Maria Lucia Parrella[1] · Giuseppina Albano[2] · Cira Perna[1] · Michele La Rocca[1]

Received: 8 July 2020 / Accepted: 13 March 2021 / Published online: 5 April 2021

Abstract

Missing data reconstruction is a critical step in the analysis and mining of spatio-temporal data. However, few studies comprehensively consider missing data patterns, sample selection and spatio-temporal relationships. To take into account the uncertainty in the point forecast, some prediction intervals may be of interest. In particular, for (possibly long) missing sequences of consecutive time points, joint prediction regions are desirable. In this paper we propose a bootstrap resampling scheme to construct joint prediction regions that approximately contain missing paths of a time components in a spatio-temporal framework, with global probability $1 - \alpha$. In many applications, considering the coverage of the whole missing sample-path might appear too restrictive. To perceive more informative inference, we also derive smaller joint prediction regions that only contain all elements of missing paths up to a small number k of them with probability $1 - \alpha$. A simulation experiment is performed to validate the empirical performance of the proposed joint bootstrap prediction and to compare it with some alternative procedures based on a simple nominal coverage correction, loosely inspired by the Bonferroni approach, which are expected to work well standard scenarios.

Keywords Spatio-temporal models · Path missing reconstruction · Joint prediction regions

✉ Giuseppina Albano
pialbano@unisa.it

Maria Lucia Parrella
mparrell@unisa.it

Cira Perna
perna@unisa.it

Michele La Rocca
larocca@unisa.it

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

[2] Dip. di Studi Politici e Sociali, Università of Salerno, Fisciano, Salerno, Italy

1 Introduction

In the last decade, data with temporal and spatial attributes are quickly accumulated and form large numbers of spatio-temporal datasets collected in diverse domains, including climate science, social sciences, economics, neuroscience, epidemiology, transportation, mobile health, and Earth sciences. In these domains, the real-world processes being studied are intrinsically spatio-temporal, and several data acquisition methodologies have been proposed to record the spatial and temporal information of every measurement in the data. For example, when dealing with environmental data generated by air quality monitoring sensors, data are simultaneously collected by monitoring stations for different sites and different time points (see Atluri et al. 2018 for a review).

In all these fields, missing data are pervasive as it often happens that, for several reasons including equipment failure or measurement errors, (possibly) long sequences of data are not correctly recorded. If these holes in data cannot be accurately estimated, the subsequent steps of data analysis and modelling might lead to incorrect results and unreasonable inference. Clearly, merely deleting the records containing missing data cannot be considered a sensible strategy since it would lead to a significant loss of initial information and would be a waste of data resources. Therefore, statistical methods which are able to accurately and efficiently interpolate missing values have been proposed in the literature. Clearly, in the context of spatio-temporal data, the problem of missing data reconstruction becomes even more challenging, given the additional complexity of the spatial and time-dependent structure, which is present in the observed datasets. However, the spatial and time dependency, when correctly modeled, can be effectively exploited to get accurate reconstruction of (possibly) long sequences of missing values.

Studies on the missing values estimation in meteorological time series go back to the 1950s. In the first studies, missing values are estimated by means of imputation or simple linear regression estimates. For the same purpose, in Young (1992) three methods, normal ratio method, multiple discriminant analysis and multiple linear regression are discussed and compared. In Eischeid et al. (2000) several estimation techniques based on spatial analysis schemes are evaluated to create a complete national daily time series of the maximum-minimum temperatures and total precipitation over the western United States. In Teegavarapu et al. (2005) an imputation method based on the inverse distance weighting method, by looking at the distance between target and reference stations, was proposed. The use of a regularized EM algorithm for the imputation of incomplete climate data is recommended in Schneider (2001), especially when the number of the observed series with missing values exceeds the sample size. Another study that proposes multiple imputation is provided in Cano et al. (2010), in which a MCMC-based procedure is suggested. Other appreciable works in the area of metereological time series are carried out by Junninen et al. (2004); Lo Presti et al. (2010); Smith et al. (2007).

Recently, some new approaches have been proposed in the literature, with clear evidence of substantial advantages compared with existing methods (see, Pollice et al. 2009; Liu et al. 2014; Yang et al. 2018 inter alia). 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 estimates obtained by using EM algorithm. This approach is able to deal with multiple variables sampled at different monitoring networks and missing data. In Parrella et al. (2019) we proposed a new procedure for estimating (even long) missing sequences in time series, which 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 et al. 2010 for an introduction and a survey of such 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, it is possible to take into account spatial correlation in the data and estimate missing sequences in one given spatial site by looking at near sites, but also by looking at the previous lags of the same station and the neighbour sites.

However, when estimating the missing values for a given site, a point forecast alone is usually not sufficient. A statement about the uncertainty contained in the point forecast, as expressed by some prediction intervals, may also be desired. Moreover, missing values are not only isolated missing points, but often there are (possibly long) missing sequences of data points in spatio-temporal databases. Hence, we usually have a path of missing values for H consecutive time points, for a given site. A path-missing reconstruction refers to the sequence of corresponding missing values imputation for the H missing time points.

On the one hand, one can construct H marginal prediction intervals by using a given method to build a prediction interval repeatedly, one period at a time. But, by design, probability statements then only apply marginally, one period at a time: the prediction interval at a specific time point t_h, for some $t_1 \leq t_h \leq t_H$, will contain the random variable representing the missing value with prespecified probability $1-\alpha$. The problem has been already addressed in the literature with effective proposals as in Alonso et al. (2008) and Alonso et al. (2013) where a bootstrap scheme is employed to construct accurate interpolation intervals.

On the other hand, a more general problem is the construction of a joint prediction region (JPR) that will contain the entire missing path with the desired probability $1-\alpha$. Clearly, stringing together marginal prediction intervals for time points t_1 up to t_H, each one at level $1-\alpha$, will not result in a JPR that contains the entire missing path with probability $1-\alpha$. Instead, like the case of prediction intervals, apart from pathological cases, the joint coverage probability of the missing sequence will be strictly less than $1-\alpha$, and decreasing in H. Despite the importance of the problem just described, the construction of JPRs for missing paths has been somewhat neglected in the literature so far.

In this paper we propose a bootstrap resampling scheme to construct JPRs that contain missing paths of a time series of interest with nominal coverage level $1-\alpha$, along the same approach used by Wolf et al. (2015) and based on the maximum predictive root. Moreover, if the missing value time horizon H is large, the applied researcher may deem the criterion that all elements of the missing path must be contained in the JPR with probability $1-\alpha$ as too strict. We also consider the more general problem

of constructing JPRs that will only contain all elements of missing paths up to a small number $k-1$, ($k = 1, 2, \ldots$) of them with probability $1-\alpha$. The choice of k must be made by the applied researcher, with respect to the given problem at hand. But it will be useful to the applied researcher to have a method available that can handle any desired value of k. In particular, the choice $k = 1$ yields a standard JPR that must contain all elements of a missing path with probability $1-\alpha$.

A simulation experiment is performed to validate the empirical performances of the proposed JPRs construction method. For the sake of comparison, we also propose two alternative resampling techniques for the construction of JPRs, based on a simple nominal coverage correction loosely inspired by the Bonferroni approach, which are expected to work well in simple and standard scenarios.

The remainder of this paper is organized as follows. Section 2 contains a short review on the spatio-temporal model used and the missing data reconstruction procedure. Section 3 describes our method to construct JPRs for missing paths in spatio-temporal datasets. Section 4 investigates the finite-sample performance via Monte Carlo simulations. An application to real data is also presented in Sect. 5. Finally, some concluding remarks close the paper.

2 Path missing values reconstruction in spatio-temporal datasets

Let $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \ldots, y_{t,p})$ be a multivariate stationary process of dimension p, assumed for simplicity with zero mean value vector, generating the observations at time t from p different locations. Following Dou et al. (2016), we assume that the process can be modeled by the following *Spatial Dynamic Panel Data* (*SDPD*) model

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

where $D(\cdot)$ are diagonal matrices with diagonal coefficients from the vectors $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, and the error process $\boldsymbol{\epsilon}_t$ is serially uncorrelated, with diagonal heteroskedastik variance-covariance matrix $\boldsymbol{\Sigma}_\epsilon$. Model (1) belongs to the family of *spatial econometric models*, so it is particularly oriented to model spatio-temporal data. The matrix $\mathbf{W}$ is called *spatial matrix* and collects the weigths used in the *spatial regression* of each time series observation with simultaneous or delayed observations of neighboring data. In particular, note that the term $D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t$ captures the pure spatial effects, since it only considers contemporary observations, the component $D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1}$ captures the pure dynamic effects, since it involves lagged observations, while $D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1}$ captures the spatial-dynamic effects. However, if one uses a correlation based matrix $\mathbf{W}$ to measure variable distances, instead of using physical distances, one can use model (1) to analyse any kind of multivariate time series, not necessarily of strictly spatial nature. Of course, spatial models rely on the spatial weight matrix $\mathbf{W}$ to specify the cross-sectional correlation. Currently, there is no well-defined theory on how to find the true spatial matrix for a given data application. Recently, there have been some proposals to make this matrix "endogenous" within the model or to select the best spatial matrix among a list of candidates, in order to increase the flexibility of the spatial model (see, for example, Gao et al. 2019; Qu et al. 2021; Zhang et al. 2018). Of course,

some additional assumptions and/or some adjustments to the model parametrization are required for identifiability. In this paper, we assume that matrix $\mathbf{W}$ is known, as usually done within this class of models, but we defer further investigations on the selection of the "best" matrix $\mathbf{W}$ for the future.

In the following, we assume that $\mathbf{y}_1, \ldots, \mathbf{y}_T$ are realizations from the stationary process defined by (1). Then, 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 superscript denotes the transpose operator.

The parameters of model (1) can be estimated following Dou et al. (2016). In particular, given stationarity, 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 equation system is

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

with $\mathbf{w}_i$ the i-th row vector of $\mathbf{W}$ and $\mathbf{e}_i$ the i-th unit vector. 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 generalized Yule-Walker estimator, available in closed form,

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

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 $\widehat{\mathbf{y}}_i = \widehat{\boldsymbol{\Sigma}}'_1\mathbf{e}_i$.

Model (1), once estimated, can be used to reconstruct sequences of missing values. To this aim, let us assume that $\widetilde{\mathbf{y}}_1, \cdots, \widetilde{\mathbf{y}}_T$ are realizations from a stationary spatio-temporal process with mean value not necessarily equal to zero. Let $\boldsymbol{\delta}_t = (\delta_{t,1}, \ldots, \delta_{t,p})$ be a vector of zeroes/ones that identifies all the missing values in the observed vector $\widetilde{\mathbf{y}}_t$, so that $\delta_{t,i} = 0$ if the observation $\widetilde{y}_{t,i}$ is missing, otherwise it is $\delta_{t,i} = 1$. In case of processes with no zero mean, model (1) can still be used for parameter estimation after a pre-processing step which centers the observed time series, i.e. $\mathbf{y}_t = \widetilde{\mathbf{y}}_t - \bar{\mathbf{y}}$. Now, because we have missing values in the observed spatio-temporal series, centering data using the initial mean $\bar{\mathbf{y}}$ (excluding missing values) is like centering using the *wrong* mean. As a result, we can have a bias in the estimation results (unless we can assume missing observations happen at random, *i.e.* they occur independently on the level of the process). To prevent this problem, we include the computation of the mean in the iterative procedure, so that mean-centering is made iteratively on the basis of the whole series (including the imputed values). See Parrella et al. (2019) for more details about this issue.

Algorithm 1 Pseudocode for missing value path imputation

Initialize the mean centered vector $\mathbf{y}_t^{(0)}, \quad t = 1, \ldots, T$ as
for $t = 1, \ldots, T$ **do**

$$\mathbf{y}_t^{(0)} = \boldsymbol{\delta}_t \circ \left(\widetilde{\mathbf{y}}_t - \overline{\mathbf{y}}^{(0)}\right), \qquad \text{with} \quad \overline{\mathbf{y}}^{(0)} = \sum_{t=1}^{T}(\delta_t \circ \widetilde{\mathbf{y}}_t)/\sum_{t=1}^{T}\delta_t,$$

($\circ$ denotes the Hadamard product; the ratio between the two vectors is intended component-wise.)
end for
Fix γ to a small given value (or fix $max.iter = 30$) and $s = 0$
repeat
$s = s + 1$
Estimate $(\widehat{\boldsymbol{\lambda}}_0^{(s-1)}, \widehat{\boldsymbol{\lambda}}_1^{(s-1)}, \widehat{\boldsymbol{\lambda}}_2^{(s-1)})$ as in Eq. (3), using the centered data $\{\mathbf{y}_1^{(s-1)}, \ldots, \mathbf{y}_T^{(s-1)}\}$;
for $t = 1, \ldots, T$ **do**

$$\widehat{\mathbf{y}}_t^{(s)} = D(\widehat{\boldsymbol{\lambda}}_0^{(s-1)})\mathbf{W}\,\mathbf{y}_t^{(s-1)} + D(\widehat{\boldsymbol{\lambda}}_1^{(s-1)})\mathbf{y}_{t-1}^{(s-1)} + D(\widehat{\boldsymbol{\lambda}}_2^{(s-1)})\mathbf{W}\,\mathbf{y}_{t-1}^{(s-1)} \tag{4}$$

$$\overline{\mathbf{y}}^{(s)} = \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{\delta}_t \circ \widetilde{\mathbf{y}}_t + (\mathbf{1} - \boldsymbol{\delta}_t) \circ (\widehat{\mathbf{y}}_t^{(s)} + \overline{\mathbf{y}}^{(s-1)})\right) \tag{5}$$

$$\mathbf{y}_t^{(s)} = \boldsymbol{\delta}_t \circ (\widetilde{\mathbf{y}}_t - \overline{\mathbf{y}}^{(s)}) + (\mathbf{1} - \boldsymbol{\delta}_t) \circ \widehat{\mathbf{y}}_t^{(s)}, \tag{6}$$

($\mathbf{1}$ is a vector of ones)
end for
until $\|\mathbf{y}_t^{(s)} - \mathbf{y}_t^{(s-1)}\|_2^2 \leq \gamma$ or $s \geq max.iter$
Reconstruct the multivariate time series as
for t=1,2,…,T **do**

$$\widetilde{\mathbf{y}}_t^{(s)} = \mathbf{y}_t^{(s)} + \overline{\mathbf{y}}^{(s)}$$

(the original missing data replaced by the estimated values.)
end for

The imputation procedure is described in details in Algorithm 1.

Note that the procedure is able to reconstruct both isolated missing values and (possibly long) sequences of missing values. In doing so, the procedure uses both the time dependence structure in each time series and the cross-dependency among time series making the procedure very effective, assuming that the spatio-temporal parametric model is correctly assumed. As an additional remark, note that the model estimation procedure is based on an estimator available in closed form making the overall procedure very convenient from a computational point of view. Therefore, our estimation procedure can still be used efficiently when the number of spatial locations is very large. In fact, note that matrix $\widehat{\mathbf{X}}_i'\widehat{\mathbf{X}}_i$ in Eq. (3) is always of order 3×3, whatever the total number of spatial locations is. So, the matrix inversion can be done very quickly. Hence, the estimation procedure requires a loop to estimate the spatial parameters for all spatial units (for $i = 1, \ldots, p$, where p is the total number of units). This loop can be easily parallelized, in case p is extremely large. This latter aspect is of great importance as we use the bootstrap to approximate the confidence bands, which itself requires additional computational burden.

3 Bootstrap joint prediction regions for missing value paths

Given the observed spatio-temporal series $\mathbf{y}_1, \cdots, \mathbf{y}_T$, we assume there can be several missing values and/or missing sequences appearing here and there in the multivariate series, at different locations and/or different time intervals. Unlike the previous section, where we used a matrix/vector notation to refer to the multivariate series, in this section we prefer to focus on a single univariate missing sequence (or missing value), in order to keep notation simple as we explain how to derive the JPR for this sequence. Of course, the same arguments must be repeated for all the missing values/sequences in the spatio-temporal series.

So, let $y_{t,i}, y_{t+1,i}, \ldots, y_{t+H_t^i-1,i}$ denote a generic sequence of H_t^i missing values, starting at time t for a given site i. Here and in the following we refer to a missing sequence of length H_t^i, but note that this notation includes the special case of isolated missing values when $H_t^i = 1$.

Let $\widehat{y}_{t,i}, \widehat{y}_{t+1,i}, \ldots, \widehat{y}_{t+H_t^i-1,i}$ be the sequence of predicted values using the procedure of the previous section, based on the multivariate data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$. Since the predictor $y_{t+h,i}$ is a function of the data and of the model with parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ we write it as $\widehat{y}_{t+h,i} = g_h(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T, \widehat{\boldsymbol{\theta}})$, with $h = 0, 1, \ldots, H_t^i - 1$.

The aim is to construct a joint prediction region (JPR) that contains the missing path of interest with probability $1 - \alpha$, at least asymptotically. Following Wolf et al. (2015), define the family-wise error rate as

$$\text{FWE} = \Pr\Big(\text{at least one of } y_{t+h,i},\, h = 0, \ldots, H_t^i - 1, \text{ is not contained in JPR}\Big).$$

When the length of the missing sequence H_t^i is large, control of the FWE may be deemed too strict as it would cause very large (and therefore uninformative) JPRs. In such a case, we suggest to use the generalized family-wise error rate (k-FWE) defined as

$$k\text{-FWE} = \Pr\Big(\text{at least } k \text{ of } y_{t+h,i},\, h = 0, \ldots, H_t^i - 1, \text{ are not contained in JPR}\Big),$$

with $k < H_t^i$, providing the applied researcher with an alternative tool in order to relax the control of the FWE.

In the next sections, we describe three different methods to derive the JPR for a missing sequence, following the k-FWE criteria: the Maximum Predictive Root method (MPR), the normal Bootstrap method (NB) and the Percentile method (PER). Our main proposal is the MPR method, while the other two methods are included for the sake of comparison, and used as much simpler alternative techniques that are expected to work reasonably well in standard scenarios. However, it is worth pointing out that all three methods have been proposed here for the first time, and all of them have interesting peculiarities and good empirical performance, as will be shown in the simulation study.

3.1 Joint prediction regions with the Maximum Predictive Root method

Let $y_{t+h,i} - \widehat{y}_{t+h,i}$, $h = 0, 1, \ldots, H_t^i - 1$, be the sequence of prediction errors. The Maximum Predictive Root (MPR) method is based on the statistic M_{k,H_t^i} defined as

$$M_{k,H_t^i} = k\text{-max}_{h=0,1,\ldots,H_t^i-1}\left(\left|y_{t+h,i} - \widehat{y}_{t+h,i}\right|\right),$$

where, as in Wolf et al. (2015), we define $k\text{- max}_{h=0,1,\ldots,H_t^i-1}(x_h)$ as the function that returns the k-th largest value of the set $\{x_0, \ldots, x_{H_t^i-1}\}$. Then a two-sided JPR for the missing sequence, that controls the k-FWE in finite samples, is given by:

$$\left[\widehat{y}_{t,i} \pm q^{i,t,k}_{(1-\alpha)}\right] \times \left[\widehat{y}_{t+1,i} \pm q^{i,t,k}_{(1-\alpha)}\right] \times \cdots \times \left[\widehat{y}_{t+H_t^i-1,i} \pm q^{i,t,k}_{(1-\alpha)}\right] \tag{7}$$

where $q^{i,t,k}_{(1-\alpha)}$ is the $1-\alpha$ quantile of the distribution of M_{k,H_t^i}. The implication is that the probability that the previous region will contain at least $H_t^i - k + 1$ elements of the missing sequence is equal to (at least) $1 - \alpha$ asymptotically.

Clearly, the JPR defined in Eq. (7) is not useful since the quantile $q^{i,t,k}_{(1-\alpha)}$ is unknown. However, it can be estimated by using some resampling scheme. Given a bootstrap pseudo series $\mathbf{y}_1^*, \mathbf{y}_2^* \ldots, \mathbf{y}_T^*$ the bootstrap counterparts of the aforementioned quantities can be defined as $y^*_{t+h,i} - \widehat{y}^*_{t+h,i}$ with $\widehat{y}^*_{t+h,i} = g_h(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_T^*, \widehat{\boldsymbol{\theta}}^*)$. Let $\widehat{q}^{i,t,k}_{(1-\alpha)}$ be the $1-\alpha$ quantile of the distribution of the statistic $M^*_{k,H_t^i} = k\text{- max}_{h=0,1,\ldots,H_t^i-1}(|y^*_{t+h,i} - \widehat{y}^*_{t+h,i}|)$. A two-sided JPR for $y_{t,i}, y_{t+1,i}, \ldots, y_{t+H_t^i-1,i}$ that controls the k-FWE in finite samples is given by:

$$\left[\widehat{y}_{t,i} \pm \widehat{q}^{i,t,k}_{(1-\alpha)}\right] \times \left[\widehat{y}_{t+1,i} \pm \widehat{q}^{i,t,k}_{(1-\alpha)}\right] \times \cdots \times \left[\widehat{y}_{t+H_t^i-1,i} \pm \widehat{q}^{i,t,k}_{(1-\alpha)}\right]. \tag{8}$$

Details are given in Algorithm 2.

The JPRs derived by the MPR method have two important advantages: first, they are proven to be asymptotically consistent under a realistic, mild high-level assumption. Second, they enjoy superior finite-sample properties, as demonstrated via extensive Monte Carlo simulations in Wolf et al. (2015). Algorithm 2 assumes a generic bootstrap method to generate the bootstrap sample. In the following we use a resampling procedure based on the residual bootstrap approach. The procedure can be implemented as detailed in Algorithm 3. The theoretical properties of the residual bootstrap scheme for time series can be derived following Choi et al. (2000).

3.2 Joint prediction regions by combining bootstrap resampling with a generalised Bonferroni technique

In order to present some alternative approaches to derive the JPRs for the missing sequences, in this section we also suggest to build them by stringing together marginal confidence intervals derived by combining a bootstrap resampling technique with a generalised version of the Bonferroni correction. Unlike the MPR method used in the

Algorithm 2 Pseudocode for joint predictive bands

Define $k\text{-}\max_{h=0,1,\ldots,H_t^i-1}(x_h)$ as the function that returns the k-th largest value of the set $\{x_0,\ldots,x_{H_t^i-1}\}$.

Define $\widetilde{\alpha}_k^{(i,t)} = \max\left\{a \in (0,1) : Pr(A_a^{H_t^i} \leq k) \geq 1-\alpha\right\}$, with $A_a^{H_t^i} \sim Binom(H_t^i, a)$.

For a given site i, let $\widehat{y}_{t,i}, \widehat{y}_{t+1,i}, \ldots, \widehat{y}_{t+H_t^i-1,i}$ be the sequence of predictions for the missing values, where $\widehat{y}_{t+h,i} = g_h(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T, \widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\lambda}}_0, \widehat{\boldsymbol{\lambda}}_1, \widehat{\boldsymbol{\lambda}}_2)$ (see Algorithm 1).

Fix $B \geq 999$, the number of bootstrap replicates, and $k < H_t^i$.

for $b = 1, \ldots, B$ **do**

Generate a bootstrap sample $\mathbf{y}_1^{*,b}, \mathbf{y}_2^{*,b}, \ldots, \mathbf{y}_T^{*,b}$ using Algorithm 3.

Compute $\widehat{y}_{t+h,i}^{*} = g_h(\mathbf{y}_1^{*,b}, \mathbf{y}_2^{*,b}, \ldots, \mathbf{y}_T^{*,b}, \widehat{\boldsymbol{\theta}}^{*,b})$ for $h = 0, 1, \ldots, H_t^i - 1$.

Compute $S_h^{*,b} = y_{t+h,i}^{*,b} - \widehat{y}_{t+h,i}^{*,b}$ for $h = 0, 1, \ldots, H_t^i - 1$.

Let $M_{k,H_t^i}^{*,b} = k\text{-}\max_{h=0,1,\ldots,H_t^i-1}\left(|S_h^{*,b}|\right)$.

end for

1) For the MPR joint prediction region (see Sect. 3.1):

- Compute the (1-α)-quantile of the empirical distribution of $M_{k,H_t^i}^{*,b}$, $b = 1, 2, \ldots, B$, as $\widehat{q}_{(1-\alpha)}^{i,t,k}$.
- Construct the the H_t^i intervals, for $h = 1, 2 \ldots, H_t^i$,

$$\left[\widehat{y}_{t+h,i} - \widehat{q}_{(1-\alpha)}^{i,t,k}, \widehat{y}_{t+h,i} + \widehat{q}_{(1-\alpha)}^{i,t,k}\right]. \tag{9}$$

2) For the NB joint prediction region (see Sect. 3.2):

- Compute the standard deviation sd_h^B of the empirical distribution of $S_h^{*,b}$, $b = 1, 2, \ldots, B$, for $h = 1, \ldots, H_t^i$.
- Using the quantiles of the standard normal density, z_γ, construct the H_t^i intervals, for $h = 1, 2 \ldots, H_t^i$,

$$\left[\widehat{y}_{t+h,i} + z_{\widetilde{\alpha}_k^{(i,t)}/2} sd_h^B, \widehat{y}_{t+h,i} + z_{1-\widetilde{\alpha}_k^{(i,t)}/2} sd_h^B\right]. \tag{10}$$

3) For the PER joint prediction region (see Sect. 3.2):

- Define the (γ)-quantile of the empirical distribution of $S_h^{*,b}$, $b = 1, 2, \ldots, B$, for $h = 1, \ldots, H_t^i$, as $\widetilde{q}_h(\gamma)$
- Construct the H_t^i intervals, for $h = 1, 2 \ldots, H_t^i$,

$$\left[\widehat{y}_{t+h,i} + \widetilde{q}_h(\widetilde{\alpha}_k^{(i,t)}/2), \widehat{y}_{t+h,i} + \widetilde{q}_h(1 - \widetilde{\alpha}_k^{(i,t)}/2)\right]. \tag{11}$$

Output: the JPRs in (9), (10) and (11).

previous section, the Bonferroni method is accused to be a rough method to derive joint confidence regions, since it does not take into account the dependence among the marginal confidence intervals and therefore may lead to more conservative joint confidence regions. However, to take advantage of the simplicity of this approach, in this paper we generalise the idea of the Bonferroni correction by adapting it to the k-FWE criteria, so that we can derive more flexible joint confidence regions that are substantially comparable with those based on the MPR method.

Algorithm 3 Pseudocode for spatio-temporal residual bootstrap

By Algorithm 1, compute the residuals $\widehat{\mathbf{e}}_t^{(s)} = \mathbf{y}_t^{(s)} - \widehat{\mathbf{y}}_t^{(s)}$, where $\mathbf{y}_t^{(s)}$ is computed by Eq. (6) and $\widehat{\mathbf{y}}_t^{(s)}$ is computed by Eq. (4). The value for the index s is taken from the last iteration of the imputation procedure described in Algorithm 1.
Compute the centered residuals as $\hat{\epsilon}_t^{(s)} = \widehat{\mathbf{e}}_t^{(s)} - \bar{\widehat{\mathbf{e}}}_T^{(s)}$.
Fix an integer value $c \geq 100$.
Obtain the bootstrap error series ϵ_t^*, with $t = -c+1, \ldots, -1, 0, 1, \ldots, T$ by drawing $T + c$ samples independently and uniformly, with replacement, from the centered residuals $\hat{\epsilon}_t^{(s)}$.
Set the starting value $\mathbf{y}_{-c}^* = \mathbf{0}$.
for $t = -c+1, \ldots, 0, 1, 2, \ldots, T$ **do**
Generate the bootstrap value $\mathbf{y}_t^*$ as

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

end for
Output: the bootstrap sample $\mathbf{y}_1^*, \ldots, \mathbf{y}_T^*$.

So, if the problem is to guarantee that the JPR will contain at least $H_t^i - k + 1$ elements of the missing sequence ($1 \leq k < H_t^i$) with global probability equal to $1-\alpha$, asymptotically, then suitable individual marginal sizes $\widetilde{\alpha}_k^{(i,t)}$ must be set. Generalizing the idea of the Bonferroni multiple scheme, we set this value equal to

$$\widetilde{\alpha}_k^{(i,t)} = \max\left\{a \in (0,1) : Pr(A_a^{H_t^i} \leq k) \geq 1-\alpha\right\} \text{ with } A_a^{H_t^i} \sim Binom(H_t^i, a).$$

Then, the individual confidence intervals with nominal confidence level $(1-\widetilde{\alpha}_k^{(i,t)})$ are derived using some bootstrap method, for each h-th value of the missing sequence. In particular, we use the normal bootstrap method (i.e. the normal approximation with bootstrap estimated standard errors) and the percentile boostrap method. The final JPR derived with these two bootstrap procedures are denoted with NB and PER, respectively.

In particular, the percentile (PER) joint bootstrap prediction intervals are given similarly as in (8) in which $\widehat{q}_{(\alpha)}^{i,t,k}$ is replaced by $\widetilde{q}_h(\widetilde{\alpha}_k^{(i,t)})$ for $h = 1, \ldots, H_t^i$, that is the $\widetilde{\alpha}_k^{(i,t)}$-quantile of the bootstrap distribution of the statistic

$$S_h^* = y_{t+h,i}^* - \widehat{y}_{t+h,i}^*.$$

Therefore, the JPR derived by the PER method is given by

$$\left[\widehat{y}_{t+h,i} + \widetilde{q}_h(\widetilde{\alpha}_k^{(i,t)}/2), \widehat{y}_{t+h,i} + \widetilde{q}_h(1-\widetilde{\alpha}_k^{(i,t)}/2)\right].$$

Finally, the normal bootstrap (NB) joint prediction intervals are derived assuming normality for the error process. For $h = 1, \ldots, H_t^i$, we have

$$\left[\widehat{y}_{t+h,i} - z_{1-\widetilde{\alpha}_k^{(i,t)}/2} sd_B(S_h^*), \widehat{y}_{t+h,i} + z_{1-\widetilde{\alpha}_k^{(i,t)}/2} sd_B(S_h^*)\right],$$

where z_γ is the γ-th percentile of the standard normal distribution and $sd_B(S_h^*)$ is the bootstrap estimated standard deviation for S_h^*.

The procedure for the NB and PER joint prediction regions can be implemented as detailed in Algorithm 2.

4 A Monte Carlo study

To validate the empirical performance of the proposed JPR in Eq. (8), we have implemented a Monte Carlo simulation study. The aim is to compare empirically the coverages and the mean lengths of the regions obtained using the normal-based method, the percentile and the Maximum Predictive Root method.

We have considered multivariate time series of dimension $p = 30$ and lengths $T = 100, 500$ and 1000. The weight matrix $\mathbf{W}$ 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 $[-0.9, 0.9]$. The error component $\boldsymbol{\varepsilon}_t$ has been generated from two different multivariate distributions. The first one is a multivariate normal distribution, with mean vector zero and diagonal variance-covariance matrix, with heteroscedastic variances $(\sigma_1^2, \ldots, \sigma_p^2)$. In particular, the standard deviations $(\sigma_1, \ldots, \sigma_p)$ have been generated randomly from a Uniform distribution $U(0.5; 1.5)$. The second one is the multivariate distribution in which the marginal distributions are pairwise independent student-t distribution with 6 degrees of freedom. Note that, in the case of normal distributed error term, the consistency of the normal based bootstrap confidence intervals is guaranteed.

In the simulation experiment both isolated missing values and missing sequences have been considered. One missing sequences with length H has been placed at location 2 and for it three different time horizons (missing sequence length) have been considered $H = 5, 10$ and 20. The number of the isolated missing values has been fixed at 10 and they have been randomly generated at other locations.

In the experiment, for all the three methods, we have considered the $k-$JPR that is the estimated confidence region which contains at least $H - k + 1$ elements. We have fixed $k = 1, 2$ and 3.

All bootstrap estimates have been computed by using $B = 999$ replicates and $N = 1000$ Monte Carlo runs.

In Table 1 the mean lengths of the joint $k-$JPRs ($k = 1, 2, 3$) by means of the normal-based (NB) method, the percentile (PER) and the Maximum Predictive Root (MPR) methods are listed for all the considered values of T and H and for both the confidence levels $1 - \alpha = 0.95$ and 0.90. In this case the distribution of the error term is the standard Gaussian.

The three methods present similar performance in terms of mean length. As expected, the amplitude of the regions increases as H increases. The results of Table 1 are confirmed from Table 2, in which the empirical coverages are shown for the same choices of Table 1. In order to evaluate if the coverages are significantly different with respect to the nominal level, we have calculated the asymptotic acceptance confidence interval at 99% . It is defined as $p_0 \pm 2.33\sqrt{\frac{p_0(1-p_0)}{N}}$ where p_0 is fixed to the nomi-

Table 1 Mean lengths of the $k-$JPRs ($k = 1, 2, 3$) with the normal-based (NB) method, the percentile (PER) and the Maximum Predictive Root (MPR) method in the case of normal distributed error term

		$1-\alpha = 0.95$			$1-\alpha = 0.90$		
		H=5	H=10	H=20	H=5	H=10	H=20
$T = 100$							
$k = 1$	NB	3.80	4.00	4.08	3.45	3.70	3.77
	PER	3.77	3.90	3.95	3.47	3.73	3.82
	MPR	3.88	4.15	4.25	3.50	3.84	4.08
$k = 2$	NB	2.65	3.04	3.22	2.37	2.79	3.04
	PER	2.68	3.11	3.42	2.39	2.86	3.22
	MPR	2.69	3.14	3.48	2.40	2.88	3.25
$k = 3$	NB	1.96	2.48	2.78	1.73	2.29	2.61
	PER	1.97	2.54	2.94	1.73	2.33	2.75
	MPR	1.98	2.55	2.97	1.73	2.34	2.77
$T = 500$							
$k = 1$	NB	3.90	4.19	4.48	3.54	3.88	4.15
	PER	3.86	4.13	4.36	3.52	3.85	4.11
	MPR	3.95	4.30	4.58	3.55	3.93	4.28
$k = 2$	NB	2.71	3.18	3.54	2.42	2.93	3.34
	PER	2.71	3.19	3.56	2.43	2.93	3.36
	MPR	2.72	3.20	3.61	2.44	2.95	3.38
$k = 3$	NB	2.01	2.60	3.05	1.77	2.40	2.87
	PER	2.01	2.61	3.08	1.77	2.40	2.89
	MPR	2.01	2.62	3.10	1.78	2.41	2.91
$T = 1000$							
$k = 1$	NB	3.91	4.21	4.52	3.55	3.89	4.19
	PER	3.87	4.13	4.39	3.53	3.86	4.12
	MPR	3.96	4.29	4.61	3.56	3.93	4.27
$k = 2$	NB	2.72	3.19	3.57	2.44	2.94	3.37
	PER	2.71	3.18	3.56	2.43	2.93	3.37
	MPR	2.72	3.20	3.62	2.44	2.95	3.39
$k = 3$	NB	2.02	2.61	3.08	1.78	2.41	2.90
	PER	2.01	2.61	3.08	1.78	2.40	2.90
	MPR	2.02	2.62	3.11	1.78	2.41	2.92

H denotes the length of the missing sequence and T the time series length. The confidence levels are fixed to 0.95 and 0.90

nal level and N is the number of Monte Carlo runs. Observed coverages inside such interval can be considered not different from the fixed nominal level. In Table 2 these values are reported in bold.

The three methods present similar empirical coverages, by varying H and by varying α. When $H = 10$, the NB and MPR method have similar performance and generally

Table 2 Empirical coverages of the $k-$JPR ($k = 1, 2, 3$) with the normal-based (NB) method, the percentile (PER) and the Maximum Predictive Root (MPR) method in the case of normal distributed error term

		$1-\alpha = 0.95$			$1-\alpha = 0.90$		
		H=5	H=10	H=20	H=5	H=10	H=20
$T = 100$							
$k = 1$	NB	0.910	0.871	0.77	0.849	0.799	0.667
	PER	0.875	0.816	0.674	0.833	0.764	0.625
	MPR	0.911	0.871	0.798	0.851	0.817	0.746
$k = 2$	NB	0.905	0.889	0.760	0.837	0.803	0.658
	PER	0.909	0.892	0.810	0.837	0.817	0.741
	MPR	0.916	0.910	0.846	0.848	0.841	0.764
$k = 3$	NP	0.912	0.873	0.745	0.862	0.788	0.637
	PER	0.910	0.883	0.804	0.855	0.796	0.711
	MPR	0.915	0.884	0.830	0.856	0.814	0.739
$T = 500$							
$k = 1$	NB	**0.942**	0.907	0.894	0.874	0.852	0.816
	PER	0.924	0.883	0.844	0.873	0.814	0.771
	MPR	**0.940**	0.915	0.894	**0.883**	0.852	0.825
$k = 2$	NB	0.930	0.922	0.881	0.875	0.842	0.82
	PER	0.926	0.914	0.886	0.873	0.842	0.818
	MPR	0.930	0.927	0.907	0.871	0.848	0.842
$k = 3$	NB	0.927	0.928	0.893	0.869	0.855	0.827
	PER	0.928	0.927	0.902	0.859	0.859	0.829
	MPR	0.924	**0.936**	0.912	0.865	0.857	0.847
$T = 1000$							
$k = 1$	NB	**0.943**	0.930	0.914	**0.905**	0.868	0.855
	PER	0.933	0.898	0.868	**0.893**	0.847	0.813
	MPR	**0.947**	**0.934**	0.920	**0.905**	0.874	0.875
$k = 2$	NB	**0.945**	0.927	0.920	**0.889**	0.862	0.858
	PER	**0.942**	0.918	0.916	**0.892**	0.856	0.861
	MPR	**0.947**	0.926	0.928	**0.889**	0.865	0.859
$k = 3$	NB	**0.950**	0.925	0.928	**0.890**	0.861	0.859
	PER	**0.947**	0.924	0.916	**0.896**	0.867	0.856
	MPR	**0.951**	0.930	0.929	**0.895**	0.870	0.868

In parenthesis the MAD; H denotes the length of the missing sequence and T the time series length. The confidence levels are fixed to 0.95 and 0.90. In bold the observed coverages inside the asymptotic acceptance interval at 99%

better than the PER method, for all the values of k and for both the considered confidence levels.

By increasing H, the coverages are worse and many of them are outside the acceptance region, with levels that are far from the nominal coverage. However, the MPR method presents better performances with respect to the other two methods. When k

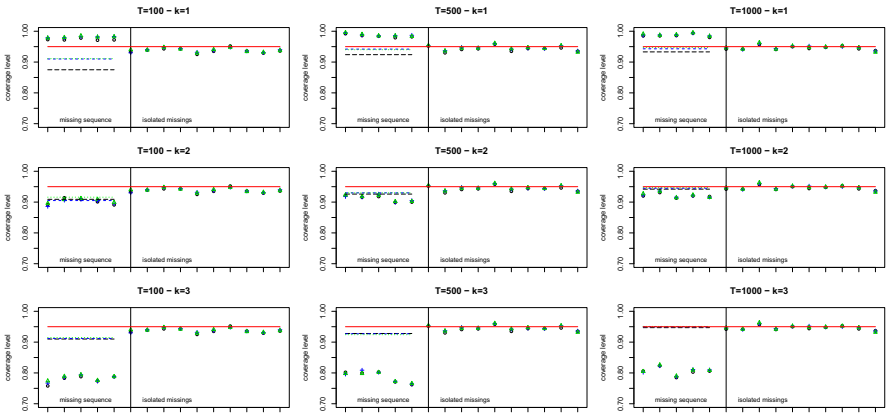

Fig. 1 Empirical coverages of the joint $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of normal distributed error term and in the presence of 15 missing values. The first 5 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missings. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

is fixed, the lengths of the regions obtained by the three methods seem to be basically comparable.

Figure 1 refers to the case $H = 5$ (the total number of missing values is 15). The three time series lengths $T = 100, 500$ and $T = 1000$ are considered (left, center and right respectively) and $k = 1, 2$ and 3 (from the top to the bottom). All the considered methods are consistent, since by increasing T and fixing k, the coverage level becomes closer and closer to the nominal one 0.95. The NB and the MPR methods seem to have similar performance. For $k = 1$ the PER method has worse performances with respect to the others since the coverage is lower for all values of T. For $k = 2$ and $k = 3$ also the PER method seems equivalent to the others. By looking at the univariate confidence intervals, the observed coverage for all the three methods is more or less similar. They reach the nominal level for the isolated missing values, already for $T = 100$ and $k = 1$.

Figure 2 refers to the case $H = 10$ (the total number of missing values is 20). The results seem to be quite similar to Fig. 1, showing a better performance of the MPR method with respect to the others, expecially for $k = 2$ and $T = 100$. Also here, the PER method presents worse performances, which however improve for $k = 2$ and $k = 3$ and become comparable with the other methods. For the univariate coverages, the estimated intervals for the isolated missing values present a coverage close to the nominal one. For the sequence of missing values, empirical coverages of the univariate intervals must not be compared with the nominal coverage $1-\alpha = 0.95$, which refers to the nominal global coverage of the whole JPR. Therefore, univariate coverages of the marginal intervals for the missing sequence must necessarily be greater than $1-\alpha$, but can be significantly reduced when $k > 1$.

Figure 3 refers to the case $H = 20$ (the total number of missing values is 30). In this last case the MPR method outperforms the other ones.

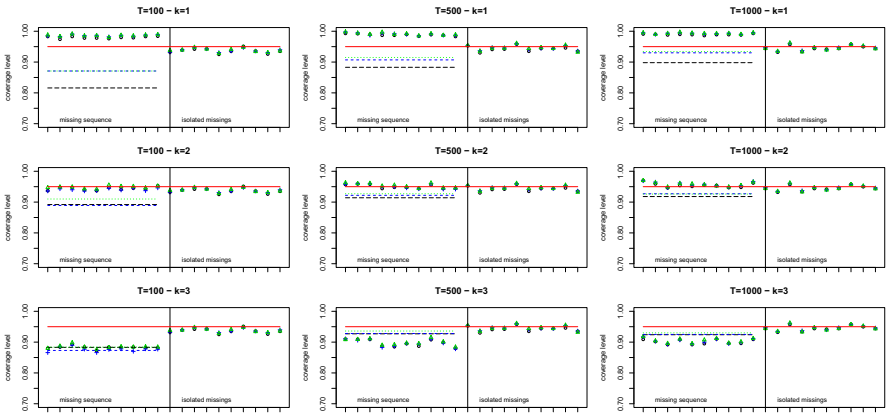

Fig. 2 Empirical coverages of the joint $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of normal distributed error term and in the presence of 20 missing values. The first 10 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missing values. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

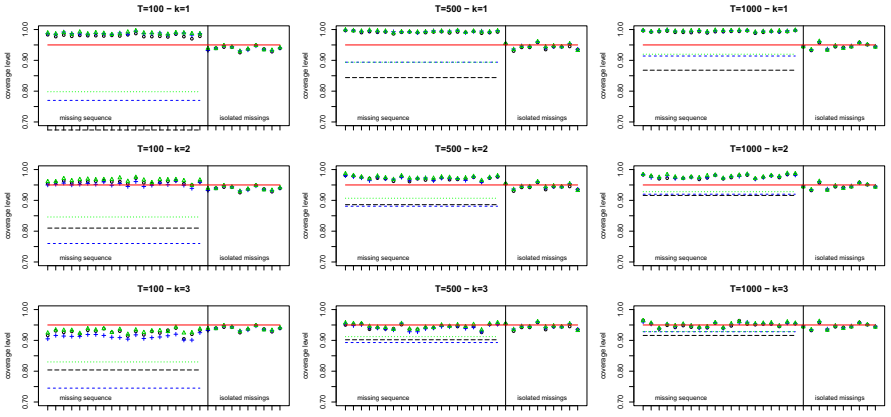

Fig. 3 Empirical coverages of the joint $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of normal distributed error term and in the presence of 30 missing values. The first 20 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missing values. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

In Table 3 the mean lengths of the joint $k-$JPRs ($k = 1, 2, 3$) when the error term is $t(6)$-distributed are shown for all the considered values of T and H. Also in this case, the three methods present similar behaviours.

In Table 4 the empirical coverages are reported for the same values of Table 2 when the error term is t-distributed. In general, the NB and PER methods present worse performances, especially for longer missing sequences. Moreover, the PER method seems to suffer more instability for the nominal coverage correction, especially

Table 3 Mean lengths of the $k-$JPRs ($k = 1, 2, 3$) with the normal-based (NB) method, the percentile (PER) and the Maximum Predictive Root (MPR) method in the case of $t_{(6)}$-distributed error term

		$1-\alpha = 0.95$			$1-\alpha = 0.90$		
		H=5	H=10	H=20	H=5	H=10	H=20
$T = 100$							
$k = 1$	NB	5.97	6.28	6.39	5.42	5.82	5.92
	PER	6.88	7.15	7.12	6.04	6.80	6.91
	MPR	7.10	8.06	8.05	5.99	6.91	7.79
$k = 2$	NB	4.15	4.77	5.05	3.72	4.39	4.76
	PER	4.155	5.07	5.92	3.63	4.53	5.33
	MPR	4.17	5.09	5.93	3.64	4.55	5.33
$k = 3$	NB	3.08	3.90	4.36	2.72	3.59	4.10
	PER	2.91	3.90	4.69	2.52	3.52	4.31
	MPR	2.92	3.92	4.72	2.54	3.53	4.33
$T = 500$							
$k = 1$	NB	6.16	6.63	7.09	5.59	6.13	6.57
	PER	7.07	8.01	8.95	6.10	7.07	7.99
	MPR	7.27	8.46	9.59	6.15	7.23	8.41
$k = 2$	NB	4.28	5.03	5.60	3.84	4.63	5.28
	PER	4.22	5.26	6.21	3.68	4.69	5.69
	MPR	4.24	5.29	6.34	3.69	4.72	5.73
$k = 3$	NB	3.17	4.11	4.84	2.80	3.79	4.55
	PER	2.94	4.02	5.02	2.55	3.63	4.61
	MPR	2.95	4.04	5.06	2.56	3.64	4.64
$T = 1000$							
$k = 1$	NB	6.18	6.66	7.16	5.61	6.17	6.63
	PER	7.07	7.99	9.08	6.09	7.06	7.97
	MPR	7.28	8.48	9.78	6.16	7.24	8.42
$k = 2$	NB	4.30	5.06	5.66	3.85	4.65	5.34
	PER	4.23	5.27	6.22	3.69	4.70	5.71
	MPR	4.25	5.30	6.35	3.70	4.72	5.75
$k = 3$	NB	3.19	4.14	4.88	2.81	3.81	4.59
	PER	2.94	4.03	5.03	2.55	3.64	4.64
	MPR	2.95	4.05	5.09	2.56	3.65	4.67

H denotes the length of the missing sequence and T the time series length. The confidence levels are fixed to 0.95 and 0.90

when moving deep into the tails. Apparently, a much greater number of Monte Carlo replicates are needed for an accurate estimate of the bootstrap percentiles used for the construction of the JPRs.

Figure 4 refers to the case $H = 5$ (the total number of missing values is 15). As in the case of normal errors, all the considered methods are consistent and again the MPR method outperforms all the others. However, for $k = 1$ and for all values of T, the NB method has worse performance, while for $k = 2$ and $k = 3$ the three methods

Table 4 Empirical coverages of the $k-$JPRs ($k = 1, 2, 3$) with the normal-based (NB) method, the percentile (PER) and the Maximum Predictive Root (MPR) method in the case of t(6)-distributed error term

		$1-\alpha = 0.95$			$1-\alpha = 0.90$		
		H=5	H=10	H=20	H=5	H=10	H=20
$T = 100$							
$k = 1$	NB	0.853	0.773	0.618	0.803	0.718	0.513
	PER	0.871	0.792	0.633	0.822	0.763	0.588
	MPR	0.908	0.873	0.776	0.836	0.81	0.744
$k = 2$	NB	0.929	0.858	0.688	**0.884**	0.785	0.606
	PER	0.926	0.873	0.783	0.867	0.792	0.695
	MPR	0.933	0.885	0.833	0.869	0.822	0.732
$k = 3$	NB	**0.938**	0.897	0.740	**0.900**	0.840	0.636
	PER	0.924	0.898	0.793	0.861	0.812	0.7
	MPR	0.928	0.894	0.825	0.865	0.822	0.727
$T = 500$							
$k = 1$	NB	0.894	0.846	0.780	0.848	0.793	0.711
	PER	0.923	0.906	0.871	0.874	0.848	0.819
	MPR	**0.944**	0.927	0.927	**0.892**	**0.884**	**0.880**
$k = 2$	NB	**0.942**	0.924	0.872	**0.902**	0.876	0.819
	PER	**0.943**	**0.935**	0.920	**0.883**	0.870	0.864
	MPR	**0.938**	**0.946**	0.931	**0.892**	**0.888**	**0.882**
$k = 3$	NB	**0.956**	**0.944**	0.909	**0.920**	**0.904**	0.867
	PER	0.931	0.929	0.922	0.867	0.877	0.867
	MPR	0.930	0.932	0.933	**0.878**	**0.878**	**0.884**
$T = 1000$							
$k = 1$	NB	0.898	0.865	0.798	0.850	0.815	0.718
	PER	**0.939**	0.915	0.882	**0.891**	**0.879**	0.824
	MPR	**0.954**	**0.953**	**0.946**	**0.889**	**0.904**	**0.893**
$k = 2$	NB	**0.959**	0.927	0.862	**0.920**	**0.884**	0.822
	PER	**0.956**	**0.940**	0.919	**0.900**	**0.883**	0.871
	MPR	**0.955**	**0.941**	**0.951**	**0.907**	**0.889**	**0.887**
$k = 3$	NB	**0.965**	**0.949**	0.912	0.940	**0.920**	0.867
	PER	**0.944**	0.933	0.930	**0.895**	**0.882**	0.866
	MPR	**0.950**	**0.942**	**0.942**	**0.900**	**0.889**	0.873

H denotes the length of the missing sequence and T the time series length. The confidence levels are fixed to 0.95 and 0.90. In bold the observed coverages inside the asymptotic acceptance interval at 99%

seem to be equivalent. By looking at the univariate confidence intervals, the observed coverages for all the three methods are quite similar, both for isolated and sequence of missing values. In particular, in all the cases the coverages for the isolated missing values are close to 0.95 for all T and for all k.

Figure 5 refers to the case $H = 10$ (the total number of missing values is 20) with $t_{(6)}-$student errors. The results seem to be quite similar to Fig. 4, showing even more clearly a better performance of the MPR method with respect to the others, expecially

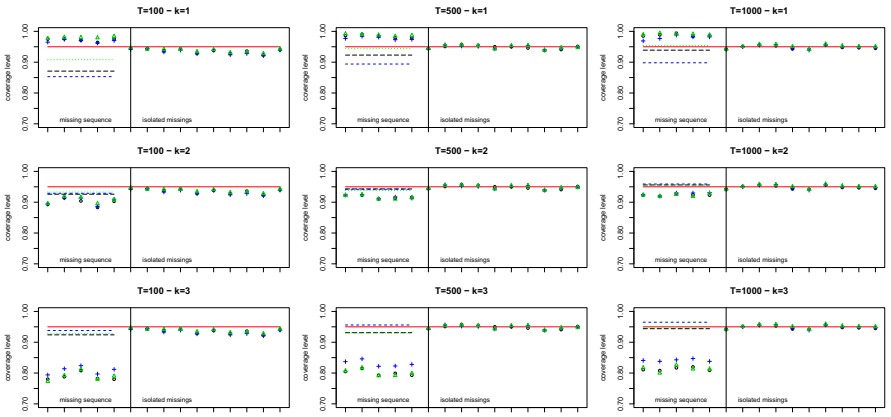

Fig. 4 Empirical coverages of the $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of $t_{(6)}-$distributed error term and in the presence of 15 missing values. The first 5 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missing values. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

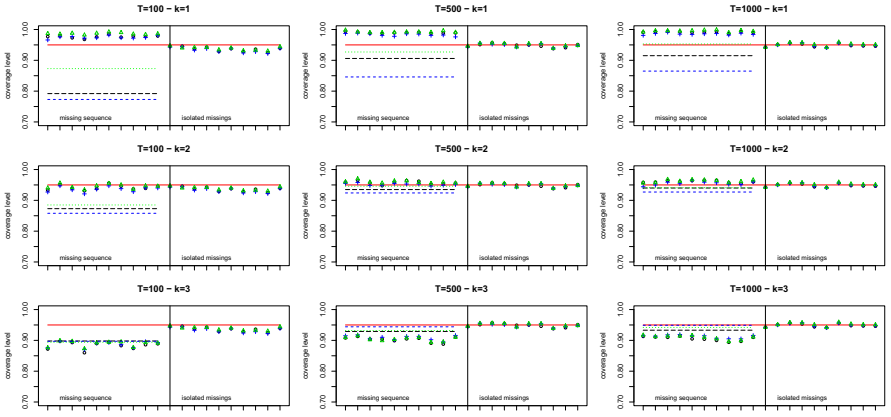

Fig. 5 Empirical coverages of the $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of $t_{(6)}-$distributed error term and in the presence of 20 missing values. The first 10 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missing. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

for $k = 1$ and $T = 100, 500$. Also here, the NB method presents worse performances for $k = 1$, which however improve for $k = 2$ and $k = 3$. Again, for the univariate case, the estimated intervals for the isolated missing values present a coverage close to the nominal one.

Figure 6 refers to $H = 20$ (the total number of missing values is 30). In this last case the MPR method sharply outperforms the other ones mainly for the time series length $T = 100$ and 500. For the univariate case, as in the previous analysis, the estimated intervals for the isolated missing values present a coverage close to the nominal one.

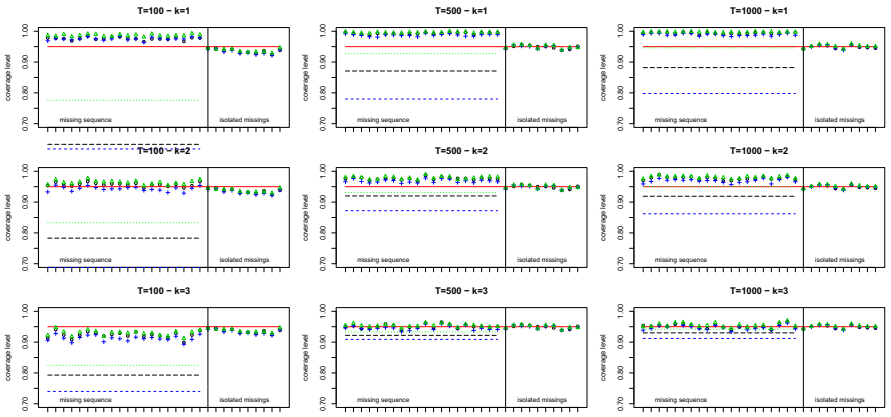

Fig. 6 Empirical coverages of the $k-$JPRs ($k = 1, 2, 3$) with the NB (blue dashed line), PER (black dashed line) and MPR (green dashed line) methods in the case of $t_{(6)}-$distributed error term and in the presence of 30 missing values. The first 20 values, on the left of the vertical line, represent the missing sequence, while the remaining values are isolated missing values. The nominal level is fixed to 0.95 (red line). The empirical coverages for the univariate NB intervals are blue "+"; the univariate PER are black "o" and the univariate MPR are green "Δ" (colour figure online)

5 An application to real data

In order to evaluate how the proposed procedure works on real data, we consider daily PM_{10} data (in $\mu g/m^3$) from 1 January 2015 to 19 October 2016 (658 days) at 24 sites in Piemonte (see http://www.arpa.piemonte.gov.it). The particulate matter PM_{10} emissions are regulated by the EU which has set two limit values for the protection of human health: the daily mean value may not exceed $50\,\mu g/m^3$ more than 35 times in a year and the annual mean value may not exceed $40\mu g/m^3$. In the considered dataset, isolated missing values as well as missing sequences are present due to the defaults in the monitoring stations. Their point reconstruction has been addressed in Parrella et al. (2019) by using model (1). The spatial matrix $\mathbf{W}$ has been set as the normalized geographical distances matrix of $\mathbf{y}_t$, i.e. its entries are $w_{ij}/\sum_{k=1}^{p} w_{kj}$, where $w_{ij} = 1/(1 + d_{ij})$, d_{ij} is the geographical distance between the i-th and j-th stations for $i \neq j$, and $w_{ii} = 0$ for $i = 1, \ldots, p$.

Here, for sake of illustration, we analyse the data from Novara-Verdi station which presents the 14.74% of missing data. In particular there is one isolated missing value on 12 January 2015 and then a sequence of missing values 42 observations long, starting from 11 March 2015. Figures 7 and 8 report the observed time series, along with the imputed values and the $k-$JPRs ($k = 1, 2, 3$) for the missing values obtained with the MPR and the NB methods respectively, in the time interval up to 30/04/2015. The confidence level is fixed at 90% and the red horizontal line is the EU limit $50\,\mu g/m^3$.

It is evident that, in both the cases, the reconstruction procedure for the missing values provides values that are quite often under the EU threshold. However, by looking at $k-$JPRs obtained with the MPR method in 7, their upper bounds is above 50 almost everywhere expecially for $k = 1$. Also for the isolated missing value, the univariate

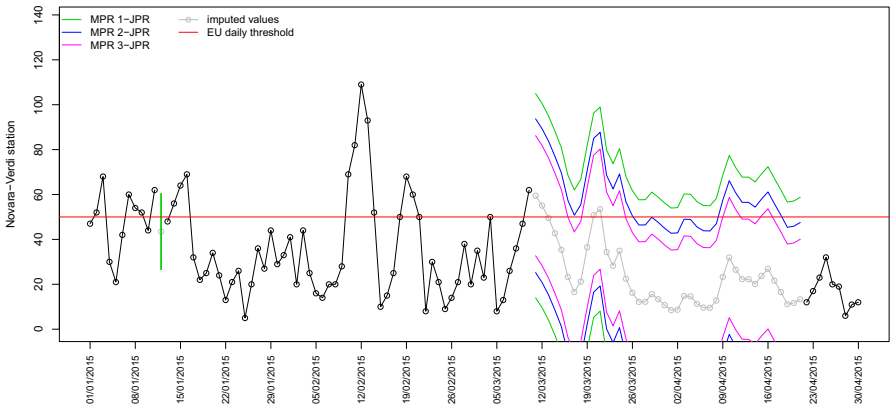

Fig. 7 PM_{10} data from Novara-Verdi station in the interval 01/01/2015–30/04/2015 (black line). The gray circles represent the imputed values for the missing values. The coloured lines are the $k-$JPRs, for $k = 1$ (green), 2 (blue) and 3 (magenta), calculated by using MPR method. Red horizontal line is the EU limit, i.e. $50\,\mu g/m^3$ (colour figure online)

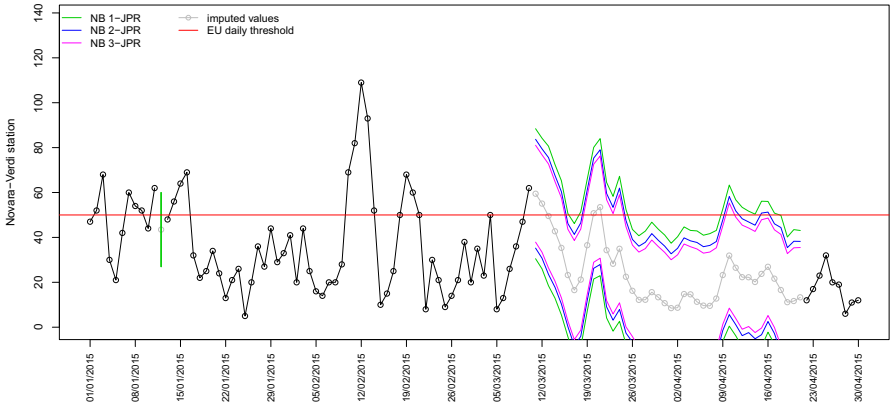

Fig. 8 PM_{10} data from Novara-Verdi station in the interval 01/01/2015–30/04/2015 (black line). The gray circles represent the imputed values for the missing values. The coloured lines are the $k-$JPRs, for $k = 1$ (green), 2 (blue) and 3 (magenta), calculated by using NB method. Red horizontal line is the EU limit, i.e. $50\,\mu g/m^3$ (colour figure online)

confidence interval includes the EU limit. Moreover, as expected, by increasing k, the lengths of the $k-$JPRs decrease.

When using the NB method (Fig. 8), by varying k, lengths are closer to each other, with respect to MPR case.

6 Concluding remarks

In this paper we deal with the construction of joint prediction intervals for missing data patterns in the context of spatio-temporal data. We have proposed a bootstrap

resampling scheme able to provide joint prediction regions that approximately contain missing paths of a time series of interest, with probability $1-\alpha$. The approach is based on maximum predictive roots proposed in Pan (2016) and Wolf et al. (2015). We have also considered JPRs that only contain all elements of missing paths up to a small number $k-1$ of them with probability $1-\alpha$, where the choice of k can be made with respect to the problem at hand.

A simulation experiment has been performed to validate the empirical performance of the proposed method based on the maximum of the predictive root statistic and to compare it with two simpler alternatives. In particular we have considered multivariate time series generated by a SDPD model (see Dou et al. 2016) with $p = 30$ and different sample sizes $T = 100$, 500 and 1000 and two distributions for the error terms, normal and student-t. In the experiment we have analysed the mean lengths of the obtained JPRs along with their empirical coverages both in the cases of isolated missing values and for missing sequences of different lengths. The simulation experiment has shown that generally the procedure based on the maximum predictive root delivers better and more stable performances for non-gaussian error terms and longer missing sequences, with similar mean lengths for the interpolation regions.

The application on real data shows that the reconstruction procedure for the missing values provides values that are in most of the cases under the EU threshold. However, the $k-$JPRs obtained with the MPR method produce intervals which quite often contain the EU limit of $50\,\mu g/m^3$, showing the importance of the additional information delivered by prediction intervals with respect to single predictions.

Funding Open access funding provided by Universitá degli Studi di Salerno within the CRUI-CARE Agreement.

References

Alonso AM, Sipols AE (2008) A time series bootstrap procedure for interpolation intervals. Comput Stat Data Anal 52:1792–1805

Alonso AM, Sipols AE, Quintas S (2013) A single-index model procedure for interpolation intervals in time series. Comput Stat 28:1463–1484

Atluri G, Karpatne A, Kumar V (2018) Spatio-temporal data mining: a survey of problems and methods. ACM Comput Surv 51(4):1–41

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

Cano S, Andreu J (2010) Using multiple imputation to simulate time series: a proposal to solve the distance effect. WSEAS Trans Comput 9(7):768–777

Choi E, Hall P (2000) Bootstrap confidence regions computed from autoregressions of arbitrary order. J R Stat Soc Ser B 62(3):461–477

Dou B, Parrella ML, Yao Q (2016) Generalized Yule–Walker estimation for spatio-temporal models with unknown diagonal coefficients. J Econom 194:369–382
Eischeid JK, Pasteris PA, Diaz HF, Lantico MSP, Lott NJ (2000) Creating a serially complete, national daily time series of temperature and precipitation for the western United States. J Appl Meteorol 39:1580–1591
Gao Z, Ma Y, Wang H, Yao Q (2019) Banded spatio-temporal autoregression. J Econom 208:211–230
Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18):2895–2907
Lee LF, Yu J (2010) Estimation of spatial autoregressive panel data models with fixed effects. J Econom 154:165–185
Liu S, Molenaar PC (2014) iVAR: a program for imputing missing data in multivariate time series using vector autoregressive models. Behav Res Methods 46(4):1138–1148
Lo Presti R, Barca E, Passarella G (2010) A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ Monit Assess 160:1–22
Pan L, Politis ND (2016) Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. J Stat Plan Inference 177:1–27
Parrella ML, Albano G, La Rocca M, Perna C (2019) Reconstructing missing data sequences in multivariate time series: an application to environmental data. Stat Method Appl 28(2):359–383
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
Qu X, Lee L, Yang C (2021) Estimation of a SAR model with endogenous spatial weights constructed by bilateral variables. J Econom 221(1):180–197
Schneider T (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J Clim 14:853–871
Smith KW, Aretxabaleta AL (2007) Expectation-maximization analysis of spatial time series. Nonlinear Proc Geophys 14(1):73–77
Teegavarapu RSV, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312:191–206
Wolf M, Wunderli D (2015) Bootstrap joint prediction regions. J Time Ser Anal 36(2):352–376
Yang H, Yang J, Han LD, Liu X, Pu L, Chin SM, Hwang HL (2018) Kriging based spatiotemporal approach for traffic volume data imputation. PloS one 13(4):1–11
Young KC (1992) A three-way model for interpolating for monthly precipitation values. Mon Weather Rev 120:2561–2569
Zhang X, Yu J (2018) Spatial weight matrix selection and model averaging for spatial autoregressive models. J Econom 203:1–18