Stochastics and Statistics

Time series interpolation via global optimization of moments fitting

Emilio Carrizosa b, Alba V. Olivares-Nadal a,b,* , Pepa Ramírez-Cobo c

a Department of Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Av. Reina Mercedes, s/n, 41012 Sevilla, Spain
b Department of Statistics and Operational Research, University of Sevilla, Spain
c Department of Statistics and Operational Research, University of Cádiz, Spain

A R T I C L E   I N F O

Article history:
Accepted 7 April 2013
Available online 16 April 2013

Keywords:
Missing values
Moments matching
Global optimization
Variable Neighborhood Search

A B S T R A C T

Most time series forecasting methods assume the series has no missing values. When missing values exist, interpolation methods, while filling in the blanks, may substantially modify the statistical pattern of the data, since critical features such as moments and autocorrelations are not necessarily preserved.

In this paper we propose to interpolate missing data in time series by solving a smooth nonconvex optimization problem which aims to preserve moments and autocorrelations. Since the problem may be multimodal, Variable Neighborhood Search is used to trade off quality of the interpolation (in terms of preservation of the statistical pattern) and computing times.

Our approach is compared with standard interpolation methods and illustrated on both simulated and real data.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The existence of missing values is a very common yet critical issue in data from a variety of fields such as engineering, physics, hydrology, finance, marketing or transportation; see [26,30,18,36,12,19,8,3].

Missing data may hide the pattern of the data, and they may considerably distort the results of any statistical analysis performed. In order to cope with data sets with missing values, two main strategies have been followed in the literature: one either modifies existing statistical techniques to accommodate the existence of missing data, e.g. [26], or one first estimates the missing values and then statistical methods are used for the completed data set.

The latter approach has been widely used, in particular, in time series analysis, addressed in this paper. A number of interpolation methods have been suggested from long ago. Some assume a statistical parametric model, such as a MA(∞), [2,32], an ARMA, [6,15,23,27], an ARIMA, [17], or they impose a specific form for the interpolator, which ranges from piecewise polynomial-based approaches (linear, nearest-neighbor, logistic, B-splines), e.g. [4,20,38,37], to sinc interpolation or wavelets (in connection to filtering and signal processing), see [35] and the references given there, and Chapter 19 in [34] for a detailed description of such methods.

Because of their simplicity, weak underlying assumptions and good empirical performance, simple polynomial-based interpolation methods such as linear, nearest neighbor, cubic or splines are implemented in general-purpose packages as Matlab or R.

In the linear interpolation method, the missing data in an interval are imputed by the straight line that passes through the interval endpoints. The nearest neighbor algorithm assigns the closest known neighbor to a missing point, leading in this way to a piecewise constant interpolant. In a similar spirit, the cubic interpolation method fills a missing data interval by the third degree polynomial that passes through the four nearest neighbors. Finally, the spline interpolation method completes the series via a cubic spline under the requirements of smoothness and existence of derivatives. More specifically, if $y_i$ and $y_{i+1}$ denote the known edges of the missing interval then, this will be completed by the function

$$ S_i(x) = a_i(x - y_i)^3 + b_i(x - y_i)^2 + c_i(x - y_i) + d_i $$

where the constants $a_i, b_i, c_i$ and $d_i$ are calculated to make the resulting function smooth enough. In other words, they are chosen according to

$$ S_i(y_i) = S_i(y_{i+1}), \quad S_i'(y_i) = S_i'(y_{i+1}), \quad S_i''(y_i) = S_i''(y_{i+1}). $$

Though not too often mentioned, such interpolation methods may yield rather unsatisfactory results, since they may even provide interpolated values out of the possible range: the values may be known, for instance, to be non-negative (rainfall, exchange rates, etc.) or to range in a given interval, such as [0,1] if the time series

* Research partially supported by Research Grants and Projects MTM2009-14039, MTM2012-36163 (Ministerio de Ciencia e Innovación, Spain) and FQM329 (Junta de Andalucía, Spain), both with EU ERDF funds. The third author is supported by Consolider “Ingenio Mathematica” through her post-doc contract.

* Corresponding author at: Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Av. Reina Mercedes, s/n, 41012 Sevilla, Spain. Tel.: +34 637 872 829.

E-mail address: aolivares@us.es (A.V. Olivares-Nadal).
represent the evolution of a ratio, (unemployment rate, disease prevalence, ...), etc. As an illustrating example, which motivated our study, consider Fig. 1, where the series of daily precipitation amounts between 1976 and 1980 at the station of Nijar (Andalucía, Spain) has been analyzed. The top panel illustrates the real series, with several missing values, due to disfunctions in the measurement equipment. The central and bottom panels show the interpolated series according to the spline and cubic interpolation, respectively. It can be seen that under both methods negative values were obtained. Moreover, by construction, there is no guarantee that the moments and autocorrelations of the so-obtained time series remain close to their sample estimates. In other words, the statistical pattern of the time series may be strongly distorted when interpolation is done.

The purpose of this paper is to explore how missing values in time series can be interpolated so that the imputed data are forced to belong to a specific interval, and thus the drawback of some interpolation methods illustrated by Fig. 1 is avoided, and the moments and autocorrelation function of the data are preserved as well.

The paper is organized as follows. Section 2 formulates the problem of interpolating missing values preserving range, moments and correlations as a nonconvex optimization problem with box constraints. A well-known global-optimization metaheuristic, namely, Variable Neighborhood Search (VNS), is customized for this problem, as discussed in Section 3. Section 4 is devoted to illustrate the performance of our approach by comparing it with the benchmark interpolation methods previously described. Finally, Section 5 presents conclusions and prospects regarding this work.

2. Problem formulation

As commented in Section 1, the purpose of this work is to develop a method for interpolating the missing values in an incomplete time series so that the range of values, as well as moments and autocorrelation coefficients, are preserved. In this section we describe how this can be naturally modeled as an optimization problem.

Consider a sequence \( y = \{y_t\}_{t=1}^N \) of real values. The index set \( \{1, 2, \ldots, N\} \) is partitioned into two sets: the index set \( B \) of times \( t \) for which the value \( y_t \) is missing, and the index set \( S = \{1, 2, \ldots, N\} \setminus B \) of times \( t \) for which \( y_t \) is known. The aim is to recover the sequence \( y \) when only the values \( \{y_t\}_{t \in S} \) are given.

Any sequence \( x = \{x_t\}_{t=1}^N \) satisfying

\[
x_t = y_t \quad \forall t \in S,
\]

interpolates the partially observed series \( y \). However, not any such interpolating sequence will yield a reasonable estimate of \( y \), as we illustrated with the example in Section 1. In order to avoid out-of-range problems, we will also impose \( x \) to satisfy

\[
a_t \leq x_t \leq b_t \quad \forall t \in B.
\]

The constants \( a_t, b_t \) satisfying \( -\infty \leq a_t \leq b_t \leq +\infty \) are assumed to be given by the user. A possible choice is obtained if we consider, for each \( t \in B \), \( a_t = \min_{s \in S} y_s \) and \( b_t = \max_{s \in S} y_s \), and thus the range of \( x \) coincides with the range of \( y \). However, if the time series is suspected to present outliers, such a choice of \( a_t \) and \( b_t \) may lead to extreme values. In these cases, \( a_t \) and \( b_t \) can be derived from Tukey's fences, though, as discussed in Section 4, this choice may also be controversial.

![Fig. 1. Nijar rainfall series from 1976 to 1980 (top panel) and imputed data according to the spline (central panel) and cubic (bottom panel) interpolation methods.](image-url)
Together with forcing \( x \) to take values in a specified range, by imposing (2), we also wish to fix its statistical patterns by making moments and autocorrelation coefficients of \( x \) match given target values:

\[
\frac{1}{N} \sum_{t=1}^{N} x_t^k = m_k \quad \forall k = 1, 2, \ldots, k_0.
\]

(3)

and

\[
\rho_j(x) = \rho_j \quad \forall j = 1, 2, \ldots, j_0.
\]

(4)

Here \( m_k \) and \( \rho_j \) are given target values for the \( k \)th moment and lag-\( j \) autocorrelation coefficient, where

\[
\rho_j(x) = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x})(x_{t+j} - \bar{x})}{\sum_{t=1}^{N-1} (x_t - \bar{x})^2}.
\]

(5)

is the estimation of the lag-\( j \) autocorrelation coefficient as defined in [9], and \( x \) is the sample mean of \( x \).

Target values \( m_k \) and \( \rho_j \) allow one to preserve the statistical pattern of the time series. They can be either estimated from the observed series \( y_t \), \( t = 1, \ldots, N \), or from a related but complete time series (even the same time series in a different time window). See Section 4.4 for a discussion on the effect of the selection of target values on the performance of the interpolation approach described in this paper.

While (2)–(4) allow us to govern the statistical pattern and the range of the series, a smoothing criterion is also considered, imposing the missing values not to significantly differ from the adjacent values. In other words, if we define \( f \) as

\[
f(x) = \sum_{t=1}^{N-1} (x_{t+1} - x_t)^2,
\]

(6)

we also wish to have \( f(x) \) as small as possible.

The previous discussion leads to us model the problem of properly interpolating the incomplete time series \( y \) as the optimization problem of finding \( x \) minimizing \( f \), as defined in (6), and satisfying the constraints (1)–(4). However, the feasible region defined by (1)–(4) is rather complex, and it may be hard for a numerical algorithm to find a feasible solution. Moreover, since analyst confidence on the target values \( m_k \) and \( \rho_j \) may be different, we find more convenient to consider (1) and (2) as hard constraints, and (3) and (4) as soft constraints. This way we obtain an optimization problem of the form

\[
\text{min } F(x) \quad \text{s.t. } \begin{cases} a_i \leq x_i \leq b_i & \forall i \in B \\ x_i = y_i & \forall i \in S. \end{cases}
\]

(P)

The objective function \( F \) in (P) is given by

\[
F(x) = f(x) + \sum_{k=1}^{k_0} \lambda_k \left( \frac{1}{N} \sum_{i=1}^{N} x_i^k - 1 \right)^2 + \sum_{j=1}^{j_0} \mu_j \left( \rho_j(x) - \rho_j \right)^2.
\]

where \( p_0 \) is a reference starting point for \( x \), so that the quotient \( f(x) \) is dimensionless, as the remaining terms in \( F \), and the parameters \( \lambda_k, k = 1, \ldots, k_0 \) and \( \mu_j, j = 1, \ldots, j_0 \) are positive scalars which trade off the deviations in the soft constraints.

Problem (P) is a nonlinear problem with very simple constraints (just box constraints), but a nonconvex objective function. A closed-form solution to (P) seems hard to obtain due to the high nonlinearity of the objective. Hence, in order to cope with (P), numerical procedures are suggested. This will be discussed in Section 3.

3. Solving the problem

Problem (P) is a smooth optimization problem. Moreover, in contrast with what happens if (1)–(4) are all considered as hard constraints, obtaining starting feasible solutions for (P) is straightforward. Hence, obtaining a local minimum for (P) is a rather simple and cheap task, achievable by standard local-search numerical routines. However, due to the nonconvexity of the objective, there is no guarantee that the output of such local-search routines is a global optimum of (P).

In order to scope from local optima, we propose to embed the local searches into a metaheuristic strategy, namely, the Variable Neighborhood Search (VNS), [29,28,10], which can successfully exploit the fact that the feasible region is rather simple, and thus neighborhoods are easily defined. It may be observed that other metaheuristics, e.g. [24,22,14,16], could have been used instead.

The scheme of the VNS algorithm is summarized in Fig. 2.

The VNS is customized to Problem (P) by defining the neighborhood structure, the random distributions for shaking, the starting solution and the stopping criterion.

Since the feasible region of (P) is box-constrained, a set of nested boxes are chosen as neighborhoods, and sampling is performed by following a uniform distribution on the boxes. As stopping criterion, an upper bound on the number of iterations allowed is given.

A more sophisticated strategy is followed for the selection of the starting point, since it is commonly accepted, and confirmed as well in our numerical experiments with (P), that choosing a good starting point is critical to guarantee an appropriate convergence speed for the VNS. Although, as already mentioned, it is unlikely to obtain a closed-form solution of problem (P), it is possible however to solve analytically a simpler relative of (P), in which constraints (2) and (4) are ignored, while, together with the interpolation constraints (1), the constraint (3) for \( k_0 = 1 \) are put as hard constraint, (or, equivalently, we set \( i_1 = +\infty \) in (P)). In other words, we consider the auxiliary problem \((P^*)\) given by:

\[
\min \sum_{i=1}^{N} (x_{i+1} - x_i)^2
\]

\[
\text{s.t. } \begin{cases} x_i = y_i & \forall i \in S. \\ \sum_{i=1}^{N} x_i = m_1. \end{cases}
\]

\((P^*)\)

Problem \((P^*)\) has a quadratic convex objective and one linear constraint. An optimal solution can be analytically derived, as discussed in the Appendix. Such optimal solution will be taken as starting solution \( x_0 \) for the VNS algorithm for solving (P).

4. Numerical illustrations

In this section we illustrate the performance of the proposed methodology and compare it with the results provided by standard interpolation methods. All results have been obtained using \( \mathbb{R} \) and, in the spirit of a reproducible research, the codes utilized in this paper are available as a stand-alone \( \mathbb{R} \) toolbox from the authors upon request.

4.1. Data and experiments description

Ten real time series from different contexts and presenting different statistical features have been selected to illustrate the performance of the proposed approach. The series considered in this Section do not present missing values (see Section 4.3, for a real incomplete series) and therefore, their sample moments and autocorrelation coefficients are known. As will be described in the experiments below, a chosen proportion of observations will be randomly removed in order to test different interpolation methods. The series, shown at the top panels of Figs. 3–7, are described next:

- exchange. This series represents the evolution of the exchange rates between the Hong Kong Dollar (HKD) and the US Dollar (USD). The number of moments and autocorrelation coefficients to be matched were \( k_0 = 3 \) and \( j_0 = 30 \), respectively. Data can be found at http://gtwavelet.bme.gatech.edu/datapro.html
• **Initialization:** Define
  
  - A neighborhood structure: a family of neighborhoods \( \{V_i(x), i = 1, \ldots, i_{\text{max}}\} \), for all feasible \( x \)
  - Random distributions on the neighborhoods \( V_i(x) \), to be used in the **Shaking step**
  - An initial solution \( x_0 \)
  - The number \( Q_{\max} \) of random points generated on each neighborhood
  - A stopping criterion

• **Repeat** the following sequence until the stopping condition is met:
  
  - Set \( i \leftarrow 1 \), \( q \leftarrow 1 \)
  - **Repeat** the following steps until \( i > i_{\text{max}} \):
    
    * **Shaking:** Generate a point \( x \) randomly from the \( i \)-th neighborhood of \( x_0 \) \( (x \in V_i(x_0)) \)
    
    * **Local search:** Apply some local search method with \( x \) as initial solution to obtain a local optimum given by \( \tilde{x} \)
    
    * **Neighborhood change:** If this local optimum is better than the incumbent, move there \( (x_0 \leftarrow \tilde{x}) \), and continue the search with \( V_1(\tilde{x}) \) \( (q \leftarrow 1) \); otherwise, set \( q \leftarrow q + 1 \). If \( q > Q_{\max} \), then set \( q \leftarrow 1 \), \( i \leftarrow i + 1 \)

Fig. 2. Pseudo-code of VNS.

- **Norwegian.** A total of 9181 consecutive fire claims values from 1972 to 1992 in a Norwegian insurance portfolio are considered here. Data are heavy-tailed (see [5,33] for a complete description, and analysis of the data concerning heavy-tailed distributions) and \( k_0 = 3 \) and \( j_0 = 10 \) were chosen. The series is available at [http://lstat.kuleuven.be/Wiley/](http://lstat.kuleuven.be/Wiley/).

- **Internet.** This time series, widely used in the literature (see [25,33], among others), represents 50,000 real interarrival times in a total of one million packet arrivals, recorded by Bellcore Morristown Research and Engineering facility. The series is highly variable and presents a non-negligible autocorrelation coefficients. Again, \( k_0 = 3 \) and \( j_0 = 10 \) have been set. Data are found at [http://www.sigcomm.org/ITA/](http://www.sigcomm.org/ITA/).

- **Coke.** This series records a total of 4128 daily Coca-Cola stock market price on dollars. Fig. 4 shows that the series is clearly non-stationary and long-range dependent. In order to explore how the moments matching approach performs if only the fitting of the autocorrelation coefficients is considered, we set \( k_0 = 0 \) and \( j_0 = 1 \). Data can be found at [http://gtwave.elet.bme.gatech.edu/datapro.html](http://gtwave.elet.bme.gatech.edu/datapro.html).

- **Cordoba.** This time series, provided by the AEMET (Spanish National Meteorology Agency, [http://www.aemet.es](http://www.aemet.es)) represents the total daily precipitation at the station of Córdoba’s airport (Spain) from 2004 to 2005. In this case \( k_0 = 3 \) and \( j_0 = 1 \) were fixed.

The following series are provided by the **Time Series Data Library** found at [http://datamarket.com/data/list/?q=provider:tsdl](http://datamarket.com/data/list/?q=provider:tsdl). All of them represent monthly observations and thus the number of autocorrelation coefficients to be matched was set to \( j_0 = 12 \). In all cases, a value of \( k_0 = 3 \) was chosen.

- **Gas.** This series represents 106 monthly average gas usage (measured in cubic feet times 100) from 1971 to 1979 at Iowa state. The series, analyzed by [1], presents a strong cyclical pattern.

- **Riverflow.** This time series records a total of 588 monthly riverflow from the Boise River (near Twin Springs, Idaho) measured in cubic meters per second. Data correspond to periods from October 1912 to September 1960. A more detailed description of the series can be found at [21].

- **Unemploy.** A number of 736 monthly civilian unemployment rates at the US from January 1948 to December 2011 are represented in this series. Data were provided by The Bureau of Labor Statistics (BLS).

- **Gnp.** This series, analyzed by [31], represents 176 quarterly US real GNP (in dollars) from the first quarter of 1947 to the first quarter of 1991.

- **Milk.** This time series relates to cow milk production. Specifically, it represents a total of 156 monthly pounds of milk per cow, from January 1962 to December 1975. It was listed and analyzed in [13]. This series is clearly non-stationary, see Fig. 7.

Next, we describe the three conducted experiments. As commented previously the series do not have missing values and therefore they were artificially **incompleted** as follows.

- **Experiment 1:** 30% of the observations, randomly selected, was deleted.

- **Experiment 2:** 15% of the observations, randomly selected, was deleted.

- **Experiment 3:** An interval of consecutive observations, representing the 30% of the total, was deleted.

The values 30% and 15% were arbitrarily selected as examples of high and moderate percentages of missing data.
4.2. Performance and comparison with benchmark approaches

This section presents the results obtained after interpolating the time series described in Section 4.1. by the interpolation method defined in this work, namely, the moments matching method (MMM, from now on), and compares the obtained performance with that provided by Linear, Nearest Neighbor, Cubic and Spline methods, which are easily found in common statistical packages.
such as R or Matlab. The chosen option ‘cubic’ makes a cubic interpolation based on the four nearest neighbors; the Spline approach is encompassed in piecewise Cubic Interpolation. For more information about Spline interpolation the reader is referred to [7].

We should point out here that, in order to implement our approach, the R-cran function optim was used with default options as local-search routine. The function implements the L-BFGS-B algorithm, which allows for box constrained problems. In practice,
penalties have been considered constant and for convenience set to $\lambda_k = \lambda = 5000$, for $k = 1, \ldots, k_1$, and $\mu_j = \mu = 4000$, for $j = 1, \ldots, j_0$. The number of VNS neighborhoods and total of random points generated at each neighbor were 7 and 2, respectively. Finally, the target values are set as the sample moments of the complete series.

Tables 1–3 show the absolute percent error from target values in Experiments 1–3 obtained under the MMM (highlighted in gray

---

**Fig. 5.** Cordoba (first column) and Gas (second column) time series. Top: Real series and proportion of data removed in Experiment 3. Central: Interpolation by MMM. Bottom: Best interpolation (if different from that obtained by the MMM), according to Table 2.
color), Linear, Nearest Neighbor, Cubic and Spline interpolation methods, for all time series described in the previous section. Recall that \( m_1, m_2, m_3 \) denote the moments of order 1, 2, and 3 of the series, and \( \rho_j \) the jth-lag autocorrelation coefficient. Imputed values are required to belong to the interval defined by the minimum and maximum values in the series. Fitted values that lie out of such a range are starred (\(*\)). Although in some series \( j_0 \) was set up to 30, for abbreviation, we only present the error when fitting \( \rho_1 \).

Several conclusions can be obtained from Tables 1–3. First, note how in the three experiments considered our approach provides
interpolated series whose moments are either very close to, or exactly match the target values. Additionally, the imputed values always remain within the required interval. On the contrary, the rest of methods present a poorer performance (the results in Experiment 1 being poorer than Experiment 2, as expected, since less information is considered). In Experiment 1, MMM provides better fits than the rest of methods in 28 out of 37 fitted coefficients, and comparable results to the best of the alternative interpolators in 3
cases. In Experiment 2, these values turn into 29 and 5 (out of 37), respectively. Finally, in Experiment 3, MMM performs better than or equal to the best of the other methods in 35 cases; see the panels of the second row of Figs. 3–7 which depict the completion of the missing central gap under the MMM approach (data in between the bands are assumed to be missing). Note that, in appearance, no big differences exist between the real and interpolated series in all cases. Also, it is of interest to note from Table 3 how the performance of classic interpolation methods gets worse in Experiment 3, due to their strong dependence on neighbor data.

In order to examine more in depth the performance of the MMM against that of other existing interpolation methods, it may be natural to apply the commonly used prediction error between the actual values of the time series and the estimated values. To this end, the absolute and quadratic differences, normalized by the smallest error, have been computed. They are shown in Table 4, in which the best obtained results are highlighted in bold style. Note that the best interpolation method according to these criteria presents an error equal to 1, and for the rest of methods this value is exceeded. Under the absolute error criterion, MMM outperforms the rest of methods in 12 out of 30 cases; presents a comparable performance in three cases and performs poorer than the best of the remaining benchmark approaches in the half of cases. If the quadratic difference criterion is used instead, MMM outperforms the other interpolators in 15 out of 30 cases, and performs similarly to the best in two cases. Panels of the third row of Figs. 3–7 show how MMM and Linear approaches perform when fitting the missing values. Fig. 8 illustrates how the rest of interpolators, namely, Nearest Neighbor (top right panel), Cubic (bottom left panel) and Spline (bottom right panel) behave for the time series in all cases. Again, we emphasize the poor performance obtained under these methods.

### 4.3. VNS versus local search

The optimization problem described in Section 2 has been solved using a global optimization technique, the VNS, which.

---

**Table 1**

<table>
<thead>
<tr>
<th>Series</th>
<th>MMM</th>
<th>Linear</th>
<th>Nearest</th>
<th>Cubic</th>
<th>Spline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.10</td>
<td>0.10</td>
<td>0.01</td>
<td>3.29</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.23</td>
<td>1.88</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>11.19</td>
<td>11.22</td>
<td>3.47</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>8.35</td>
<td>8.22</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>9.75</td>
<td>1.90</td>
<td>2.19</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>1.02</td>
<td>0.87</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>4.30</td>
<td>2.10</td>
<td>88.33</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>8.07</td>
<td>3.49</td>
<td>71.09</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>24.15</td>
<td>15.94</td>
<td>28.36</td>
<td></td>
</tr>
<tr>
<td>Coke</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>Cordoba</td>
<td>0.00</td>
<td>22.32</td>
<td>22.32</td>
<td>17.62</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>35.32</td>
<td>27.19</td>
<td>27.76</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>1.21</td>
<td>0.41</td>
<td>25.16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>3.36</td>
<td>3.94</td>
<td>2603</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>0.04</td>
<td>8.27</td>
<td>70.53</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.00</td>
<td>1.28</td>
<td>38.73</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>7.88</td>
<td>0.64</td>
<td>1837</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>15.40</td>
<td>0.26</td>
<td>68.38</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>9.24</td>
<td>3.21</td>
<td>40.15</td>
<td></td>
</tr>
<tr>
<td>Unemploy</td>
<td>0.07</td>
<td>0.07</td>
<td>0.02</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.08</td>
<td>0.13</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.27</td>
<td>0.32</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.36</td>
<td>0.38</td>
<td>0.02</td>
<td>6.91</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.40</td>
<td>6.72</td>
<td>8.90</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.14</td>
<td>12.09</td>
<td>8.09</td>
<td>2264</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
<td>10.62</td>
<td>5.39</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.08</td>
<td>26.08</td>
<td>17.39</td>
<td>40.66</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.04</td>
<td>0.19</td>
<td>0.25</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.31</td>
<td>0.49</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.37</td>
<td>0.73</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.32</td>
<td>2.35</td>
<td>2.12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.07</td>
<td>0.07</td>
<td>0.01</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.06</td>
<td>0.20</td>
<td>0.34</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**

<table>
<thead>
<tr>
<th>Series</th>
<th>MMM</th>
<th>Linear</th>
<th>Nearest</th>
<th>Cubic</th>
<th>Spline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.10</td>
<td>0.10</td>
<td>0.01</td>
<td>3.29</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.23</td>
<td>1.88</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>11.19</td>
<td>11.22</td>
<td>3.47</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>8.35</td>
<td>8.22</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>9.75</td>
<td>1.90</td>
<td>2.19</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>1.02</td>
<td>0.87</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>4.30</td>
<td>2.10</td>
<td>88.33</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>8.07</td>
<td>3.49</td>
<td>71.09</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>24.15</td>
<td>15.94</td>
<td>28.36</td>
<td></td>
</tr>
<tr>
<td>Coke</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>Cordoba</td>
<td>0.00</td>
<td>22.32</td>
<td>22.32</td>
<td>17.62</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>35.32</td>
<td>27.19</td>
<td>27.76</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>1.21</td>
<td>0.41</td>
<td>25.16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>3.36</td>
<td>3.94</td>
<td>2603</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>0.04</td>
<td>8.27</td>
<td>70.53</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.00</td>
<td>1.28</td>
<td>38.73</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>7.88</td>
<td>0.64</td>
<td>1837</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>15.40</td>
<td>0.26</td>
<td>68.38</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>9.24</td>
<td>3.21</td>
<td>40.15</td>
<td></td>
</tr>
<tr>
<td>Unemploy</td>
<td>0.07</td>
<td>0.07</td>
<td>0.02</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.08</td>
<td>0.13</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.27</td>
<td>0.32</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.36</td>
<td>0.38</td>
<td>0.02</td>
<td>6.91</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.40</td>
<td>6.72</td>
<td>8.90</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.14</td>
<td>12.09</td>
<td>8.09</td>
<td>2264</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
<td>10.62</td>
<td>5.39</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.08</td>
<td>26.08</td>
<td>17.39</td>
<td>40.66</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.04</td>
<td>0.19</td>
<td>0.25</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.31</td>
<td>0.49</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.37</td>
<td>0.73</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.32</td>
<td>2.35</td>
<td>2.12</td>
<td></td>
</tr>
</tbody>
</table>
4.4. The choice of target values

As commented in Section 1, the MMM approach for completing a time series with missing values needs for the specification of target values, to which the moments and autocorrelation coefficients are matched. In the previous Section where the series were artificially made incomplete by erasing some records, these target values were set as the empirical moments of the complete series (being the sample autocorrelation coefficients defined as in (5)). However, if the series really presents missing data, as will happen in practice, these sample moments may be seriously distorted. This section aims to investigate how the choice of target values influences the results of the proposed interpolation method.

We first consider an example where 1000 synthetic data were simulated from an AR(1) process with parameters \( m_1 = 1 \) and \( \rho_1 = 0.6 \). Experiments 1, 2 and 3, as described in Section 4.1, were carried out with \( K_0 = J_0 = 1 \), and the target values fixed as deviations from \( m_1 = 1 \) and \( \rho_1 = 0.6 \). Specifically, 0%, 5%, 10% and 15% deviations from the theoretical values were considered. Obtaining sample estimates for the autocorrelation coefficients is not a trivial task since the available information consists in different sequences spaced in time. Two different estimates for \( \rho_j \), particularized for \( j = 1 \), were also considered. First, a weighted autocorrelation coefficient estimator is proposed as follows: the series is split into sub-series of consecutive values with length higher than \( j \); for each such sub-series, its sample \( j \)th lag autocorrelation coefficient \( \hat{\rho}_j \) is computed according to (5), and then the estimator is given by

\[
\hat{\rho}_j = \sum_{\ell} \omega_{n, \ell} \hat{\rho}_j, 
\]

where \( \omega_{n, \ell} \) represents the weight of the \( \ell \)th sub-series. However, a restriction worth to be mentioned while considering the \( j \)th lag autocorrelation is that it cannot be calculated if the length of the interval is less than \( j \). Moreover, (5) does not behave properly when the length of the interval is not much larger than \( j \). This could lead to unreliable estimations, thus the \( j \)th autocorrelation of the longest interval of known data can be considered as an estimator instead of the weighted one. Other estimates for \( \rho_j \) suggested in the literature can be found in [11].

Table 6 shows the predictive errors, when interpolating using MMM, computed as the sum of the (absolute/squared) differences between the interpolated and real values, under the assortment of target values commented previously. In the last two columns the target values \( m_1 \) are chosen as the sample mean of the uncompleted series, and the target values \( \rho_1 \) are computed as the weighted estimator (7) for \( j = 1 \), and the sample estimator (5) of the longest complete interval, respectively.

From Table 6 it can be deduced that, in this example, the estimator of the autocorrelation function based on the longest known interval produces better results than the weighted one. It could be also said that, as expected, the larger the intervals of known data, the better performance is obtained by both estimators. Finally, and as expected too, the larger the deviations from the real values \( m_1 \) and \( \rho_1 \), the poorer the results obtained under both absolute and squared errors criterion.

We conclude this section by interpolating via the MMM a time series which actually had missing values, and the pattern of such missing values is unknown. The series was described in Section 1 and represents the daily precipitation amounts in Nijar, Spain, from 1976 to 1980. It was shown in Section 1 how classic methods as Spline and Cubic failed in interpolating the series properly due to the fact that the imputed values severely violated the range of admissible values. Values of \( K_0 = J_0 = 3 \) were set; since the sample moments and autocorrelation coefficients are unknown, the target values \( m_1, m_2 \) and \( m_3 \) were chosen as the sample moments of the available data, and the target values \( \rho_1, \rho_2 \) and \( \rho_3 \) were computed.
as the weighted estimators, and as the sample estimators (5) from the longest interval.

Nijar has sub-desertic Mediterranean climate and therefore rainfall days are very rare, although when they occur precipitation amounts may be extreme. This leads to interquartile ranges close to zero. In order to implement our approach we have considered lower and upper bounds as the minimum and maximum of the known data, as in the previous examples. From top panel of Fig. 9, it can be seen that the series does not present isolated missing data but located missing intervals, which motivates the use of the weighted estimator (7). Central and bottom panels show the completed series via MMM. In the central panel, the estimator (7) has been used while in the bottom panel, only the estimator corresponding to the longest interval has been considered. It can be observed from both figures how imputed values representing precipitation amounts are non-negative and do not exceed the

<table>
<thead>
<tr>
<th>Series</th>
<th>Abs diff.</th>
<th>MMM Linear</th>
<th>Nearest</th>
<th>Cubic</th>
<th>Spline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>1 abs</td>
<td>1.04</td>
<td>1.00</td>
<td>1.27</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>1 quad</td>
<td>1.05</td>
<td>1.00</td>
<td>1.75</td>
<td>31.36</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>1.06</td>
<td>2.62</td>
<td>11.16</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.00</td>
<td>2.42</td>
<td>2.42</td>
<td>1.14 x 10^6</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>3.08</td>
<td>4.03</td>
<td>1.76 x 10^6</td>
</tr>
<tr>
<td>Norwegian</td>
<td>1 abs</td>
<td>6.81</td>
<td>1.08</td>
<td>1.00</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>2.24</td>
<td>1.11</td>
<td>1.00</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>4.33</td>
<td>1.07</td>
<td>1.00</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>2.22</td>
<td>1.00</td>
<td>1.00</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.95</td>
<td>1.00</td>
<td>1.01</td>
<td>3.96 x 10^6</td>
</tr>
<tr>
<td>Internet</td>
<td>1 abs</td>
<td>1.36</td>
<td>1.00</td>
<td>1.17</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.45</td>
<td>1.00</td>
<td>1.21</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.79</td>
<td>1.00</td>
<td>1.46</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.25</td>
<td>1.00</td>
<td>1.02</td>
<td>1.99 x 10^6</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.29</td>
<td>1.00</td>
<td>1.00</td>
<td>3.88 x 10^6</td>
</tr>
<tr>
<td>Coke</td>
<td>1 abs</td>
<td>1.00</td>
<td>1.00</td>
<td>1.27</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.00</td>
<td>1.00</td>
<td>1.51</td>
<td>519.50</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>1.30</td>
<td>1.60</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.00</td>
<td>1.66</td>
<td>3.22</td>
<td>1.95 x 10^6</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>3.22</td>
<td>11.83</td>
<td>6.51 x 10^8</td>
</tr>
<tr>
<td>Cordoba</td>
<td>1 abs</td>
<td>1.48</td>
<td>1.00</td>
<td>1.02</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.29</td>
<td>1.00</td>
<td>1.03</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>1.08</td>
<td>1.74</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.97</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Gas</td>
<td>1 abs</td>
<td>1.00</td>
<td>2.04</td>
<td>4.53</td>
<td>18.30</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.00</td>
<td>4.02</td>
<td>19.29</td>
<td>1795.20</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>7.42</td>
<td>23.21</td>
<td>19.96</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.00</td>
<td>39.43</td>
<td>376.53</td>
<td>301.51</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>5.32</td>
<td>4.20</td>
<td>1.15 x 10^9</td>
</tr>
<tr>
<td>Riverflow</td>
<td>1 abs</td>
<td>1.00</td>
<td>1.22</td>
<td>1.51</td>
<td>5.69</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.00</td>
<td>1.43</td>
<td>2.27</td>
<td>322.56</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>1.22</td>
<td>1.99</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.00</td>
<td>1.51</td>
<td>4.40</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>1.07</td>
<td>1.08</td>
<td>1.31 x 10^6</td>
</tr>
<tr>
<td>Unemploy</td>
<td>1 abs</td>
<td>1.06</td>
<td>1.00</td>
<td>1.36</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.06</td>
<td>1.00</td>
<td>1.77</td>
<td>27.69</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.12</td>
<td>1.00</td>
<td>1.48</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.09</td>
<td>1.00</td>
<td>1.69</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>1.27</td>
<td>1.26</td>
<td>1.27 x 10^6</td>
</tr>
<tr>
<td>Gnp</td>
<td>1 abs</td>
<td>1.11</td>
<td>1.00</td>
<td>1.00</td>
<td>5.62</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.35</td>
<td>1.00</td>
<td>1.01</td>
<td>205.98</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.12</td>
<td>1.00</td>
<td>1.27</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.35</td>
<td>1.00</td>
<td>1.01</td>
<td>2.93 x 10^9</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.78</td>
<td>1.00</td>
<td>1.01</td>
<td>1.27 x 10^9</td>
</tr>
<tr>
<td>Milk</td>
<td>1 abs</td>
<td>1.00</td>
<td>1.92</td>
<td>4.38</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>2 abs</td>
<td>1.00</td>
<td>3.42</td>
<td>14.60</td>
<td>16.13</td>
</tr>
<tr>
<td></td>
<td>2 quad</td>
<td>1.00</td>
<td>2.25</td>
<td>2.95</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td>3 abs</td>
<td>1.00</td>
<td>6.94</td>
<td>10.08</td>
<td>34.70</td>
</tr>
<tr>
<td></td>
<td>3 quad</td>
<td>1.00</td>
<td>5.56</td>
<td>5.61</td>
<td>3.03 x 10^4</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>26.60</td>
<td>27.36</td>
<td>1.23 x 10^4</td>
<td>521.75</td>
</tr>
</tbody>
</table>
maximum of the known series. Moreover, it is interesting to note that not all imputed data are close to zero; there exists a large imputed value, of the order of the larger values in the observed series, in the case of the weighted estimator interpolation. Note that, although the longest interval is completed with drain days in central panel, precipitations are allowed on other time instants.

5. Concluding remarks and extensions

This work has considered the problem of interpolating missing values in a time series so that moments and autocorrelation coefficients are fitted to target values. This is done via a smooth non-convex optimization problem, solved via a VNS approach in a continuous space.
By several numerical examples, the suitability of the new approach has been highlighted, in comparison with classic interpolation methods, which clearly present a poorer performance. Regarding the optimization problem to be solved, the choice of a proper starting point and the use of a global optimization routine are shown to be relevant.

For simplicity, just univariate time series have been considered, though the model naturally extends to the multivariate case, in which also the correlation between the different time series is governed by setting target values. Experimental analysis of such extension deserve further attention.

**Appendix A. Closed-form solution of \((P^*)\)**

In this section, how to obtain the optimal solution of \((P^*)\) is shown in detail. The problem to be solved is:

\[
\begin{align*}
\min & \quad \sum_{i=1}^{N-1} (x_{i+1} - x_i)^2 \\
\text{s.t.} & \quad x_i = y_i, \quad \forall i \in S, \\
& \quad \sum_{i=1}^{N} m_i = m_1.
\end{align*}
\]

The Lagrangian function is given by

\[
L(x, \lambda) = \sum_{i=1}^{N-1} (x_{i+1} - x_i)^2 + \lambda \left( \sum_{i=1}^{N} m_i - m_1 \right).
\]

As shown next, the solution to \((P^*)\) depends on the number of intervals containing missing values and their lengths. In order to solve \((P^*)\), three cases are considered: the length of the interval of missing values is greater than or equal to 3 (case 1) or it is just 2 or just 1 (case 2).

First, we first address the problem for a unique interval, and then we will extend the obtained results to the case of several intervals of missing values.

**Case 1.a) A unique interval of missing values.**

Assume a unique interval of consecutive missing values, given by \([x_{p}, \ldots, x_{p+r}]\), of length \(R = r + 1 \geq 3\). The partial derivatives of the Lagrangian function are given by

\[
\begin{align*}
\frac{\partial L}{\partial x_p} &= 4x_p - 2y_{p-1} - 2x_{p+1} + \frac{\lambda}{N}, \\
\frac{\partial L}{\partial x_{p+j}} &= 4x_{p+j} - 2x_{p+j-1} - 2x_{p+j+1} + \frac{\lambda}{N}, \quad \text{if } 1 \leq j \leq r - 1, \\
\frac{\partial L}{\partial x_{p+r}} &= 4 + x_{p+r} - 2x_{p+r-1} - 2y_{p+r+1} + \frac{\lambda}{N}, \quad \text{if } j = r.
\end{align*}
\]

Making the Lagrangian function equal to zero, we obtain

\[
x_{p+j} = (j + 1)x_p - jy_{p-1} + \frac{j(j + 1)}{4N} \lambda, \quad \text{for } 1 \leq j \leq r - 1.
\]

On the other hand, \(x_{p+r}\) can be calculated in two different ways. First, from the partial derivative of (8) with respect to \(x_{p+r}\), we conclude that

\[
x_{p+r} = (r + 1)x_p - r(y_{p-1} + \frac{r(r + 1)}{4N} \lambda).
\]

Also, from the partial derivative of (8) with respect to \(x_{p+r}\),

\[
x_{p+r} = \frac{r - 1}{2}x_p - \frac{r - 1}{2}y_{p-1} + y_{p+1} + \frac{r^2 - r - 2}{8N} \lambda.
\]

Making (10) and (11) equal, we obtain

\[
x_p = \frac{r + 1}{r + 2}y_{p-1} - \frac{1}{r + 2}y_{p-1} + \frac{r^2 - 3r}{4N(r + 2)} \lambda.
\]

From the active constraint

\[
\sum_{i=1}^{N} m_i = m_1,
\]

it follows that

\[
x_p + \sum_{j=1}^{r-1} x_{p+j} + x_{p+r} = c,
\]

where \(c = Nm_1 - \sum_{i \in S} y_i\) from now on. Substituting (9), (11) and (12) into (13), yields

\[
\lambda = \frac{2c - (r + 1)y_{p-1} - (r + 1)y_{p+r-1}}{\gamma},
\]

where

\[
\gamma = -\frac{r^3 - 2r^2 - 3r - 2 + 2\tau}{4N} \quad \text{and} \quad \tau = \sum_{j=1}^{r-1} j(j + 1).
\]

**Case 1.b) \(L_1 \geq 2\) intervals of missing values.**

Assume in this case a number \(L_1 \geq 2\) of intervals of missing values, noted \(I_1, \ldots, I_{L_1}\). An analysis similar to that in the previous case shows that, in the interval \(I_0\) of length \(R_0 = r_0 + 1\), for \(1 \leq n \leq L_1\),

\[
\begin{align*}
x_{p_n} &= \frac{r_0 + 1}{r_0 + 2}y_{p_n} - \frac{1}{r_0 + 2}y_{p_n - 1} + \frac{r_0^2 - 3r_0 - 2}{4N(r_0 + 2)} \lambda, \\
x_{p_n + r_0} &= (r_0 + 1)x_{p_n} - r_0y_{p_n} + \frac{r_0(r + 1)}{4N} \lambda, \\
x_{p_n + r_0 + r_1} &= \frac{r_1}{2}x_{p_n} - \frac{r_1}{2}y_{p_n} + \frac{1}{2}y_{p_n + r_0} + \frac{r_1^2 - 2r_1}{8N} \lambda, \\
x_{p_n + j} &= (j + 1)x_{p_n} - jy_{p_n} + \frac{j(j + 1)}{4N} \lambda, \quad 1 \leq j \leq r_n,
\end{align*}
\]

where

\[
\lambda = \frac{2c - \sum_{n=1}^{L_1} (r_n + 1)y_{p_n} - \sum_{n=1}^{L_1} (r_n + 1)y_{p_n + r_n - 1}}{\gamma_n},
\]

where \(\gamma_n = -\frac{r_n^2 - 2r_n - 2r_n + 2r_n}{4N} \) and \(r_n = \sum_{j=1}^{r_n - 1} j(j + 1)\).
Now we address \((P^*)\) for case 2, which corresponds to intervals of missing values of length 1 or 2.

(Case 2.a) A unique interval of missing values of length 2
Assume a single interval \(T\) of missing values \(x_p\) and \(x_{p+1}\), for \(2 \leq p \leq N - 2\).
In the same manner as in the previous calculations we can see that
\[
\begin{align*}
x_p &= \frac{2}{3}y_{p-1} + \frac{1}{3}y_{p+2} - \frac{1}{2N} \lambda, \\
x_{p+1} &= \frac{2}{3}y_{p+2} + \frac{1}{3}y_{p-1} - \frac{1}{2N} \lambda,
\end{align*}
\]
where in this case
\[
\lambda = \frac{-c + y_{p-1} + y_{p+2}}{2N}.
\]

(Case 2.b) \(I_2\) intervals of missing values of length 2.
Assume \(I_1, \ldots, I_{I_2}\) intervals of two missing values in the time series \(x\). The reasoning above applies to this case to obtain, for \(1 \leq n \leq I_2\)
\[
\begin{align*}
x_p &= \frac{2}{3}y_{p_n-1} + \frac{1}{3}y_{p_n+2} - \frac{1}{2N} \lambda, \\
x_{p+1} &= \frac{2}{3}y_{p_n+2} + \frac{1}{3}y_{p_n-1} - \frac{1}{2N} \lambda,
\end{align*}
\]
where
\[
\lambda = \frac{-c - \sum_{n=1}^{I_2} (y_{p_n-1} + y_{p_n+2})}{2N}.
\]

(Case 2.c) A unique missing data.
When there exists one single missing data \(x_p\) in the time series \(x\), it is easily seen that
\[
x_p = \frac{1}{2}y_{p-1} + \frac{1}{2}y_{p+1} - \frac{1}{4N} \lambda,
\]
for \(2 \leq p \leq N - 1\), where
\[
\lambda = \frac{-2c - y_{p-1} - y_{p+1}}{2N}.
\]

(Case 2.d) \(I_3\) isolated missing values.
Finally, assume a set of isolated missing data \((x_i)_{i \in I}\), where the length of \(I\) is \(I_3\). Again, it is straightforward to obtain
\[
x_p = \frac{1}{2}y_{p-1} + \frac{1}{2}y_{p+1} - \frac{1}{4N} \lambda \quad \text{con} \quad I_3,\]
for any \(2 \leq j \leq N - 1\) and where
\[
\lambda = \frac{-2c - \sum_{n=1}^{I_3} (y_{p_n-1} - y_{p_n+1})}{2N},
\]

We end this Appendix by putting all the previous results together, giving a closed-form expression for the solution to \((P^*)\). Consider a time series which has \(I_1\) intervals of missing data of length less than or equal to 3, \(I_2\) intervals of length 2, and \(I_3\) isolated missing values. Let \(|I|\) represent the number of missing values. The expressions for the missing data are the same derived previously, where the only difference is the value of \(\lambda\), which has to be the same for all expressions. Taking into account the active constraint to preserve the mean
\[
\sum_{n=1}^{I_1} \sum_{j=1}^{I_2} (x_{p_n} + x_{p_n+j} + x_{p_{n+j}}) + \sum_{n=1}^{I_2} (x_{p_n} + x_{p_{n+1}}) + \sum_{n=1}^{I_3} x_{p_n} = c,
\]
we obtain
\[
\lambda = \frac{2c - \sum_{n=1}^{I_1} ((r_n + 1)y_{p_n-1} + (r_n + 1)y_{p_n+r_n+1}) - \sum_{n=1}^{I_2} (2y_{p_n-1} + 2y_{p_n+2}) - \sum_{n=1}^{I_3} (y_{p_n-1} + y_{p_n+1})}{\sum_{n=1}^{I_1} r_n - \frac{2}{N} - \frac{I_3}{N}},
\]
where \( c = Nm_n - \sum s_i y_i \) and

\[
\gamma_n = -\frac{r_n^2 - 2r_n^2 - 3r_n - 2 + 2r_n}{4N}, \quad \tau_n = \sum_{j=1}^{n-1} j(j + 1).
\]

References


