European Journal of Operational Research 230 (2013) 97–112

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

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] Departamento de 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

ARTICLE INFO

Article history:
Received 14 September 2012
Accepted 7 April 2013
Available online 16 April 2013

Keywords:
Missing values
Moments matching
Global optimization
Variable Neighborhood Search

ABSTRACT

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.

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 $MA(\infty)$, [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

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

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).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.04.008

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 Níjar (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 $\mathbf{y} = \{y_t\}_{t=1}^N$ 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 $S = \{1,2,\ldots,N\}\backslash B$ of times t for which y_t is known. The aim is to recover the sequence $\mathbf{y}$ when only the values $\{y_t\}_{t\in S}$ are given.

Any sequence $\mathbf{x} = \{x_t\}_{t=1}^N$ satisfying

$$x_t = y_t \qquad \forall t \in S, \tag{1}$$

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

$$a_t \leqslant x_t \leqslant b_t \qquad \forall t \in B. \tag{2}$$

The constants a_t, b_t satisfying $-\infty \leqslant a_t \leqslant b_t \leqslant +\infty$ are assumed to be given by the user. A possible choice is obtained if we consider, for each $t \in B$, $a_t = \min_{n\in S} y_n$ and $b_t = \max_{n\in S} y_n$, and thus the range of $\mathbf{x}$ coincides with the range of $\mathbf{y}$. 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.

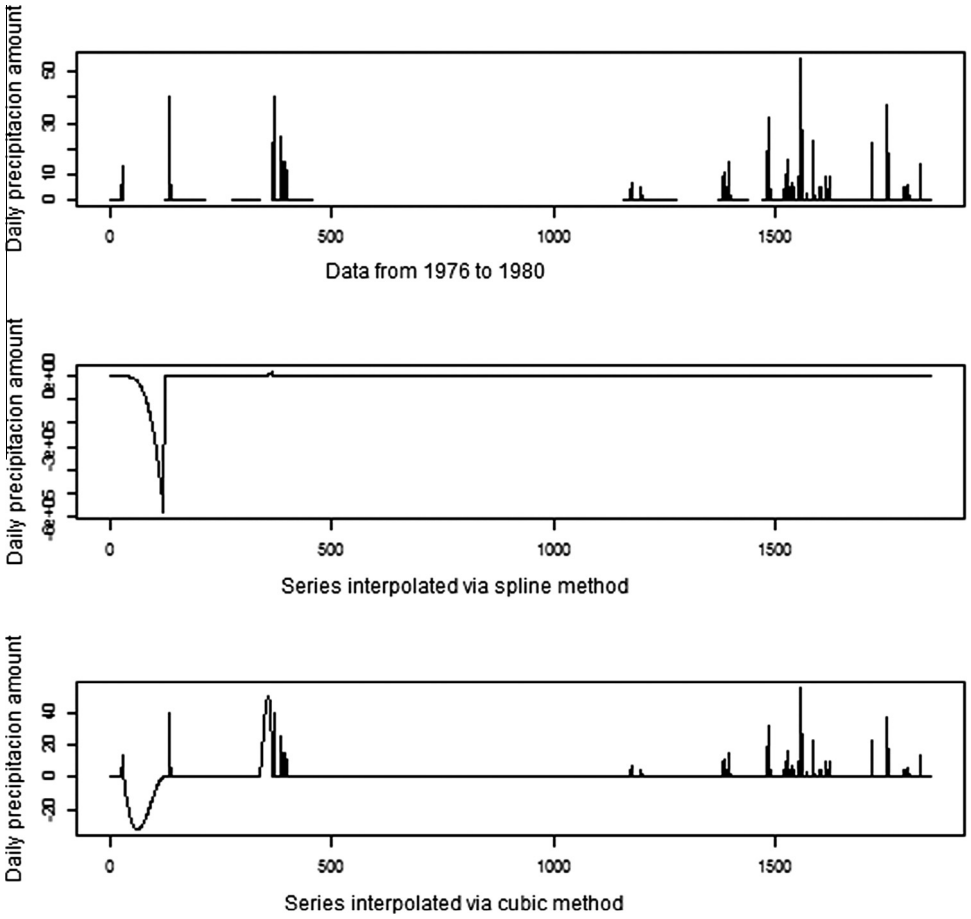

Fig. 1. Nijar rainfall series from 1976 to 1980 (top panel) and imputed data according to the spline (central panel) and cubic (bottom panel) interpolation methods.

Together with forcing $\mathbf{x}$ 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 $\mathbf{x}$ match given target values:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^k = m_k \quad \forall k = 1, 2, \ldots, k_0, \tag{3}$$

and

$$\rho_j(\mathbf{x}) = \rho_j \quad \forall j = 1, 2, \ldots, j_0. \tag{4}$$

Here m_k and ρ_j are given target values for the kth moment and lag-j autocorrelation coefficient, where

$$\rho_j(\mathbf{x}) = \frac{\sum_{t=1}^{N-j}(x_t - \bar{x})(x_{t+j} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}, \tag{5}$$

is the estimation of the lag-j autocorrelation coefficient as defined in [9], and $\bar{x}$ is the sample mean of $\mathbf{x}$.

Target values m_k and ρ_j allow one to preserve the statistical pattern of the time series. They can be either estimated from the observed series $\{y_t\}_{t\in S}$, 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

$$f(\mathbf{x}) = \sum_{i=1}^{N-1}(x_{i+1} - x_i)^2, \tag{6}$$

we also wish to have $f(\mathbf{x})$ as small as possible.

The previous discussion leads us to model the problem of properly interpolating the incomplete time series $\mathbf{y}$ as the optimization problem of finding $\mathbf{x}$ 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 ρ_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

$$\begin{array}{ll} \min & F(\mathbf{x}) \\ \text{s.t} & \begin{cases} a_t \leqslant x_t \leqslant b_t & \forall i \in B \\ x_t = y_t & \forall i \in S. \end{cases} \end{array} \tag{P}$$

The objective function F in (P) is given by

$$F(\mathbf{x}) = \frac{f(\mathbf{x})}{f(\mathbf{p_0})} + \sum_{k=1}^{k_0}\lambda_k\left(\frac{\frac{1}{N}\sum_{i=1}^{N}x_i^k}{m_k} - 1\right)^2 + \sum_{j=0}^{j_0}\mu_j\left(\frac{\rho_j(\mathbf{x})}{\rho_j} - 1\right)^2,$$

where $\mathbf{p}_0$ is a reference starting point for $\mathbf{x}$, so that the quotient $\frac{f(\mathbf{x})}{f(\mathbf{p_0})}$ is dimensionless, as the remaining terms in F, and the parameters λ_k, $k = 1, \ldots, k_0$ and μ_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 scape 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 $\lambda_1 = +\infty$ in (P)). In other words, we consider the auxiliary problem (P^*) given by:

$$\begin{array}{ll} \min & \sum_{i=1}^{N-1}(x_{i+1} - x_i)^2 \\ \text{s.t.} & \begin{cases} x_t = y_t & \forall t \in S, \\ \frac{\sum_{i=1}^{N} x_i}{N} = m_1 \end{cases} \end{array} \tag{P*}$$

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 $\mathbf{x}_0$ 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 R and, in the spirit of a reproducible research, the codes utilized in this paper are available as a stand-alone R toolbox from the authors upon request.

4.1. Data and experiments description

Ten real times 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 $\{V_i(\mathbf{x}), i = 1, ..., i_{max}\}$, for all feasible $\mathbf{x}$
  - Random distributions on the neighborhoods $V_i(\mathbf{x})$, to be used in the *Shaking* step
  - An initial solution $\mathbf{x_0}$
  - 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 $i > i_{max}$:
    - Shaking: Generate a point $\mathbf{x}$ randomly from the i-th neighborhood of $\mathbf{x_0}$ $(\mathbf{x} \in V_i(\mathbf{x_0}))$
    - Local search: Apply some local search method with $\mathbf{x}$ as initial solution to obtain a local optimum given by $\tilde{\mathbf{x}}$
    - Neighborhood change: If this local optimum is better than the incumbent, move there $(\mathbf{x_0} \leftarrow \tilde{\mathbf{x}})$, and continue the search with $V_1(\tilde{\mathbf{x}})$ $(q \leftarrow 1)$; 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.

- **N**orwegian. 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/
- **I**nternet. 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 the *InternetTraffic Archive*: http://www.sigcomm.org/ITA/
- **C**oke. 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://gtwavelet.bme.gatech.edu/datapro.html
- **C**ordoba. This time series, provided by the AEMET (Spanish National Meteorology Agency, 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. 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.

- **G**as. 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.
- **R**iverflow. 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].
- **U**nemploy. 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).
- **G**np. 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.
- **M**ilk. 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.

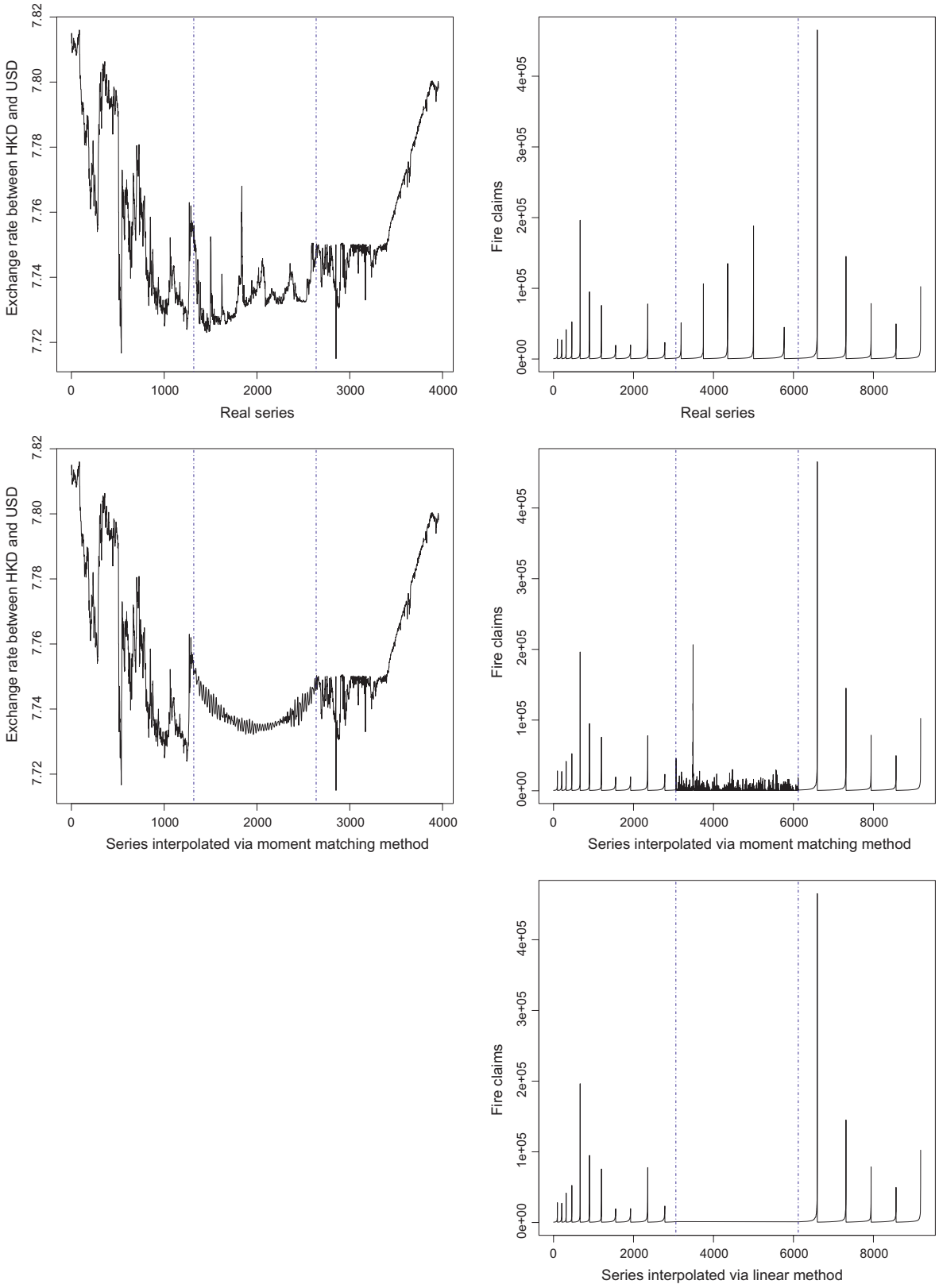

Fig. 3. `Exchange` (first column) and `Norwegian` (second column) time series. *Top*: Real series and proportion of data removed in Experiment 3. *Central*: Interpolation by MMM. *Bottom*: Best interpolation (if different from that obtained by the MMM), according to Table 2.

4.2. Performance and comparison with benchmark approaches

This section presents the results obtained after interpolating the times 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

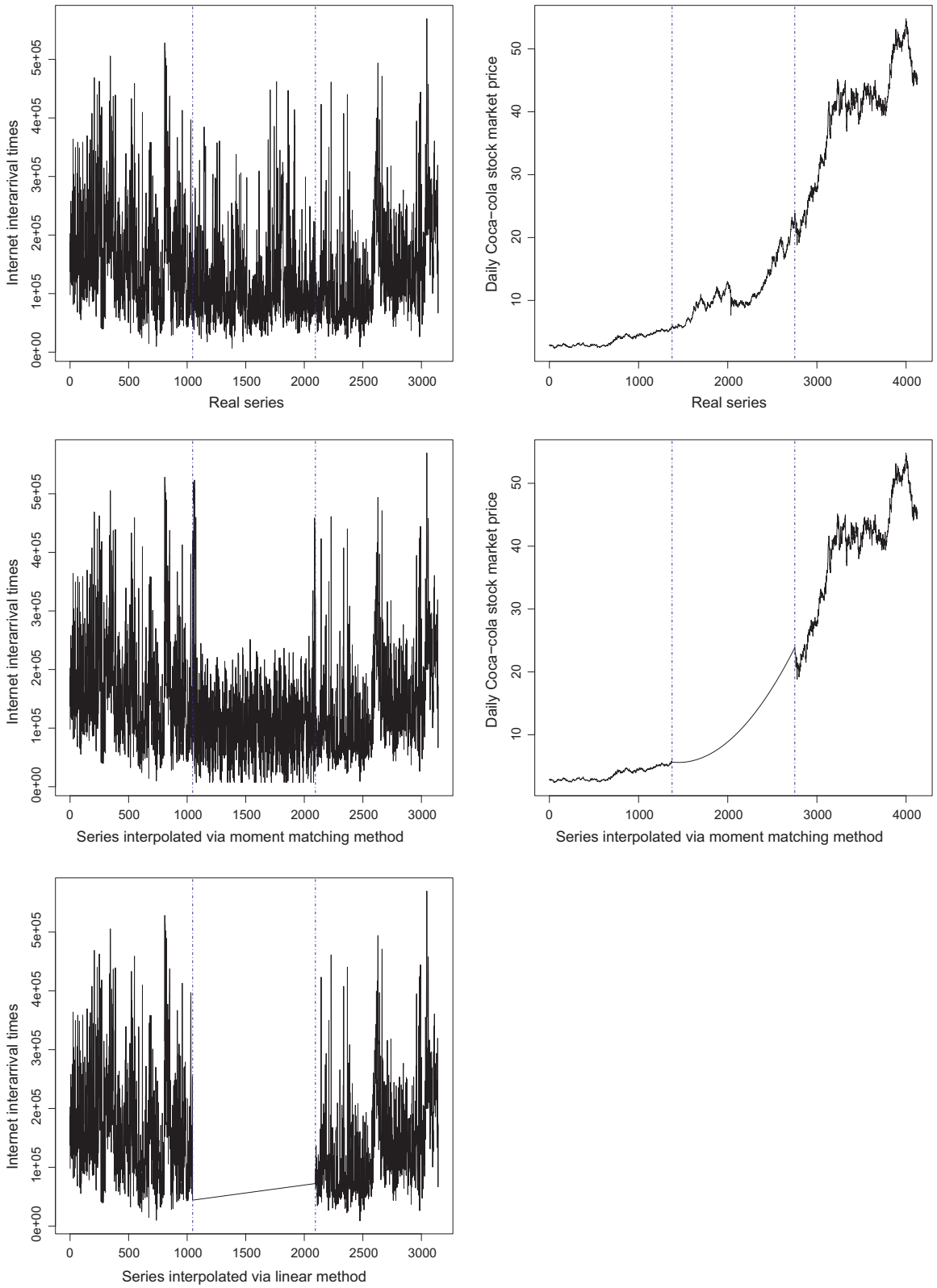

Fig. 4. `Internet` (first column) and `Coke` (second column) time series. *Top*: Real series and proportion of data removed in Experiment 3. *Central*: Interpolation by MMM. *Bottom*: Best interpolation (if different from that obtained by the MMM), according to Table 2.

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,

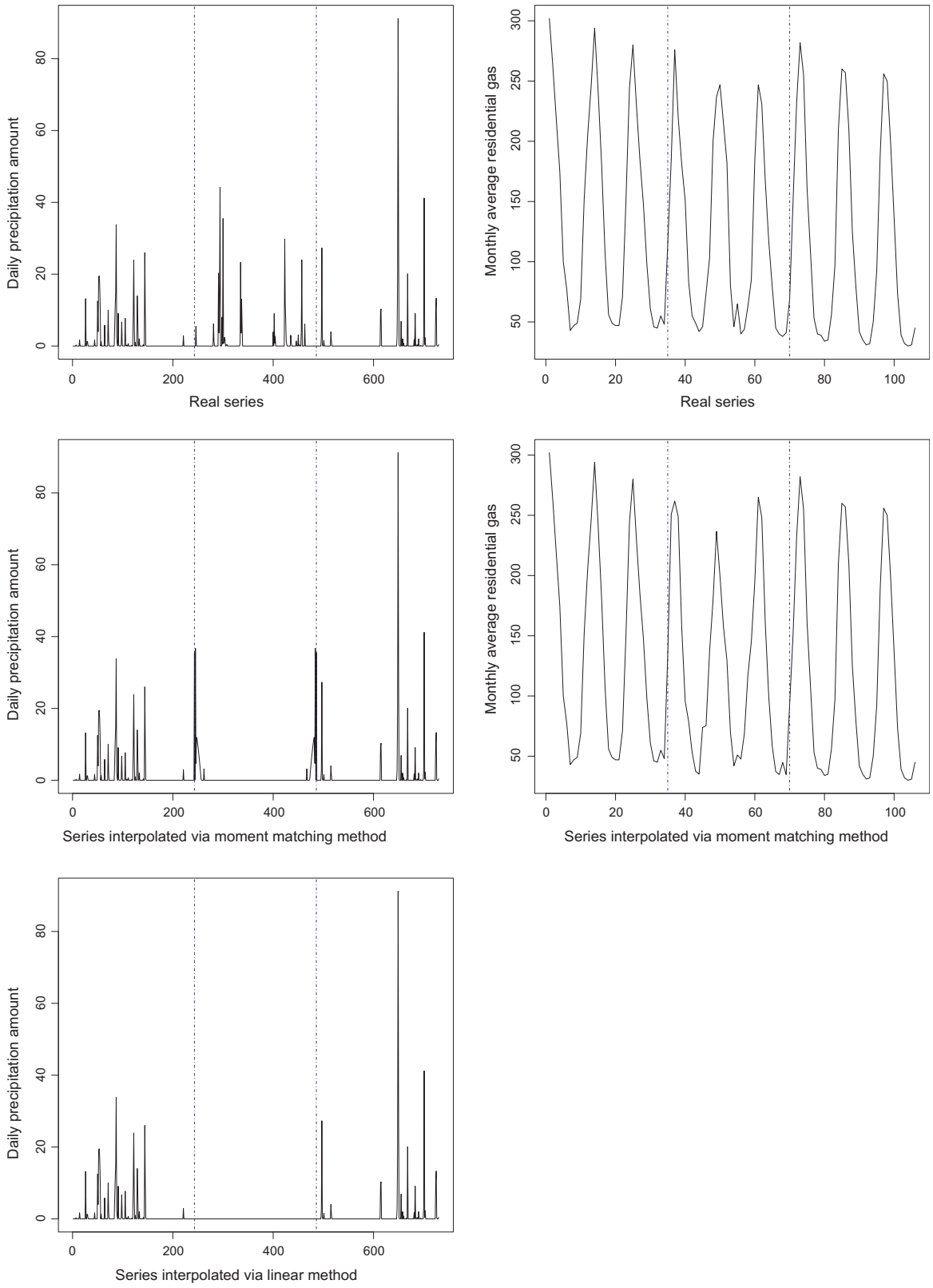

Fig. 5. `Cordoba` (first column) and `Gas` (second column) time series. *Top*: Real series and proportion of data removed in Experiment 3. *Central*: Interpolation by MMM. *Bottom*: Best interpolation (if different from that obtained by the MMM), according to Table 2.

penalties have been considered constant and for convenience set to $\lambda_k = \lambda = 5000$, for $k = 1, \ldots, k_0$, 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

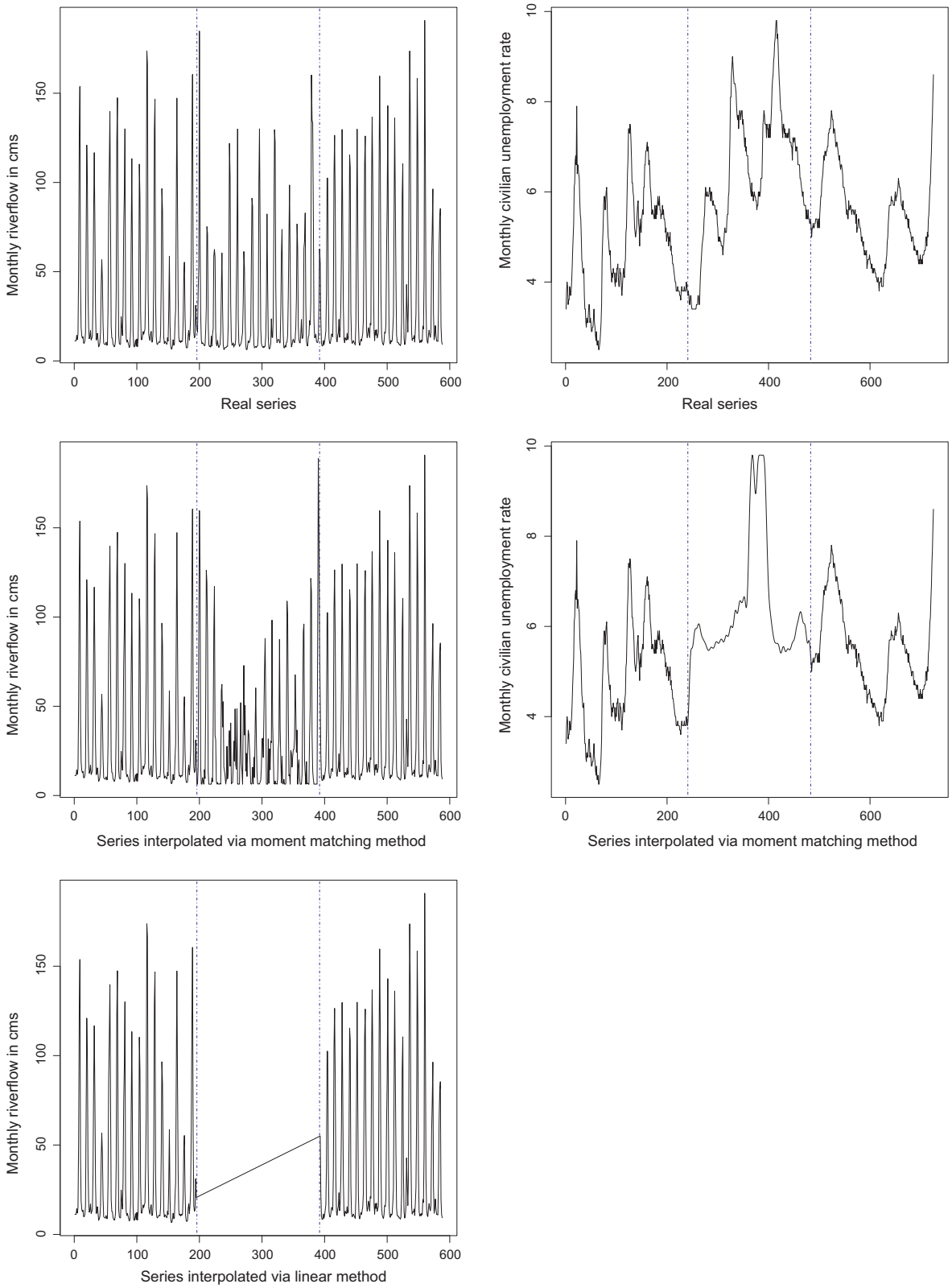

Fig. 6. `Riverflow` (first column) and `Unemploy` (second column) time series. *Top*: Real series and proportion of data removed in Experiment 3. *Central*: Interpolation by MMM. *Bottom*: Best interpolation (if different from that obtained by the MMM), according to Table 2.

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 ρ_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 ρ_1.

Several conclusions can be obtained from Tables 1–3. First, note how in the three experiments considered our approach provides

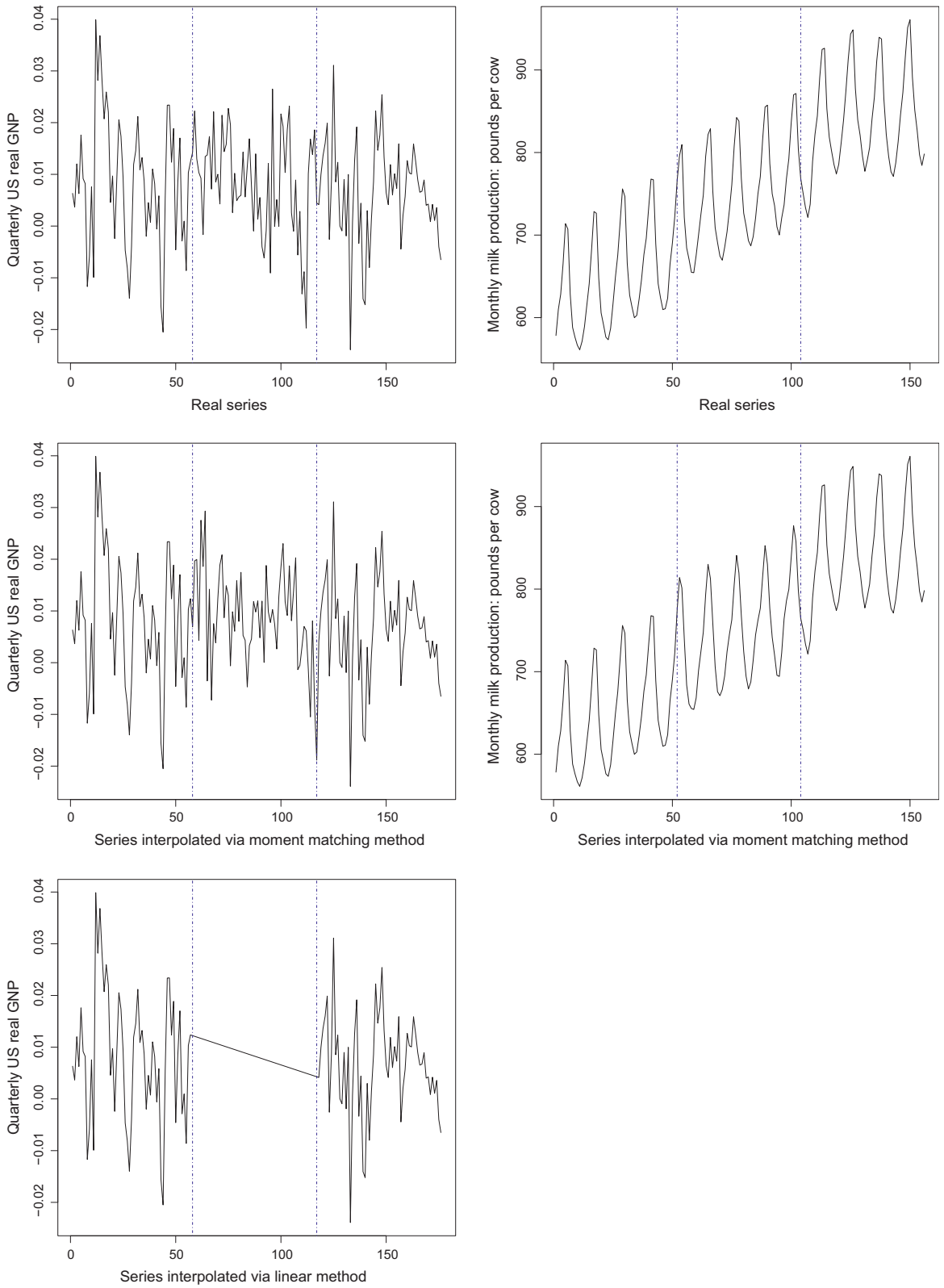

Fig. 7. `Gnp` (first column) and `Milk` (second column) time series. *Top*: Real series and proportion of data removed in Experiment 3. *Central*: Interpolation by MMM. *Bottom*: Best interpolation (if different from that obtained by the MMM), according to Table 2.

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

Table 1
Absolute percent error from target values in Experiment 1.

Series		MMM	Linear	Nearest	Cubic	Spline
Exchange	m_1	0.00	0.00	0.00	0.00*	0.00*
	m_2	0.00	0.00	0.00	0.00	0.00
	m_3	0.00	0.00	0.00	0.00	0.00
	ρ_1	0.10	0.10	0.01	3.29	0.09
Norwegian	m_1	0.00	1.23	1.88	2.61*	1.04*
	m_2	0.00	11.19	11.22	3.47	3.95
	m_3	0.00	8.35	8.22	1.09	5.15
	ρ_1	0.00	9.75	1.90	2.19	14.09
Internet	m_1	0.09	1.02	0.87	0.32*	1.45*
	m_2	0.06	4.30	2.10	88.33	1.58
	m_3	0.00	8.07	3.49	71.09	1.95
	ρ_1	0.03	24.15	15.94	28.36	24.09
Coke	ρ_1	0.01	0.01	0.00	2.00*	0.00
Cordoba	m_1	0.00	22.32	22.32	17.62*	20.71*
	m_2	0.03	57.59	52.40	51.13	48.77
	m_3	0.00	81.13	77.54	77.99	76.72
	ρ_1	0.02	35.32	27.19	27.76	52.24
Gas	m_1	0.02	1.21	0.41	25.16*	0.53*
	m_2	0.01	0.80	2.17	260.11	0.43
	m_3	0.02	3.36	3.94	2603	0.49
	ρ_1	0.08	0.04	8.27	70.53	0.72
Riverflow	m_1	0.01	0.00	1.28	38.73*	2.34*
	m_2	0.02	7.88	0.64	1837	1.51
	m_3	0.00	15.40	0.26	68,389	2.76
	ρ_1	0.03	9.24	3.21	40.15	11.63
Unemploy	m_1	0.07	0.01	0.02	0.41*	0.02
	m_2	0.04	0.08	0.13	0.45	0.11
	m_3	0.06	0.27	0.32	0.35	0.25
	ρ_1	0.36	0.38	0.02	6.91	0.33
Gnp	m_1	0.40	6.72	8.90	100*	1.12*
	m_2	0.14	12.09	8.09	2264	8.20
	m_3	0.42	10.62	5.39	100	26.95
	ρ_1	0.08	26.08	17.39	40.66	37.86
Milk	m_1	0.04	0.19	0.25	0.51	0.03
	m_2	0.02	0.31	0.49	0.95	0.04
	m_3	0.03	0.37	0.73	1.36	0.03
	ρ_1	0.03	0.32	2.35	2.12	0.23

Table 2
Absolute percent error from target values in Experiment 2.

Series		MMM	Linear	Nearest	Cubic	Spline
Exchange	m_1	0.00	0.00	0.00	0.00	0.00
	m_2	0.00	0.00	0.00	0.00	0.00
	m_3	0.00	0.00	0.00	0.00	0.00
	ρ_1	0.03	0.03	0.02	0.36	0.03
Norwegian	m_1	0.01	0.98	1.94	0.17	0.93*
	m_2	0.03	7.33	9.37	0.98	2.93
	m_3	0.01	5.69	6.38	0.33	3.20
	ρ_1	0.03	6.40	1.98	7.70	8.63
Internet	m_1	0.07	0.32	0.31	0.07*	0.49*
	m_2	0.05	1.77	0.73	1.13	0.52
	m_3	0.00	3.47	1.30	5.65	0.57
	ρ_1	0.02	12.57	7.82	4.34	12.70
Coke	ρ_1	0.00	0.00	0.00	0.04*	0.00
Cordoba	m_1	0.00	12.80	15.52	14.22*	16.78*
	m_2	0.02	41.36	40.42	31.54	29.84
	m_3	0.00	68.50	67.11	62.80	65.46
	ρ_1	0.01	11.37	5.83	4.88	28.26
Gas	m_1	0.02	0.81	1.46	1.23	0.03
	m_2	0.00	0.39	3.39	1.15	0.14
	m_3	0.02	0.27	4.81	0.66	0.26
	ρ_1	0.03	0.23	4.57	3.63	0.35
Riverflow	m_1	0.02	0.08	1.21	1.50*	0.98*
	m_2	0.03	2.89	2.86	9.64	0.99
	m_3	0.01	5.16	3.75	36.57	0.31
	ρ_1	0.03	2.11	2.53	19.63	3.09
Unemploy	m_1	0.01	0.03	0.08	0.02	0.06*
	m_2	0.00	0.07	0.13	0.04	0.11
	m_3	0.03	0.12	0.15	0.06	0.17
	ρ_1	0.21	0.24	0.04	0.00	0.21
Gnp	m_1	0.26	6.76	6.50	4.04	10.77
	m_2	0.54	0.08	0.54	0.37	8.69
	m_3	0.14	2.51	3.23	3.48	15.63
	ρ_1	0.18	18.24	11.89	12.50	22.90
Milk	m_1	0.02	0.20	0.34	0.45	0.08
	m_2	0.01	0.47	0.73	0.99	0.20
	m_3	0.04	0.81	1.16	1.62	0.36
	ρ_1	0.06	0.50	1.24	3.13	0.32

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. Marked disparities between classic interpolation methods are also observed, note for example the results from Table 3 between Cubic and the remaining interpolators. Finally, it is worth pointing out that, for Spline and Cubic interpolators, it is a rule more than an exception to impute data outside the natural range of values and in some cases quite extreme values are obtained.

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 the interpolated series in Experiment 3 by the best method (if different from MMM), according to the quadratic error criterion. Note that in these cases the linear interpolation method is the approach that always outperforms the MMM. However, note too from Figs. 3–7 that, in spite of presenting a larger prediction error, the MMM approach provides a more reasonable fit in terms of the statistical pattern of the series than the linear method; see for example the cases of Internet, Riverflow and Gnp.

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 Internet. 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 3
Absolute percent error from target values in Experiment 3.

Series		MMM	Linear	Nearest	Cubic	Spline
Exchange	m_1	0.02	0.07	0.07	$3.26 \times 10^{6*}$	0.61*
	m_2	0.04	0.14	0.14	7.27×10^{11}	1.25
	m_3	0.06	0.20	0.20	1.99×10^{17}	1.91
	ρ_1	0.01	0.05	0.05	0.12	0.42
Norwegian	m_1	0.00	15.76	15.76	$8.81 \times 10^{9*}$	68.31*
	m_2	0.02	23.95	23.94	4.01×10^{17}	19.62
	m_3	0.00	10.43	10.43	2.89×10^{24}	10.53
	ρ_1	0.01	10.74	10.74	103.74	2.07
Internet	m_1	0.07	12.68	12.68	$2.80 \times 10^{9*}$	7847.26*
	m_2	0.05	17.62	17.45	3.89×10^{17}	1.72×10^{6}
	m_3	0.00	17.96	17.84	5.16×10^{25}	3.21×10^{8}
	ρ_1	0.02	20.37	20.71	73.48	74.15
Coke	ρ_1	0.00	0.00	0.01	0.24*	0.01
Cordoba	m_1	0.00	32.34	32.34	32.34	32.24
	m_2	0.02	26.68	26.68	26.68	26.68
	m_3	0.00	16.80	16.80	16.80	16.80
	ρ_1	0.01	17.53	17.53	17.53	17.50
Gas	m_1	0.08	5.62	5.62	68,355*	87.30*
	m_2	0.05	14.26	11.47	2.20×10^{8}	57.58
	m_3	0.00	19.66	15.79	7.23×10^{11}	155
	ρ_1	0.09	2.75	3.82	8.96	18.45
Riverflow	m_1	0.02	8.76	8.67	$3.55 \times 10^{7*}$	288.76*
	m_2	0.01	5.65	3.13	3.77×10^{13}	3434
	m_3	0.01	15.97	13.34	3.37×10^{19}	21824
	ρ_1	0.03	1.89	4.75	50.73	53.07
Unemploy	m_1	0.18	11.17	11.18	$1.12 \times 10^{7*}$	19.66*
	m_2	0.06	22.67	22.36	8.06×10^{12}	70.75
	m_3	0.03	33.65	32.95	6.71×10^{18}	186
	ρ_1	0.11	1.01	0.87	0.04	1.45
Gnp	m_1	1.22	1.03	1.03	$9.64 \times 10^{5*}$	249*
	m_2	1.16	18.10	15.87	2.14×10^{10}	536
	m_3	0.54	21.56	18.26	7.79×10^{14}	2211
	ρ_1	0.13	15.68	21.75	147.26	151.86
Milk	m_1	0.01	0.67	0.68	12208*	10.34
	m_2	0.00	1.52	1.54	9.74×10^{6}	24.00
	m_3	0.00	2.52	2.53	9.38×10^{9}	41.64
	ρ_1	0.16	1.56	1.56	0.75	4.80

escapes from local optima, though at the expense of higher running times. Therefore, it is natural to gauge the benefits of such an approach instead of considering a simple local search, usually implemented by all statistical packages. To look more closely at this problem, the values of the objective function obtained with both a single run of the R-cran command optim, and the VNS approach (seven neighborhoods and two randomly generated points on each) are depicted in Table 5. The same starting point $\mathbf{x_0}$, namely the solution to Problem (P^*) discussed in Section 3, which preserves the sample mean, was used under both approaches. The improvements of the VNS on the local search are expressed as percentages. It can be observed that in eight/nine/six out of 10 series in Experiments 1/2/3, the results are similar, possibly due to the proper choice of the starting point. However, in the case of the heavy-tailed series Norwegian, and the highly variable series Riverflow and Gnp the VNS significantly improved the results in Experiment 3. Such an improvement in accuracy are obtained at the expense of an increase in running times. Indeed, the median running time for MMM among the ten considered series was, for the prototype code implemented in R, about 5.8 minutes, in contrast to the couple of seconds taken by the classic methods. In other words, local search, with an appropriate choice of the starting solution, as the one proposed here, gives a quick and usually good solution, though the accuracy is substantially improved when more computational effort is made by plugging local search into a VNS method.

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 uncomplete 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 were 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 ρ_j, particularized for $j = 1$, were also considered. First, a *weighted autocorrelation coefficient estimator* is proposed as follows: the series is split into subseries of consecutive values with length higher than j; for each such sub-series, its sample jth lag autocorrelation coefficient $\bar{\rho}_{j,n}$ is computed according to (5), and then the estimator is given by

$$\hat{\rho}_j = \sum_n \omega_n \bar{\rho}_{j,n}, \tag{7}$$

where ω_n represents the weight of the nth sub-series. However, a restriction worth to be mentioned while considering the jth 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 jth autocorrelation of the longest interval of known data can be considered as an estimator instead of the weighted one. Other estimates for ρ_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 ρ_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 ρ_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 Níjar, 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 valued 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 ρ_1, ρ_2 and ρ_3 were computed

Table 4
Absolute and quadratic differences.

Series	Exp.	diff.	MMM	Linear	Nearest	Cubic	Spline
Exchange	1	abs	1.04	**1.00**	1.27	2.94	1.14
		quad	1.05	**1.00**	1.75	31.36	1.15
	2	abs	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		quad	**1.00**	1.06	2.62	11.16	1.15
	3	abs	**1.00**	2.42	2.42	1.14×10^{8}	25.51
		quad	**1.00**	3.98	4.03	1.75×10^{16}	584.11
Norwegian	1	abs	6.81	1.08	**1.00**	2.47	1.73
		quad	2.24	1.11	**1.00**	3.24	1.47
	2	abs	4.33	1.07	**1.00**	1.74	1.63
		quad	2.22	**1.00**	**1.00**	2.04	1.32
	3	abs	1.95	**1.00**	1.01	3.96×10^{8}	3.09
		quad	1.97	**1.00**	**1.00**	1.76×10^{16}	1.55
Internet	1	abs	1.36	**1.00**	1.17	2.01	1.25
		quad	1.73	**1.00**	1.38	18.74	1.51
	2	abs	1.45	**1.00**	1.21	1.23	1.23
		quad	1.79	**1.00**	1.46	1.89	1.42
	3	abs	1.25	**1.00**	1.02	1.99×10^{8}	556.13
		quad	1.29	**1.00**	**1.00**	3.88×10^{16}	172175.02
Coke	1	abs	**1.00**	**1.00**	1.27	3.84	1.17
		quad	**1.00**	**1.00**	1.51	519.50	1.36
	2	abs	**1.00**	**1.00**	1.30	1.81	1.08
		quad	**1.00**	**1.00**	1.70	16.04	1.11
	3	abs	**1.00**	1.66	3.22	1.95×10^{9}	3.81
		quad	**1.00**	3.22	11.83	6.51×10^{18}	13.52
Cordoba	1	abs	1.48	**1.00**	1.02	1.10	1.18
		quad	1.56	1.02	1.03	1.01	**1.00**
	2	abs	1.29	**1.00**	1.03	1.25	1.39
		quad	**1.00**	1.08	1.07	1.74	1.09
	3	abs	1.97	**1.00**	**1.00**	**1.00**	**1.00**
		quad	1.99	**1.00**	**1.00**	**1.00**	**1.00**
Gas	1	abs	**1.00**	2.04	4.53	18.30	1.28
		quad	**1.00**	4.02	19.29	1795.20	1.68
	2	abs	**1.00**	7.42	23.21	19.96	4.73
		quad	**1.00**	39.43	376.53	301.51	24.74
	3	abs	**1.00**	3.52	4.20	1.15×10^{4}	14.65
		quad	**1.00**	10.03	14.06	1.75×10^{8}	153.15
Riverflow	1	abs	**1.00**	1.22	1.51	5.69	1.27
		quad	**1.00**	1.43	2.27	322.56	1.44
	2	abs	**1.00**	1.22	1.99	2.26	1.17
		quad	**1.00**	1.51	4.40	10.18	1.37
	3	abs	**1.00**	1.07	1.08	1.31×10^{6}	15.31
		quad	1.41	**1.00**	1.11	2.31×10^{12}	197.80
Unemploy	1	abs	1.06	**1.00**	1.36	2.56	1.12
		quad	1.06	**1.00**	1.77	27.69	1.14
	2	abs	1.12	**1.00**	1.48	1.75	1.17
		quad	1.09	**1.00**	1.69	2.86	1.27
	3	abs	**1.00**	1.27	1.26	1.23×10^{6}	2.52
		quad	**1.00**	1.42	1.46	2.16×10^{12}	6.17
Gnp	1	abs	1.11	**1.00**	**1.00**	5.62	1.46
		quad	1.35	**1.00**	1.01	205.98	2.10
	2	abs	**1.00**	1.31	1.27	1.33	1.80
		quad	**1.00**	2.19	2.00	2.42	4.34
	3	abs	1.35	**1.00**	**1.00**	2.93×10^{4}	7.56
		quad	1.78	**1.00**	1.01	1.27×10^{9}	44.57
Milk	1	abs	**1.00**	1.92	4.38	3.97	1.33
		quad	**1.00**	3.42	14.60	16.13	1.98
	2	abs	**1.00**	2.25	2.95	5.20	1.56
		quad	**1.00**	6.94	10.08	34.70	3.23
	3	abs	**1.00**	5.56	5.61	3.03×10^{4}	26.30
		quad	**1.00**	26.60	27.36	1.23×10^{9}	521.75

as the weighted estimators, and as the sample estimators (5) from the longest interval.

Níjar 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

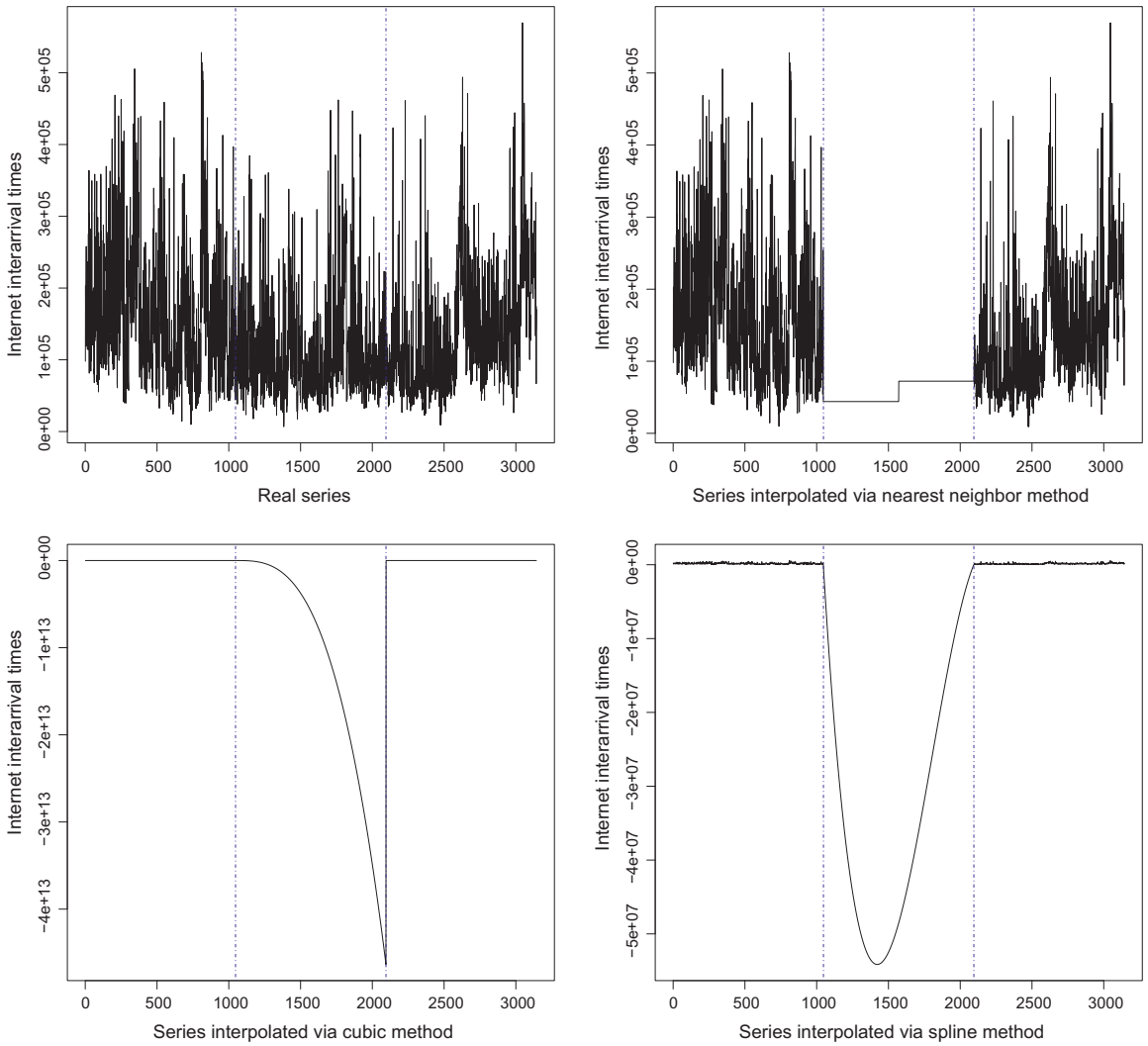

Fig. 8. *Top left:* Real `Internet` series and proportion of data removed in Experiment 3. *Top right:* interpolation by Nearest Neighbor. *Bottom left:* interpolation by Cubic. *Bottom right:* interpolation by Spline.

Table 5
Comparison between a local search versus VNS approach in terms of objective function values.

	Experiment 1			Experiment 2			Experiment 3		
	`optim`	VNS	Improv. (%)	`optim`	VNS	Improv. (%)	`optim`	VNS	Improv. (%)
`Exchange`	1.03	1.03	0.01	1.00	1.00	0.00	1.02	1.02	0.01
`Norwegian`	2.15	1.83	14.69	1.15	1.15	0.00	1025.46	1.18	99.88
`Internet`	1.65	1.64	0.86	1.26	1.25	0.16	2.59	1.39	46.14
`Coke`	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.35
`Cordoba`	2.87	2.87	0.04	1.81	1.81	0.00	1.46	1.46	0.00
`Gas`	1.10	1.09	0.70	1.03	1.00	0.00	1.58	1.56	1.11
`Riverflow`	1.40	1.40	0.00	1.09	1.09	0.02	3.46	1.36	60.47
`Unemploy`	1.10	1.10	0.00	1.05	1.05	0.00	1.21	1.17	3.68
`Gnp`	14.39	1.60	88.82	5.04	1.38	72.55	33.42	3.72	88.87
`Milk`	1.05	1.05	0.00	1.09	1.09	0.00	1.52	1.51	0.41

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 nonconvex optimization problem, solved via a VNS approach in a continuous space.

Table 6
Interpolation by MMM under different target values for a simulated series from an AR (1) process.

Experiment		0% Dev	5% Dev	10% Dev	15% Dev	Weighted	Longest
1	abs diff	249.47	244.46	248.20	268.97	460.26	346.92
	quad diff	330.33	319.25	332.18	388.24	1150.74	649.42
2	abs diff	106.96	107.63	141.59	221.19	232.91	110.59
	quad diff	125.84	126.14	203.88	559.14	621.91	136.50
3	abs diff	329.84	339.47	376.51	413.83	326.37	328.98
	quad diff	508.79	528.34	633.66	758.52	502.15	508.77

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 ($P^\star$)

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

$$\begin{cases} \min & \sum_{i=1}^{N-1}(x_{i+1}-x_i)^2 & \\ \text{s.t.} & x_i = y_i, & \forall i \in S, \\ & \frac{\sum_{i=1}^{N} x_i}{N} = m_1. & \end{cases} \qquad (P^{\star\star})$$

The Lagrangian function is given by

$$L(x,\lambda) = \sum_{i=1}^{N-1}(x_{i+1}-x_i)^2 + \lambda\left(\frac{\sum_{i=1}^{N} x_i}{N} - m_1\right). \qquad (8)$$

As shown next, the solution to ($P^\star$) depends on the number of intervals containing missing values and their lengths. In order to solve ($P^\star$), 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 $R = r + 1 \geqslant 3$. The partial derivatives of the Lagrangian function are given by

$$\begin{cases} \frac{\partial L}{\partial x_p} = 4x_p - 2y_{p-1} - 2x_{p+1} + \frac{\lambda}{N}, & \text{if } j = 0, \\ \frac{\partial L}{\partial x_{p+j}} = 4x_{p+j} - 2x_{p+j-1} - 2x_{p+j+1} + \frac{\lambda}{N}, & \text{if } 1 \leqslant j \leqslant r-1, \\ \frac{\partial L}{\partial x_{p+r}} = 4 + x_{p+r} - 2x_{p+r-1} - 2y_{p+r+1} + \frac{\lambda}{N}, & \text{if } j = r. \end{cases}$$

Making the Lagrangian function equal to zero, we obtain

$$x_{p+j} = (j+1)x_p - jy_{p-1} + \frac{j(j+1)}{4N}\lambda, \qquad (9)$$

for $1 \leqslant j \leqslant r-1$. 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-1}, we conclude that

$$x_{p+r} = (r+1)x_p - ry_{p-1} + \frac{r(r+1)}{4N}\lambda. \qquad (10)$$

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

$$x_{p+r} = \frac{r}{2}x_p - \frac{r-1}{2}y_{p-1} + \frac{1}{2}y_{p+r+1} + \frac{r^2-r-2}{8N}\lambda. \qquad (11)$$

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

$$x_p = \frac{r+1}{r+2}y_{p-1} - \frac{1}{r+2}y_{p-1} + \frac{-r^2-3r-2}{4N(r+2)}\lambda. \qquad (12)$$

From the active constraint

$$\frac{\sum_{i=1}^{N} x_i}{N} = m_1,$$

it follows that

$$x_p + \sum_{j=1}^{r-1} x_{p+j} + x_{p+r} = c, \qquad (13)$$

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) $L_1 \geqslant 2$ intervals of missing values.
Assume in this case a number $L_1 \geqslant 2$ 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_n of length $R_n = r_n + 1$, for $1 \leqslant n \leqslant L_1$,

$$\begin{cases} x_{p_n} = \frac{r_n+1}{r_n+2}y_{p_n-1} - \frac{1}{r_n+2}y_{p_n-1} + \frac{-r_n^2-3r_n-2}{4N(r_n+2)}\lambda \\ x_{p_n+r_n} = (r_n+1)x_{p_n} - r_n y_{p_n-1} + \frac{r_n(r_n+1)}{4N}\lambda \\ x_{p_n+r_n} = \frac{r_n}{2}x_{p_n} - \frac{r_n-1}{2}y_{p_n-1} + \frac{1}{2}y_{p_n+r_n+1} + \frac{r_n^2-r_n-2}{8N}\lambda \\ x_{p_n+j} = (j+1)x_{p_n} - jy_{p_n-1} + \frac{j(j+1)}{4N}\lambda, \quad 1 \leqslant j \leqslant r_n, \end{cases}$$

where

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

where $\gamma_n = \frac{-r_n^3-2r_n^2-3r_n-2+2\tau_n}{4N}$ and $\tau_n = \sum_{j=1}^{r_n-1} j(j+1)$.

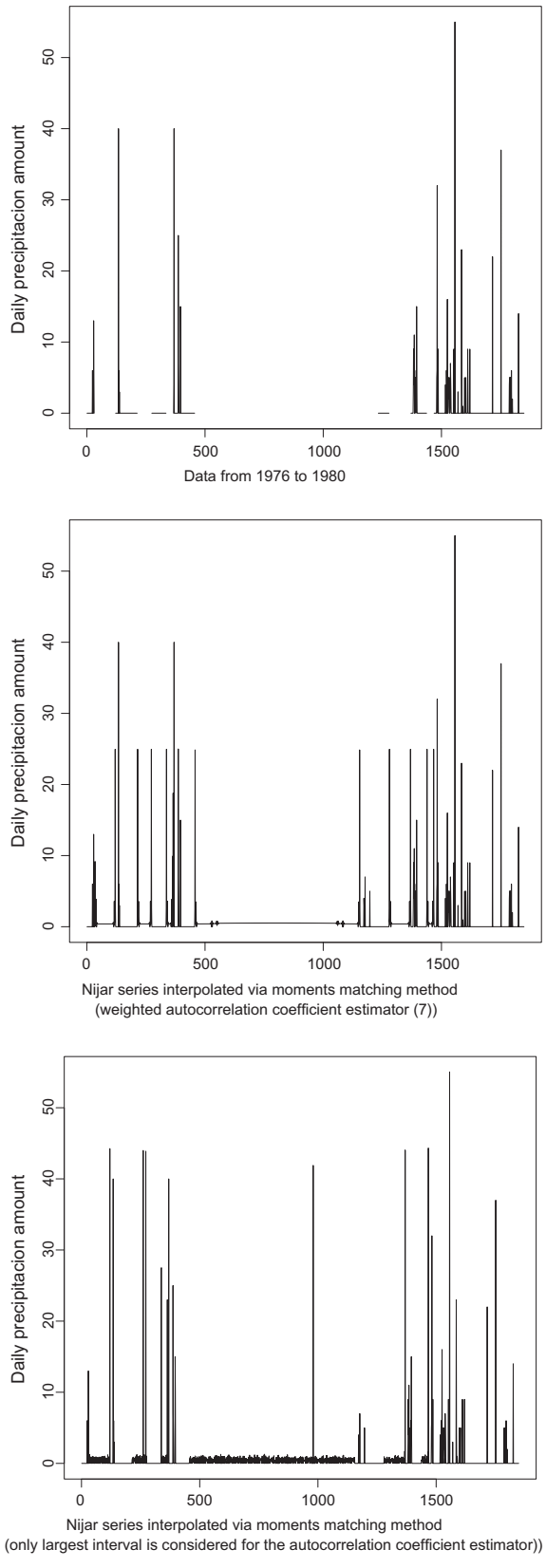

Fig. 9. Interpolation of Nijar rainfall amounts data via MMM.

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 $2 \leqslant p \leqslant N-2$.
In the same manner as in the previous calculations we can see that

$$x_p = \frac{2}{3}y_{p-1} + \frac{1}{3}y_{p+2} - \frac{1}{2N}\lambda,$$
$$x_{p+1} = \frac{2}{3}y_{p+2} + \frac{1}{3}y_{p-1} - \frac{1}{2N}\lambda,$$

where in this case

$$\lambda = (-c + y_{p-1} + y_{p+2})N,$$

(Case 2.b) L_2 intervals of missing values of length 2.
Assume $I_1, \ldots, I_{L_2}$ intervals of two missing values in the time series $\mathbf{x}$. The reasoning above applies to this case to obtain, for $1 \leqslant n \leqslant L_2$,

$$x_{p_n} = \frac{2}{3}y_{p_n-1} + \frac{1}{3}y_{p_n+2} - \frac{1}{2N}\lambda,$$
$$x_{p_n+1} = \frac{2}{3}y_{p_n+2} + \frac{1}{3}y_{p_n-1} - \frac{1}{2N}\lambda,$$

where

$$\lambda = \frac{c - \sum_{n=1}^{L_2}(y_{p_n-1} + y_{p_n+1})}{\frac{-L_2}{N}},$$

(Case 2.c) A unique missing data.
When there exists one single missing data x_p in the time series $\mathbf{x}$, it is easily seen that

$$x_p = \frac{1}{2}y_{p-1} + \frac{1}{2}y_{p+1} - \frac{1}{4N}\lambda, \tag{14}$$

for $2 \leqslant p \leqslant N-1$, where

$$\lambda = (-2c - y_{p-1} - y_{p+1})2N,$$

(Case 2.d) L_3 isolated missing data.
Finally, assume a set of isolated missing data $\{x_i\}_{i\in\mathcal{I}}$, where the length of $\mathcal{I}$ is L_3. Again, it is straightforward to obtain

$$x_{p_n} = \frac{1}{2}y_{p_n-1} + \frac{1}{2}y_{p_n+1} - \frac{1}{4N}\lambda \quad con \quad p_n \epsilon I,$$

for any $2 \leqslant j \leqslant N-1$ and where

$$\lambda = \frac{2c - \sum_{n=1}^{L_3}(y_{p_n-1} - y_{p_n+1})}{\frac{-L_3}{2N}},$$

We end this Appendix by putting all the previous results together, giving a closed-form expression for the solution to ($P^\star$). Consider a time series which has L_1 intervals of missing data of length less than or equal to 3, L_2 intervals of length 2, and L_3 isolated missing values. Let $|B|$ 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 λ, which has to be the same for all expressions. Taking into account the active constraint to preserve the mean

$$\sum_{n=1}^{L_1}\sum_{j=1}^{n}(x_{p_n} + x_{p_n+j} + x_{p_n+r_n}) + \sum_{n=1}^{L_2}(x_{p_n} + x_{p_n+1}) + \sum_{n=1}^{L_3}x_{p_n} = c,$$

we obtain

$$\lambda = \frac{2c - \sum_{n=1}^{L_1}((r_n+1)y_{p_n-1} + (r_n+1)y_{p_n+r_n+1}) - \sum_{n=1}^{L_2}(2y_{p_n-1} + 2y_{p_n+2}) - \sum_{n=1}^{L_3}(y_{p_n-1} + y_{p_n+1})}{\sum_{n=1}^{L_1}\gamma_n - \frac{2L_2}{N} - \frac{L_3}{N}},$$

where $c = Nm_1 - \sum_{i \in S} y_i$ and

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

References

[1] B. Abraham, J. Ledolter, Statistical methods for forecasting, Wiley Series in Probability and Statistics (1983).
[2] A. Alonso, A. Sipols, A time series bootstrap procedure for interpolation intervals, Computational Statistics & Data Analysis 52 (2008) 1792–1805.
[3] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences (2013), http://dx.doi.org/10.1016/j.ins.2013.01.021.
[4] Z. Bar-Joseph, Analyzing time series gene expression data, Bioinformatics 20 (2004) 2493–2503.
[5] J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, Statistics of extremes: theory and applications, Wiley, 2004.
[6] S. Beveridge, Least squares estimation of missing values in time series, Communications in Statistics Theory Methods 21 (1992) 3479–3496.
[7] C. Boor, A Practical Guide to Splines, Springer-Verlag, 1978.
[8] E. Borensztein, J. Zetterlmeyer, T. Philippon, Monetary independence in emerging markets: does the exchange rate regime make a difference?, International Monetary Found (2001)
[9] G. Box, G. Jenkins, G. Reinsel, Time Series Analysis: Forecasting and Control, Prentice Hall, Engewood Cliffs, NJ, 1994.
[10] E. Carrizosa, M. Dražić, V. Dražić, N. Mladenović, Gaussian variable neighborhood search for continuous optimization, Computers and Operations Research 39 (2012) 2206–2213.
[11] C. Chatfield, The Analysis of Time Series: An Introduction. Texts in Statistical Science Series, Taylor and Francis Group, 2004.
[12] Z. Chen, Z. Fan, M. Sun, A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, European Journal of Operational Research 223 (2012) 461–472.
[13] J. Cryer, Time Series Analysis, Duxbury Press, Boston, 1986.
[14] M. Dorigo, Optimization, Learning and Natural Algorithms, Ph.D. thesis Politecnico di Milano, Italy, 1992.
[15] W. Dunsmuir, B. Murtagh, Least absolute deviation estimation of stationary time series models, European Journal of Operational Research 67 (1993) 272–277.
[16] F. Glover, Heuristics for integer programming using surrogate constraints, Decision Sciences 8 (1977) 156–166.
[17] V. Gómez, A. Maravall, D. Peña, Missing observations in ARIMA models: Skipping approach versus additive outlier approach, Journal of Econometrics 88 (1999) 341–363.
[18] P.G. Gould, A.B. Koehler, J.K. Ord, R.D. Snyder, R.J. Hyndman, F. Vahid-Araghi, Forecasting time series with multiple seasonal patterns, European Journal of Operational Research 191 (2008) 207–222.
[19] A.D. Gronewold, C.A. Stow, J.L. Crooks, T.S. Hunter, Quantifying parameter uncertainty and assessing the skill of exponential dispersion rainfall simulation models, International Journal of Climatology 33 (2012) 746–757, http://dx.doi.org/10.1002/joc.3469.
[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Prediction, Inference and Data Mining, Springer Verlag, New York, 2009.
[21] K. Hipel, A. McLeod, Time Series Modelling of Water Resources and Environmental SYSTEMs, Elsevier, 1994.
[22] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.
[23] R. Jones, Maximum likelihood fitting of ARMA models to time series with missing observations, Technometrics 22 (1980) 389–395.
[24] S. Kirkpatrick Jr., C. GGelatt, M. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[25] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the selfsimilar nature of ethernet traffic (extended version), IEEE/ACM Transactions on Networking 2 (1994) 1–15.
[26] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, 2002.
[27] G.M. Ljung, A note on the estimation of missing values in time series, Communications in Statistics – Simulation and Computation 18 (1989) 459–465.
[28] N. Mladenović, M. Dražić, V. Kovačevic-Vujčić, M. Čangalović, General variable neighborhood search for the continuous optimization, European Journal of Operational Research 191 (2008) 753–770.
[29] N. Mladenović, P. Hansen, Variable neighborhood search, Computers and Operations Research 24 (1997) 1097–1100.
[30] W. Palma, Long-memory time series, Theory and Methods, Wiley, 2007.
[31] D. Peña, G. Tiao, R. Tsay, A course in time series analysis, Wiley Series in Probability and Statistics (2001).
[32] M. Pourahmadi, Estimation and interpolation of missing values of a stationary time series, Journal of Time Series Analysis 10 (1989) 149–169.
[33] P. Ramirez-Cobo, R. Lillo, S. Wilson, M. Wiper, Bayesian inference for double Pareto lognormal queues, The Annals of Applied Statistics 4 (2010) 1533–1557.
[34] A. Stuart, H. Panjer, G. Willmot, Loss Models: From Data to Decisions, Wiley, 2008.
[35] P. Thévenaz, T. Blu, M. Unser, Interpolation revisited, IEEE Transactions on Medical Imaging 19 (2000) 739–758.
[36] N.F. Thornhill, M.M. Naim, An exploratory study to identify rogue seasonality in a steel company's supply network using spectral principal component analysis, European Journal of Operational Research 172 (2006) 146–162.
[37] W. Wong, M. Xia, W. Chu, Adaptive neural network model for time-series forecasting, European Journal of Operational Research 207 (2010) 807–816.
[38] E. Zivot, J. Wang, Modeling Financial Time Series with S-PLUS, Birkhäuser, 2005.