

A comparison of clustering approaches for the study of the temporal coherence of multiple time series

Francesco Finazzi · Ruth Haggarty ·
Claire Miller · Marian Scott · Alessandro Fassò

© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract Two approaches for clustering of time series have been considered. The first is a novel approach based on a modification of classic state-space modelling while the second is based on functional clustering. For the latter, both k-means and complete-linkage hierarchical clustering algorithms are adopted. The two approaches are compared using a simulation study, and are applied to lake surface water temperature for 256 lakes globally for 5 years of data, to investigate information obtained from each approach.

Keywords State space · Expectation maximization · Functional data analysis · Splines

Electronic supplementary material The online version of this article (doi:10.1007/s00477-014-0931-2) contains supplementary material, which is available to authorized users.

F. Finazzi (✉)
Department of Management, Economics and Quantitative
Methods, University of Bergamo, via dei Caniana 2,
24127 Bergamo, Italy
e-mail: francesco.finazzi@unibg.it

R. Haggarty · C. Miller · M. Scott
School of Mathematics and Statistics, University of Glasgow, 15
University Gardens, Glasgow G12 8QW, UK
e-mail: ruth.haggarty@glasgow.ac.uk

C. Miller
e-mail: claire.miller@glasgow.ac.uk

M. Scott
e-mail: marian.scott@glasgow.ac.uk

A. Fassò
Department of Engineering, University of Bergamo, viale
Marconi 5, 24044 Dalmine, Italy
e-mail: alessandro.fasso@unibg.it

1 Introduction

In environmental and ecological sciences, the correlation or synchrony between major fluctuations in a set of time series is often described as temporal coherence, (Lansac-Tha 2008; Livingstone 2010; Salisbury et al. 2011). If synchronous or coherent temporal patterns are observed, then this may indicate the existence of common drivers and pressures. Increasingly within ecology, there is a need for statistical models which do not regard the individual time series separately but rather recognise that common drivers will impact at regional and sub-regional spatial scales. Commonly it is the case that the sites at which the time series' are measured are spatially registered, so that identification of a set of temporally coherent time series can be further explored spatially.

In this brief introduction, we focus on the freshwater environment, specifically lakes. Globally, lakes are considered as sensitive indicators of environmental change, impacted by both natural and anthropogenic drivers of change. In particular the impact of climate change on freshwater resources is critical and IPCC, UNEP and EEA have all recognised the sensitivity of the global water cycle to climate change and other pressures. Improved understanding of the observed changes is key to better management of aquatic resources. Such changes include synchrony in the fluctuations observed, and also in the changing seasonal patterns. Studies exploring the temporal coherence of lakes in terms of hydrological features (flow), bio-geochemistry (pH, alkalinity, chlorophyll, sulphates and nitrates, organic carbon) and temperature are widely undertaken. Each of these variables in turn respond to global and regional covariates such as the North Atlantic Oscillation, land management, global temperature and precipitation.

Classically in the ecological literature, the focus has often been on a small number of time series, and the analysis to find common patterns has used a pairwise approach (often with a simple correlation coefficient). Other approaches have made use of cross-wavelet analysis (Grinsted et al. 2004; Labat 2010; Franco-Villoria et al. 2012) but still with a focus on a pairwise approach.

In a multiple time series setting, dynamic factor analysis has been used (Calder 2007; Lopes et al. 2011; Muoz-Carpena 2005) to identify common latent trends and for prediction. In this work, we focus on clustering as an approach to study the temporal coherence of multiple time series, with a view to establishing methods that are appropriate for any number of time series. In particular, a novel clustering algorithm based on a modification to the approach of state-space modelling is proposed and is compared to functional clustering considering both k-means and complete-linkage hierarchical algorithms. The idea of combining state-space modelling and clustering is not new and has been considered in Costa and Goncalves (2011). The approach developed in Costa and Goncalves (2011), however, seems to be suitable for small numbers of time series, it is based on univariate models and does not provide a way to estimate the optimal number of clusters. The clustering approaches in this paper are illustrated on a global lake temperature data set (see MacCallum and Merchant 2013).

The rest of the paper is organized as follows: in Section 2, the concept of temporal coherence is defined. Sections 3 and 4 describe the state-space model at the basis of the clustering approach and its estimation by means of a modified version of the EM algorithm. Section 5 introduces the functional clustering approach considering both the k-means and the complete-linkage hierarchical algorithms. Section 6 compares the novel clustering approach with functional clustering and compares the performances of both the approaches when a simulated data set is considered. Section 7 describes the clustering result for the global lake temperature data set while conclusions are given in Section 8.

2 Study of temporal coherence

In this paper, we consider a set of time series to be jointly coherent when, apart from random noise, they share the same temporal pattern along the entire temporal frame of observation. In particular, the term “temporal pattern” refers to the direction of variation of the time series, and the fixed characteristics of the time series, such as the overall mean and the overall variability, are not considered to be discriminant. For this reason, only standardized time series will be analysed.

A natural way to study temporal coherence is to group the time series into a suitable number of coherency clusters, that is, two time series belong to the same cluster if they are coherent with each other. The coherency study, therefore, consists in the estimation of both the number of clusters and the membership of each time series with respect to the clusters.

Although the paper deals with spatially registered time series, the spatial correlation across time series is not explicitly modelled or forced in any way. The approaches discussed in this paper, instead, enable spatio-temporal data to be modelled where the interest is natural clusters of seasonal patterns. When the results of these approaches are mapped in geographic space they enable better understanding of the spatial context of the underlying natural processes.

3 State space modelling

State-space modelling is a time series analysis technique which is used to identify latent common temporal patterns in time series. The minimal state-space model is the following

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{K}\mathbf{z}(t) + \boldsymbol{\varepsilon}(t) \\ \mathbf{z}(t) &= \mathbf{G}\mathbf{z}(t-1) + \boldsymbol{\eta}(t) \end{aligned} \quad (1)$$

where $\mathbf{y}(t)$ is the $N \times 1$ observation vector and $\mathbf{z}(t) = (z_1(t), \dots, z_p(t))'$ is the $p \times 1$ state vector, with $\mathbf{z}(0) \sim N(\mathbf{v}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0$ is a known variance-covariance matrix. The matrix \mathbf{K} is a $N \times p$ matrix of coefficients while \mathbf{G} is a $p \times p$ stable transition matrix. Finally, $\boldsymbol{\varepsilon}(t) \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_N)$ is the $N \times 1$ measurement error vector while $\boldsymbol{\eta}(t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ is the $p \times 1$ innovation vector. If $\mathbf{y}(t)$ includes missing data, then $\mathbf{y}(t) = \mathbf{L}_t(\mathbf{y}^{(1)}(t)', \mathbf{y}^{(2)}(t)')'$, where $\mathbf{y}^{(1)}(t)$ and $\mathbf{y}^{(2)}(t)$ denotes the sub-vectors of the data and the missing data respectively while \mathbf{L}_t is the permutation matrix at time t . Moreover, $n_t^{(1)} + n_t^{(2)} = N$ and $n_t^{(2)}$ denotes the number of missing values in $\mathbf{y}(t)$. Model (1) is completely characterized by the parameter set $\Psi = \{\mathbf{K}, \mathbf{G}, \boldsymbol{\Sigma}_\eta, \sigma_\varepsilon^2, \mathbf{v}_0\}$.

The idea behind the state-space model of Eq. (1) is to model each time series $\{y_i(t)\}$, $i = 1, \dots, N$ as a linear combination of the latent time series $\{z_j(t)\}$, $j = 1, \dots, p$, with weights of the linear combinations given by the row \mathbf{k}_i of \mathbf{K} .

3.1 Model estimation

Given the $N \times T$ matrix $\mathbf{Y} = (\mathbf{y}(1), \dots, \mathbf{y}(T))$, the estimation problem consists in estimating both the parameter set

Ψ and the latent time series $\{\mathbf{z}(t)\}$. The Expectation Maximization algorithm in conjunction with the Kalman smoother algorithm represents a well known and largely accepted solution to the estimation problem within the maximum likelihood framework (see Shumway and Stoffer (2006)). In order to make the model identifiable, however, constraints must be imposed on the parameter set. For instance, (Fassò and Finazzi 2011) consider a matrix \mathbf{K} of fixed coefficients, (Mardia et al. 1998) estimate \mathbf{K} using empirical orthogonal functions, (Calder 2007) use known smoothing kernel convolution weights while Zuur et al. (2007) introduce restrictions on \mathbf{K} , Σ_η or \mathbf{v}_0 .

Assuming, for the moment, that no constraints are imposed on Ψ , the closed form updating formulas at iteration m of the EM algorithm are the following

$$\begin{aligned} \hat{\Sigma}_\eta^{(m)} &= \frac{1}{T} \left(\mathbf{S}_{11} - \mathbf{S}_{10} \hat{\mathbf{G}}^{(m-1)} \mathbf{S}'_{10} \right) \\ \hat{\mathbf{G}}^{(m)} &= \mathbf{S}_{10} \mathbf{S}_{00}^{-1} \\ (\hat{\sigma}_\varepsilon^2)^{(m)} &= \frac{1}{NT} \text{tr} \sum_{t=1}^T \mathbf{L}_t \begin{pmatrix} \mathbf{M}_t^{(m-1)} & \mathbf{0}_{n_t^{(1)} \times n_t^{(2)}} \\ \mathbf{0}_{n_t^{(2)} \times n_t^{(1)}} & (\hat{\sigma}_\varepsilon^2)^{(m-1)} \mathbf{I}_{n_t^{(2)}} \end{pmatrix} \mathbf{L}_t' \quad (2) \\ \hat{\mathbf{K}}^{(m)} &= \left(\sum_{t=1}^T \mathbf{L}_t \begin{pmatrix} \mathbf{y}^{(1)}(t) \cdot (\mathbf{z}_t^T)' \\ \mathbf{0}_{n_t^{(2)} \times 1} \end{pmatrix} \right) \mathbf{S}_{11}^{-1} \end{aligned}$$

where

$$\begin{aligned} \mathbf{S}_{11} &= \sum_{t=1}^T \mathbf{z}_t^T (\mathbf{z}_t^T)' + \mathbf{P}_t^T \\ \mathbf{S}_{10} &= \sum_{t=1}^T \mathbf{z}_t^T (\mathbf{z}_{t-1}^T)' + \mathbf{P}_{t,t-1}^T \\ \mathbf{S}_{00} &= \sum_{t=1}^T \mathbf{z}_{t-1}^T (\mathbf{z}_{t-1}^T)' + \mathbf{P}_{t-1}^T \\ \mathbf{M}_t^{(m-1)} &= \left(\mathbf{y}^{(1)}(t) - \mathbf{L}_t \hat{\mathbf{K}}^{(m-1)} \mathbf{z}_t^T \right) \left(\mathbf{y}^{(1)}(t) - \mathbf{L}_t \hat{\mathbf{K}}^{(m-1)} \mathbf{z}_t^T \right)' \\ &\quad + \mathbf{L}_t \hat{\mathbf{K}}^{(m-1)} \mathbf{P}_t^T \left(\mathbf{L}_t \hat{\mathbf{K}}^{(m-1)} \right)' \end{aligned} \quad (3)$$

and where

$$\begin{aligned} \mathbf{z}_t^T &= \mathbb{E}_{\Psi^{(m-1)}}(\mathbf{z}(t) \mid \mathbf{Y}) \\ \mathbf{P}_{t-h}^T &= \text{Var}_{\Psi^{(m-1)}}(\mathbf{z}(t-h) \mid \mathbf{Y}); \quad h = 0, 1 \\ \mathbf{P}_{t,t-1}^T &= \text{cov}_{\Psi^{(m-1)}}(\mathbf{z}(t), \mathbf{z}(t-1) \mid \mathbf{Y}) \end{aligned}$$

are the output of the Kalman smoother at iteration $m - 1$ of the EM algorithm. Note that, in Eq. (2), $\mathbf{I}_{n_t^{(2)}}$ is the identity matrix of dimension $n_t^{(2)}$ and $\mathbf{0}_{n_t^{(1)} \times n_t^{(2)}}$ is the matrix of all zeros of dimension $n_t^{(1)} \times n_t^{(2)}$. At convergence, the EM algorithm provides the estimated model parameter set $\hat{\Psi} = \{ \hat{\mathbf{K}}, \hat{\mathbf{G}}, \hat{\Sigma}_\eta, \hat{\sigma}_\varepsilon^2, \hat{\mathbf{v}}_0 \}$.

4 A novel model-based clustering approach

In classic state-space modelling, the $p \ll N$ components of the latent vector $\mathbf{z}(t)$ represent the common temporal trends and the role of the matrix \mathbf{K} is to express each time series $\{y_i(t)\}$ as a linear combination of the common trends. If the aim is to cluster the N time series with respect to their temporal coherence, the role of the j -th component of $\mathbf{z}(t)$ is to describe only the time series of the j -th cluster. Assuming standardized time series, this is equivalent to requiring the matrix \mathbf{K} to have elements which can only be zeros and ones. In particular, each row \mathbf{k}_i of \mathbf{K} contains a single element equal to one and the position of this element identifies the membership of the time series with respect to the clusters.

At this point, it is important to note that the updating formula of Eq. (2) is not able to provide such a constrained matrix. In principle, the maximum likelihood estimation of Ψ by means of the EM algorithm can be carried out considering the constrained parameter space but it is not easy to derive closed form estimation formula. For each iteration of the EM algorithm, on the other hand, an exhaustive search of the constrained matrix \mathbf{K} that maximizes the likelihood (conditional on the other model parameters) is prohibitive as the space $\mathcal{K} \ni \mathbf{K}$ of all the $N \times p$ constrained matrices contains p^N elements. Since, in practical applications, N can be large (10^2 – 10^6), we believe that even relying on optimization methods (such as the simulated annealing algorithm) is not enough to obtain estimation results in a reasonable time since the optimization method should be applied for each iteration of the EM algorithm. In the next paragraph, the classic EM algorithm is adjusted so that the estimated matrix $\hat{\mathbf{K}}$ meets the above mentioned constraint but the computational burden of model estimation is not increased.

4.1 The modified EM algorithm

In order to adapt the EM algorithm so that a constrained matrix $\hat{\mathbf{K}} \in \mathcal{K}$ is estimated, it is useful to understand how the matrix $\hat{\mathbf{K}}^{(m)}$ is derived at the iteration m of the EM algorithm. By considering Eq. (2), it can be noted that the ij -th element $\hat{k}_{ij}^{(m)}$ of $\hat{\mathbf{K}}^{(m)}$ is obtained by evaluating a weighted cross-covariance between the observed time series $\{y_i(t)\}$ and the estimated time series $\{z_{j,t}^T\} = \{\mathbb{E}_{\Psi^{(m-1)}}(z_j(t) \mid \mathbf{Y})\}$. In the trivial case of $N = p = 1$ and no missing data, in fact, the scalar $\hat{\mathbf{K}}^{(m)} \equiv \hat{k}^{(m)}$ is given by

$$\hat{k}^{(m)} = \frac{\sum_{t=1}^T y(t) \cdot z_t^T}{\sum_{t=1}^T (z_t^T)^2 + p_t^T}$$

Intuitively, $\hat{k}_{ij}^{(m)}$ is high (low) when the cross-covariance between $y(t)$ and z_t^T is high (low).

In order to estimate \mathbf{K} such that $\hat{\mathbf{K}} \in \mathcal{K}$, the following strategy is considered. At iteration m of the EM algorithm, the ij -th element of $\hat{\mathbf{K}}^{(m)}$ is given by

$$k_{ij}^{(m)} = \frac{c_{ij}}{\sum_{j=1}^p c_{ij}} \tag{4}$$

where

$$c_{ij} = \begin{cases} \langle \{y_i(t)\}, \{z_{j,t}^T\} \rangle^{f(m)} & \text{if } \langle \{y_i(t)\}, \{z_{j,t}^T\} \rangle > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In Eq. (5), $\langle \cdot, \cdot \rangle$ is the correlation operator while $f(m)$ is a function of iteration number m . In the simplest case, $f(m) \equiv m$ so that $\langle \{y_i(t)\}, \{z_{j,t}^T\} \rangle, j = 1, \dots, p$ is raised to the power of m which is appropriate here. Other choices of $f(m)$ are possible provided that $f(m)$ monotonically increases with m .

When $\langle \{y_i(t)\}, \{z_{j,t}^T\} \rangle$ is raised to a power greater than one, the differences between the correlations are amplified and, for each row vector $\mathbf{k}_i^{(m)}$, due to the normalization in Eq. (4), only one element of $\mathbf{k}_i^{(m)}$ converges to 1 when $m \rightarrow \infty$. Even if, in general, $\hat{\mathbf{K}}^{(m)} \notin \mathcal{K}$, in practice, with the exception of rounding errors, the matrix $\hat{\mathbf{K}}^{(m)}$ converges to an element of the space \mathcal{K} after a small number of iterations.

Once the parameters $\hat{\Psi}$ are estimated, the matrix $\hat{\mathbf{K}}$ directly gives the membership of the N time series with respect to the p clusters. The role of the exponent $f(m)$ in Eq. (5) is similar to the ‘‘temperature’’ parameter of the simulated annealing algorithm. In particular, $f(m)$ is gradually increased with m in order to avoid convergence to poor local maxima of the likelihood function. This is necessary for two reasons: first, the matrix \mathbf{K} is jointly estimated with the rest of the model parameters in Ψ and with the latent $\{z(t)\}$. Secondly, \mathbf{K} is randomly generated when the initial value $\Psi^{(0)}$ of Ψ is set.

Note that the estimation heuristic defined by (4) and (5) does not guarantee that the EM algorithm converges to a global maximum of the likelihood function. However, the same holds for the unconstrained parameter set Ψ and the standard EM algorithm. Moreover, the same estimation heuristic does not guarantee that the likelihood of the observed data does not decrease when moving from $\hat{\Psi}^{(m)}$ to $\hat{\Psi}^{(m+1)}$, a condition which is satisfied by the standard EM

algorithm. Nonetheless, the heuristic is able to provide sound estimation results at the same computational burden of the standard EM algorithm. Poor local maxima can be avoided by repeatedly perturbing $\Psi^{(0)}$ and by considering the estimated parameter set $\hat{\Psi}$ related to the highest likelihood. Finally, it is worth noting that, as soon as the matrix $\hat{\mathbf{K}}^{(m)}$ stabilizes, the algorithm proceeds as the standard EM algorithm with all its properties.

5 Functional clustering

In the functional clustering approach, time series are described in terms of linear combinations of basis functions. The coefficient vectors of the linear combinations are then clustered using a suitable clustering algorithm, here the k-means and complete-linkage hierarchical algorithms will be implemented.

The observed time series are described through the following model

$$y_i(t) = G_i(t) + \varepsilon_i(t)$$

where G_i is a smooth curve and ε_i is an independent random error term, $i = 1, \dots, N$.

The curve G_i is a spline function of degree d (see de Boor (2001)). Since any spline function can be expressed as a linear combination of B-splines, the following functional form for the spline $s_i(t; \beta_i)$ is considered:

$$s_i(t; \beta_i) = \sum_{l=1}^{K+d-1} \beta_{i,l} B_l(t)$$

where $\beta_i = (\beta_{i,1}, \dots, \beta_{i,K+d-1})'$ is a vector of real-valued coefficients, $(B_1(t), \dots, B_{K+d-1}(t))$ is the B-spline basis functions and K is the number of knots and d is the degree of the polynomial.

As detailed in Ignaccolo et al. (2008), the β_i vector is estimated by means of the least squares method and the G_i curve is approximated by $\hat{G}_i(t) = s_i(t; \hat{\beta}_i)$.

If the polynomial degree d , the number of knots K and the knot positions are the same for all the time series, then the B-spline basis functions are fixed and the spline coefficients β_i describe the same features for each of the time series.

Two clustering algorithms are considered here, namely the k-means algorithm and the complete-linkage hierarchical algorithm.

5.1 K-means algorithm

Functional clustering based on the k-means algorithm has been introduced in Abraham et al. (2003) and a similar approach which used partitioning around medoids rather than means has been applied in Ignaccolo et al. (2008). K-

means is applied to the spline coefficient vectors in the \mathfrak{R}^{K+d-1} space and the clustering result directly provides the clustering of the time series. For a given number of clusters, in order to reduce the influence of the starting values, the k-means algorithm is applied M times.

5.2 Complete-linkage hierarchical algorithm

In the complete-linkage hierarchical clustering algorithm (Henderson 2006), the distance between the curves $G_i(t)$, $i = 1, \dots, N$ is first estimated. The distance between two curves (denoted i and q) can be written as

$$d_{iq} = (\beta_i - \beta_q)' W (\beta_i - \beta_q) \tag{6}$$

In the above expression, W is a symmetric matrix the elements of which are given by $w_{l,m} = \int B_l(t) B_m(t)' dt$, with $l, m = 1, \dots, K + d - 1$. For each set of basis functions, W can be evaluated using numerical integration, if necessary, and the functional distance matrix D with entries d_{iq} can be computed. Standard linkage criteria for hierarchical clustering can then be applied to the elements of D .

5.3 Stopping criteria

Well developed methods exist as to how to choose the optimal number of clusters. The L-curve and gap statistic (Tibshirani et al. 2001) approaches are considered here. Both the gap statistic and L-curve use the within cluster dispersion, W_j , to determine the number of clusters. For the L-curve approach a plot of W_j versus j is produced. As the number of clusters increases, W_j will decrease monotonically. However, the first value of j at which W_j reaches a minimum and stabilises indicates where there has been the largest increase in goodness of fit and hence which is the optimum number of clusters. The gap statistic compares the average within cluster dispersion for the observed data, with the average within cluster dispersion for a null reference distribution which assumes there is no clustering within the sites.

The L-curve is easy to compute but differences between the estimates for different numbers of clusters are not normalized for comparison and often the shape is uninformative regarding the optimal number of clusters. The gap statistic is time consuming as a result of the simulations required. However, can provide clearer guidance for the optimal number of clusters.

6 Simulation study

In order to compare the clustering approaches discussed above, a simulation study is carried out. The aim of the simulation study is to show that the novel model-based

Table 1 Model-based clustering results for the simulated data set. Observed data log-likelihood and number of empty clusters with respect to number of clusters

No. of clusters	2	3	4	5	6
Log-likelihood	10'948	12'643	12'936	13'053	13'055
No. of empty clust.	0	0	0	0	1
No. of clusters	7	8	9	10	
Log-likelihood	13'048	13'052	13'052	13'055	
No. of empty clust.	2	3	4	5	

approach performs as well as the classic clustering approach based on functional data analysis and that it can be used to detect small differences between clusters. As the main focus of interest in this work is to investigate clusters which are primarily based on differences in phenologies of the time series rather than long term trends, the following simulation model is considered.

6.1 Data generation

Five clusters are simulated by generating, for each cluster, $n_j = 5, 10, 20, 40, 80$ time series considering the equation

$$y_{k_j}^j(t) = \sin\left(\frac{2\pi}{52}t + (j-1)\varphi\right) + \varepsilon_j; \quad t = 1, \dots, 260$$

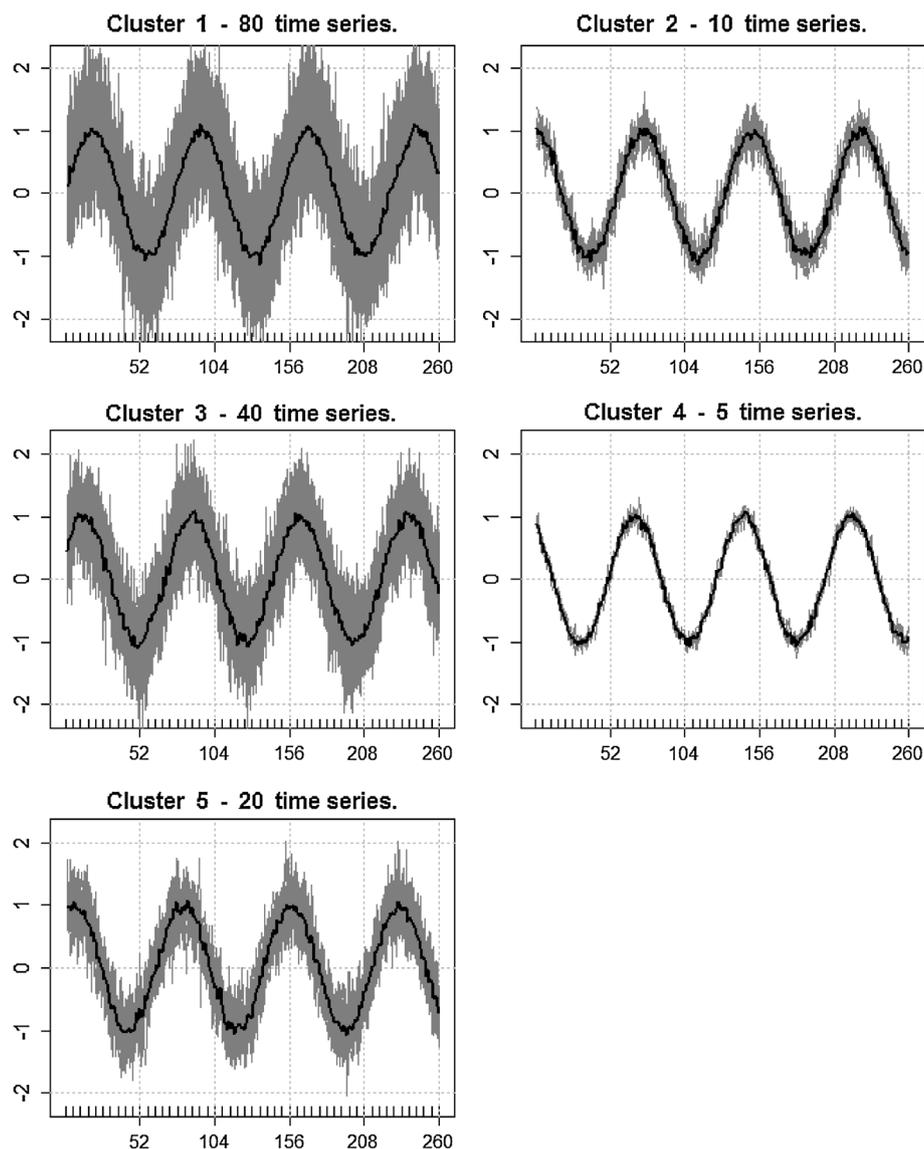
where $\varphi = \pi/6$ is a constant phase, t is time in weeks with data simulated for 5 years, $\varepsilon_j \sim N(0, \sigma_j^2)$ is a *i.i.d.* random noise with standard deviation $\sigma_j = 0.1, 0.2, 0.3, 0.4, 0.5$ and $k_j = 1, \dots, n_j$ for $j = 1, \dots, 5$. Each cluster, thus, is characterized by a different number of time series, a different phase of the sine function and a different noise variance.

6.2 Model-based clustering

The model-based clustering approach is used here to cluster the data set simulated in the previous paragraph. In order to estimate both the number of clusters p and the cluster membership, for each $p = 2, \dots, 10$ the model is estimated 50 times by perturbing the initial values of the model parameters. The solution that gives the highest observed data log-likelihood is retained. The log-likelihood of the retained solution is reported in Table 1 as a function of p where it can be noted that the log likelihood stabilizes at $p = 5$, the number of the actual clusters. In particular, for $p > 5$, the extra-clusters are empty, that is, $p - 5$ columns of the matrix $\hat{\mathbf{K}}$ are vectors of all zeros.

For $p = 5$, the number of time series in each cluster is exactly equal to $n_j, j = 1, \dots, 5$ and the cluster membership of each time series is exactly as simulated. The result is depicted in Fig. 1 in terms of both the time series of each cluster and, for each cluster, the average time series.

Fig. 1 Simulated data clustering result using the model-based approach. Individual time series (*light line*) and cluster average (*dark line*)



Two features of the model-based clustering approach are worth discussing further. First, the approach provides an accurate result even when the clusters are heterogeneous in terms of number of time series in each cluster. Secondly, the clusters are allowed to be empty, a result which is used in the identification of the optimum number of clusters. When the optimum number of clusters is identified, any additional clusters are, in fact, supposed to be empty and when an empty cluster is added the change in observed log-likelihood is negligible. Finally, the result does not depend on the choice of parameters such as the number of knots or the spline order as in functional clustering.

6.3 Functional clustering

In order to cluster the simulated time series of paragraph 6.1 using the functional clustering approach, $K = 54$

equally spaced knots are defined over the temporal range $[1, 260]$ and cubic splines ($d = 3$) are considered. This provides approximately 1 knot every 4/5 weeks and this choice enables key features of the data to be captured while eliminating local variability.

The k-means algorithm is applied to the spline coefficient vectors $M = 10$ times in order to reduce the influence of the starting values.

K-means and complete-linkage hierarchical algorithms identify four clusters as optimal via the gap statistic and the L-curve (Fig. 3), with the curve classifications under both functional clustering approaches being the same. Note that the L-curve does not actually stabilize after a given number of clusters. Nonetheless, the last large step in the L-curve occurs when moving from three to four clusters. These four clusters (see Fig. 2) are comprised of three groups of curves being correctly classified as clusters where $n_j = 20, 40, 80$,

Fig. 2 Simulated data clustering result using the functional clustering approach (both k-means and complete-linkage hierarchical algorithms). Spline for each time series (*light line*) and cluster average (*dark line*)

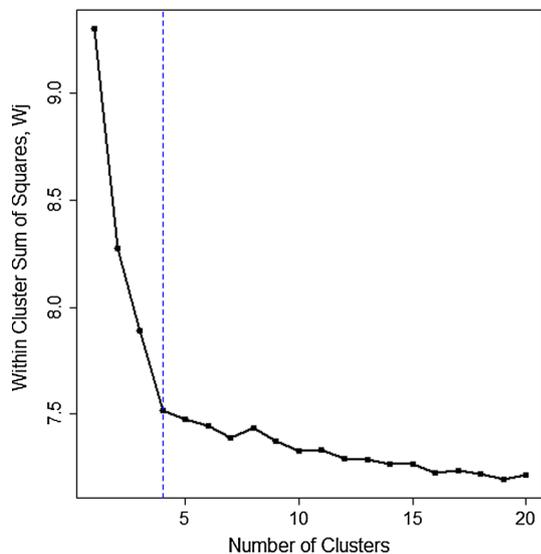
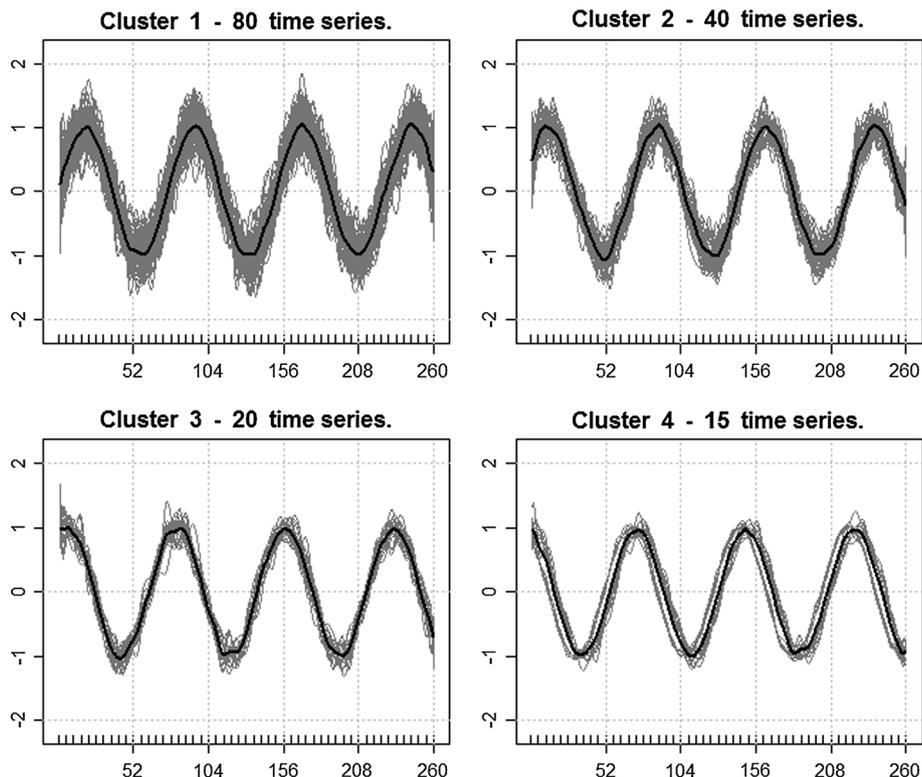


Fig. 3 L-curve related to the simulated data set

while the fourth is a combination of the simulated clusters with 5 and 10 curves. The fifth cluster is not identified using the functional clustering approaches. This is probably due to the fact that the fifth cluster only includes 5 curves and the L-curve does not seem to be very sensitive to detecting clusters with a small number of curves when differences between the clusters are small.

7 ARC-Lake data analysis

The ESA ARC-Lake project (<http://www.geos.ed.ac.uk/arclake/>) aims to exploit the scanning capability of the Along Track Scanning Radiometers (ATSRs) instrument on-board the Envisat satellite in order to derive observations of the lake surface water temperature (LSWT), for major lakes, globally, for the temporal period 1991–2010 in order to demonstrate the usefulness of these observations to climate science and to the study of climate change.

When the LSWT is analysed in order to study climate change, a fundamental aspect is to understand which lakes are temporally coherent with each other. If a global change is underway, it should be easier to detect the common change by analysing groups of temporally coherent lakes instead of all the lakes as a whole. In this section, therefore, the above developed clustering approaches are applied to the LSWT time series of the ARC-Lake data set in order to cluster the lakes into homogeneous groups with respect to their temporal coherence.

The data product ALIDxxxx_PLREC9D_TS366LM (see MacCallum and Merchant (2013)) includes the daily lake-average LSWT for 256 lakes around the globe and it is considered for data analysis.

The length of the time series represents a crucial aspect as the longer the time series the higher the probability that the time series differ at some instants in time. For this reason, the

Table 2 ARC-Lake data set clustering result using the model-based approach. Observed data log-likelihood and number of empty clusters

No. of clusters	2	3	4	5
Log-likelihood	14'437	22'478	25'644	30'663
# empty clust.	0	0	0	0
# clusters	6	7	8	9
Log-likelihood	32'897	34'045	36'101	38'330
# empty clust.	0	0	0	0
# clusters	10	11	12	13
Log-likelihood	39'568	40'928	40'925	40'938
# empty clust.	0	0	1	2

LSWT for the period 2006–2010 is considered as 5 years is a short period of time when compared to the dynamics of global change. The LSWT is averaged over seven days as there is a relatively small amount of variability at the daily level if compared to the long-term variability.

Since the lakes differ both in altitude above mean sea level and in volume, the time series of each lake is standardized to have zero mean and unit variance. This allows the removal of local effects not related to the global or regional climatology. Lakes from the same region but characterized by different altitudes, in fact, may have a different overall average LSWT, while lakes different in size may have a different

inertia and thus a different variability. Nonetheless, they should exhibit the same temporal pattern.

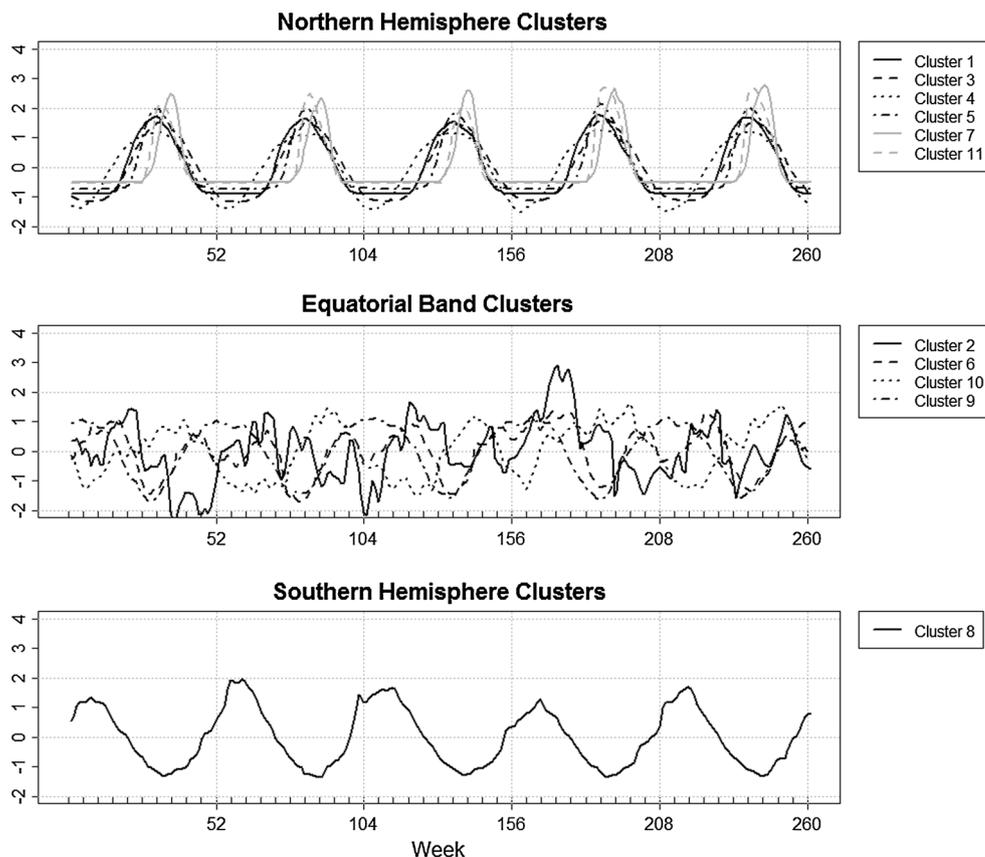
7.1 Model-based clustering

Model (1) is fitted with both \mathbf{G} and Σ_η constrained to be diagonal matrices. Model estimation is carried out using the D-STEM software (see Finazzi and Fassò (2014)) available at code.google.com/p/d-stem/.

Starting from $p = 2$, the model-based clustering technique is applied to the ARC-Lake data set and p is increased until an empty cluster is obtained. As in the simulation study, for each value of p the model is estimated 50 times and the estimation result related to the highest log-likelihood is retained. The average computing time for model estimation is around 90 seconds on a standard laptop machine. From Table 2 it can be noted that the log-likelihood stops substantially increasing between $p = 11$ and $p = 12$. In particular, the solution related to $p = 12$ is characterized by an empty cluster. Thus, $p = 11$ is considered as the optimum number of clusters. The number of time series in each cluster is reported in Table 3. Note that one cluster only includes one time series.

The clustering result displayed as estimated cluster averages is shown in Fig. 4. The average time series are given by $\{z_t^T\} = \{\mathbb{E}_\varphi(\mathbf{z}(t) | \mathbf{Y})\}$.

Fig. 4 ARC-Lake data set clustering result using the model-based approach - Estimated cluster averages divided by latitude bands



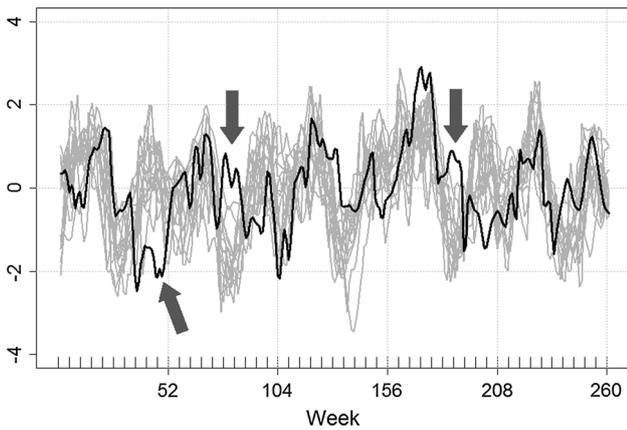


Fig. 5 ARC-Lake data set clustering result using the model-based approach - Temporal discrepancies (arrows) between cluster 2 and 6

Figure 5 shows the time series of the singleton cluster and cluster 6. Although the two clusters have many similarities, they are also characterized by differences that prevent them from being in the same cluster. The arrows in Fig. 5 identify the discrepancies between the singleton cluster and the time series of cluster 6.

The clustering result is represented on the map of Fig. 6 for Central and South America. This and subsequent

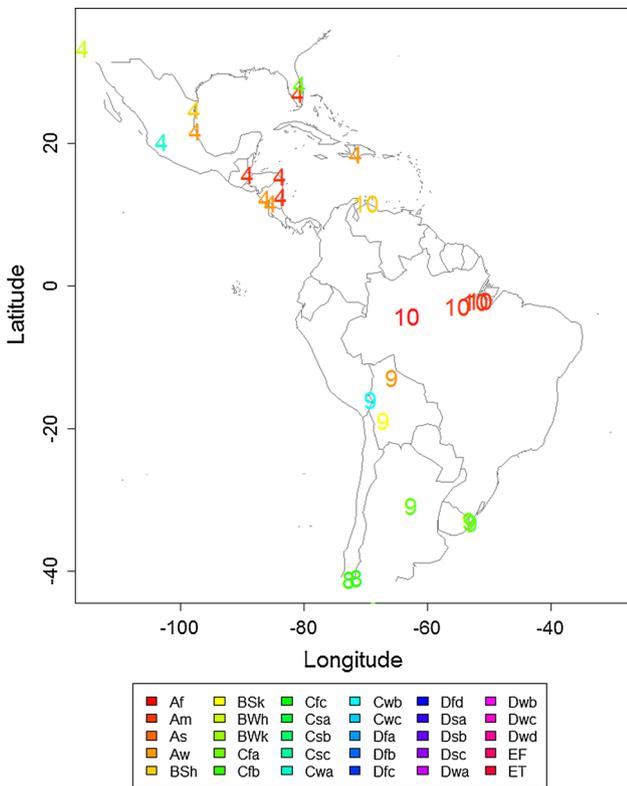


Fig. 6 ARC-Lake data set clustering result using the model-based approach - Spatial distribution of the clusters in Central and South America. Numbers represent the cluster membership while colours are related to the Köppen classification (see legend)

figures focus on a small area to facilitate the comparison across the clustering approaches. The reader may refer to the supplementary material for the global maps. The numbers displayed on the map describe the cluster membership of the lakes while the colour of the number is related to the Köppen climate classification (Peel et al. 2007). The Köppen classification, however, is based on both temperature and precipitation while the climate boundaries are defined by the local vegetation. The classification, thus, can give a hint on the spatial distribution of the clusters but the clusters are not expected to perfectly match the climate zones. For further information on the cluster classification codes see: <http://koeppen-geiger.vu-wien.ac.at/>.

The singleton cluster is related to the volcanic lake Toba in Sumatra, which surrounds the resurgent dome of the old volcano now known as Samosir island. Since the volcano is inactive or at least dormant, the LSWT discrepancies are probably due to local climatic conditions. Possibly, the large area of the resurgent dome may interfere with the remote sensing reading of the LSWT for this lake.

7.2 Functional clustering

As in the simulation study, time series are described using cubic splines considering $K = 54$ equidistant knots. K-means and complete-linkage hierarchical algorithms are subsequently applied.

For both the algorithms, the L-curves decrease smoothly and so are uninformative as to the optimal number of clusters. The gap statistic identifies 11 clusters as optimal for the k-means algorithm and 7 clusters when the complete-linkage hierarchical algorithm is applied. The number of curves in each cluster are included in Table 3 for both the algorithms.

A graphical sensitivity analysis was used to assess the influence of the number of knots/basis functions on the statistically optimal number of clusters identified by each method. The L-curve was computed for a broad range of potential numbers of basis functions. Within a reasonable range of the number of basis functions, the choice had little effect on the shape of the L-curve/gap statistic and hence the number of clusters chosen. At the more extreme values, when very few or many basis functions were used there was a difference in the number of clusters identified as optimal. The approach we decided on was to choose a number of basis functions whereby the key features of the data were captured by the curve fitted but local variation was not incorporated.

Figure(s) 7 and 8 represent the spatial distribution of the clusters for k-means and complete-linkage hierarchical clustering algorithms, respectively in Central and South America. The cluster averages curves are reported in Figs.

Table 3 Number of time series/curves in each cluster given by the three approaches

State-space	60	45	44	32	27	15	14	8	5	5	1
K-means	47	36	35	32	25	23	21	15	12	5	5
Complete-linkage	136	44	32	25	12	5	2				

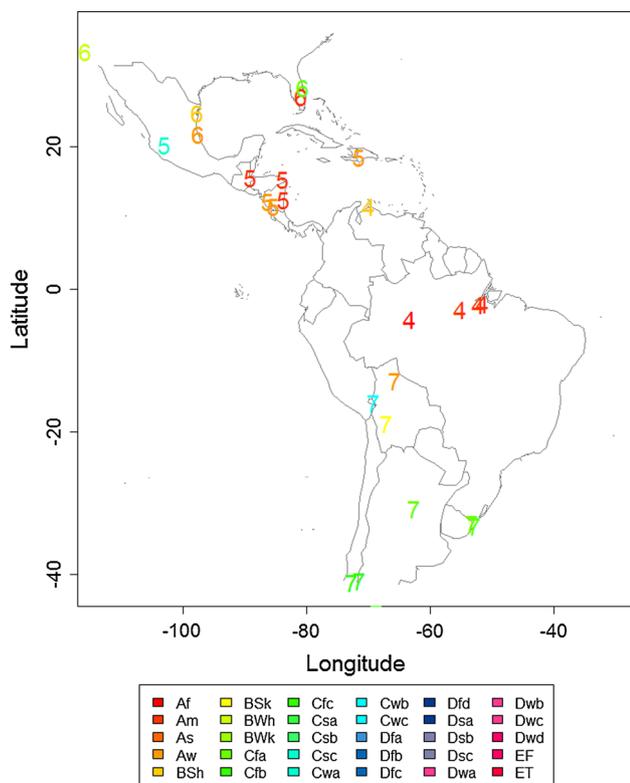


Fig. 7 ARC-Lake data set clustering result using the functional clustering approach and k-means algorithm - Spatial distribution of the clusters in Central and South America. Numbers represent the cluster membership while colours are related to the Köppen classification (see legend)

9 and 10 divided by latitude bands. In this area of Central and South America the number of clusters and the distribution are very similar. the main difference is that the k-means algorithm identifies a slightly larger number of clusters in Central America. Cluster numbers are simply labels and cannot be compared across the approaches.

7.3 Result comparison

In order to attempt to quantify how similar the clustering results are, the Adjusted Rand Index (ARI) is computed for all pairs of clustering algorithms. The ARI, which is developed in Hubert (1985), is a measure of agreement between two partitions which is corrected for the possibility that agreement between two sets of clusters may

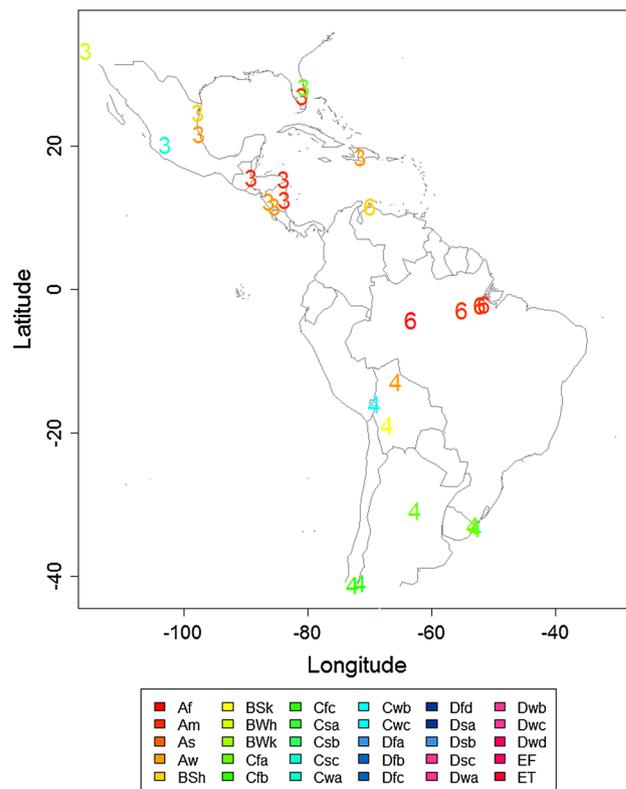


Fig. 8 ARC-Lake data set clustering result using the functional clustering approach and the complete-linkage hierarchical algorithm - Spatial distribution of the clusters in Central and South America. Numbers represent the cluster membership while colours are related to the Köppen classification (see legend)

simply be due to chance. It is an index which is based upon counting the pairs of curves on which two clusterings agree or disagree and is bounded at 1, corresponding to perfect agreement. A value of 0 indicates no agreement.

The ARI for the k-means algorithm and the model-based approach is 0.72, indicating a large degree of agreement between the partitions. For the complete-linkage hierarchical and the model-based approach the ARI is 0.48, which again indicates a general degree of agreement between the clusters identified. The ARI value for the two functional clustering algorithms is slightly lower at 0.38, however, this smaller value may be due to the different numbers of clusters, and the presence of a cluster containing two unusual curves which is identified using the complete-linkage hierarchical algorithm, but not by k-means.

Even if each algorithm provides a different clustering result, the temporal patterns they identify are similar. Figure 11 shows a comparison of the results with respect to two clusters which are apparently the same although they have different labels. The cluster averages are very similar with a propensity of the model-based approach to detect high frequency features of the temporal pattern.

Fig. 9 ARC-Lake data set clustering result using the functional clustering approach and the k-means algorithm - Estimated cluster averages divided by latitude bands

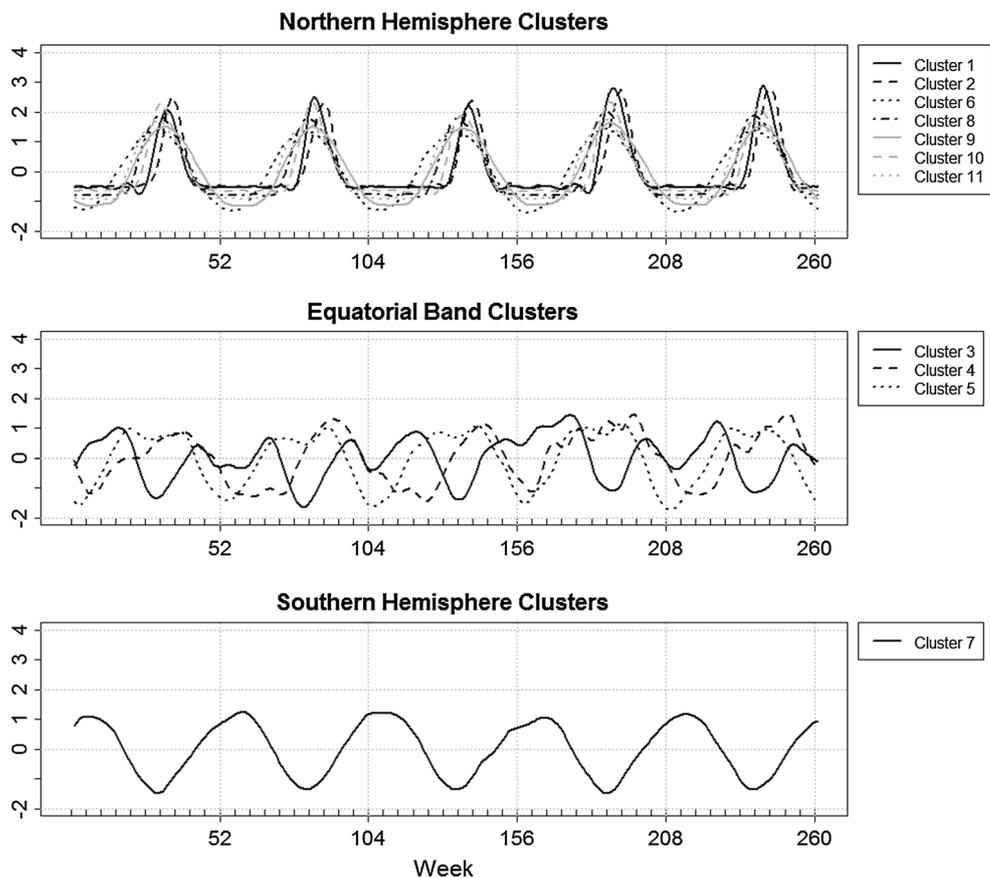
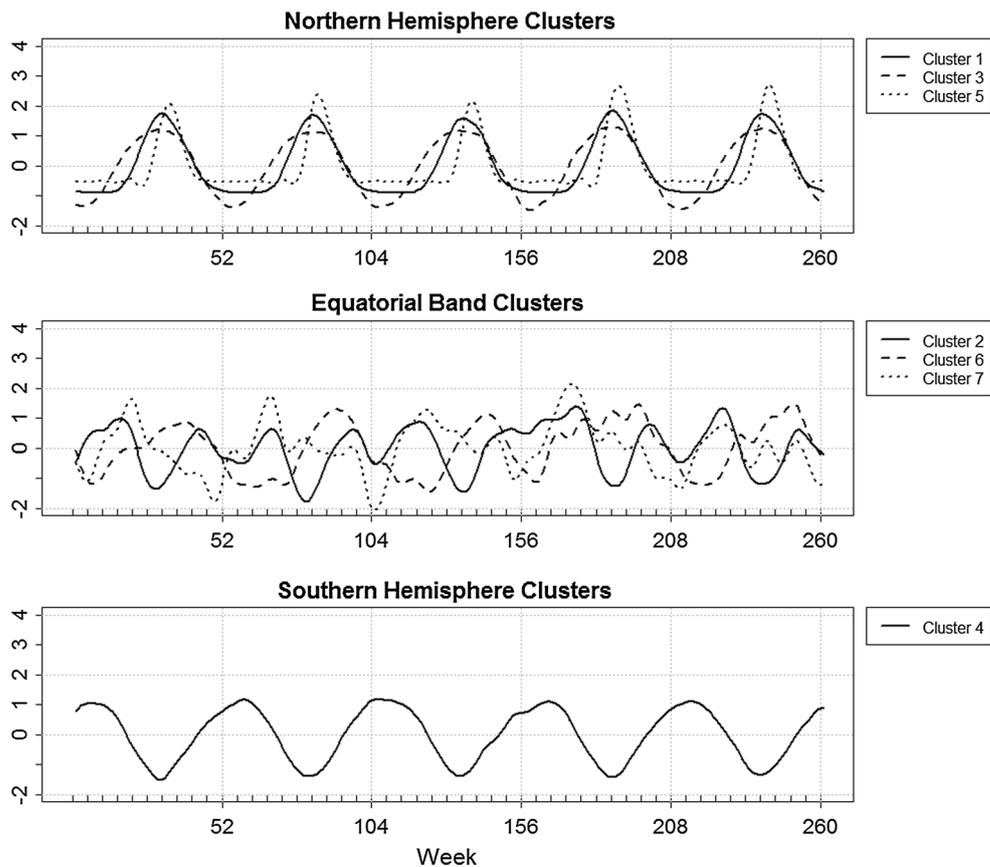


Fig. 10 ARC-Lake data set clustering result using the functional clustering approach and the complete-linkage hierarchical algorithm - Estimated cluster averages divided by latitude bands



This is due to the fact that the model-based approach does not involve any kind of smoothing of the original time series.

For the northern and the southern hemispheres it can be noted that the main difference between the cluster averages is the time of the peaks. In particular, the lower the distance to the Earth poles the later the peak in the LSWT. Lakes close to the Arctic Circle are characterized by periods of time during which the lake surface is frozen and the temperature is considered to be 0°C . This produces the flat “segments” which can be seen in the left images of Fig. 11. Although it is clearly an artefact produced by the measurement process, the model-based approach can reproduce the temporal pattern accurately. The basis functions used in the functional clustering approach, on the other hand, produce ripples where the time series should be flat. Both the clustering approaches highlight the difference between the Central Africa lakes and the Eastern Africa lakes. The time series related to the Central Africa lakes, in particular, present a double peak in winter due to the drier period in January and February which characterizes the Central Africa region.

8 Conclusions

The study of the temporal coherence of ecological time series is an important aspect of understanding the synchrony of major fluctuations in the attributes of interest and their relationships to common drivers and pressures. This is an extremely important issue in many fields, including weather and climate, made more challenging by the development of sensor networks and earth observation systems, which deliver very large data sets at high spatial and temporal frequencies. The statistical requirements in this context include models that are suitable for high dimensional noisy data with spatial and temporal correlations and software that is computationally efficient and able to handle large data sets. The new approach to state-space modelling proposed here which enables clustering, has been illustrated to successfully cluster both simulated and LSWT time series’ and to provide clustering results which are consistent with those given by functional clustering approaches. In terms of data processing, the model-based approach does not require the observed time series to be converted into curves and thus the clustering result is not

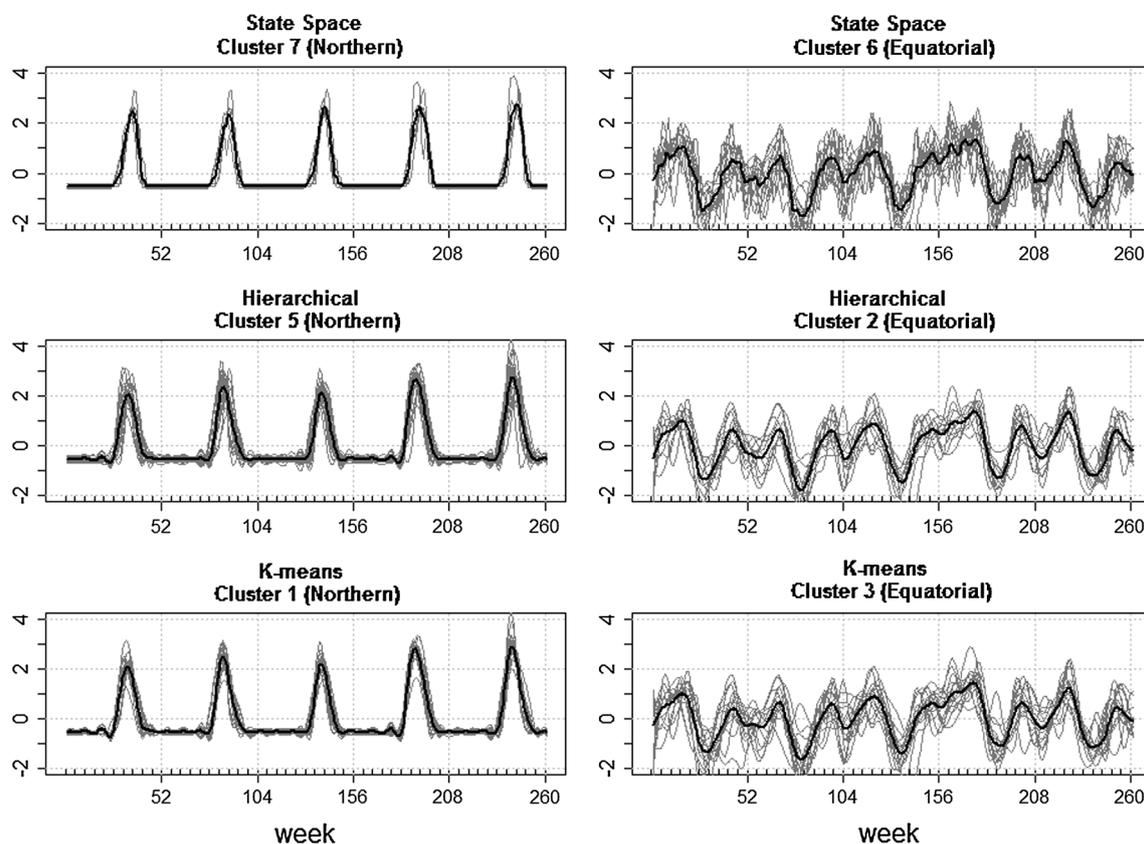


Fig. 11 Subset of the ARC-Lake data set clustering result using the model-based and the functional clustering approaches. Time series/splines (light line) and cluster average (dark line)

influenced by the choice of the spline order, the number of knots and their positions. On the other hand, smoothing can be useful when highly noisy time series are to be clustered, in which case the model-based approach might overestimate the number of clusters.

Spatial correlation can be introduced in order to avoid the proliferation of clusters when considering noisy time series. The simulation study developed in this work, however, has shown that both the clustering approaches are robust with respect to moderate levels of noise.

The approaches have been used on standardized time series as the main aim was to study their temporal coherence. If the interest is on the actual (non-standardized) time series, functional clustering can be applied straightforwardly while the model-based approach would require the introduction of additional model parameters.

The length of the time series is recognised to have an influence on the clustering result. Longer time series are expected to group into a larger number of clusters as the longer the time series the higher the probability they differ at some time point or time period. The choice of the time series length is strictly related to the aim of the analysis and to some features of the time series such as stationarity, seasonality and trends.

Future developments, driven by applications, will include a multivariable model and models which include covariates with differing spatial and temporal support and scale.

Acknowledgments Haggarty, Scott and Miller were partly funded for this work through the NERC GloboLakes project (NE/J022810/1). Finazzi was partially funded by the FIRB2012 project “Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions” (RBFR12URQJ). The authors gratefully acknowledge the ARC lake project for access to the data.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Abraham C, Cornillon PA, Matzner-Lber E, Molinari N (2003) Unsupervised curve clustering using b-splines. *Scand J Stat* 30(3):581–595
- Calder C (2007) Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. *Environ Ecol Stat* 14(3):229–247. doi:10.1007/s10651-007-0019-y
- Costa M, Goncalves A (2011) Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stoch Environ Res Risk Assess* 25(2):151–163. doi:10.1007/s00477-010-0429-5
- de Boor C (2001) A practical guide to splines. No. 27 in *Applied Mathematical Sciences*. Springer, New York

- Fassò A, Finazzi F (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics* 22(6):735–748. doi:10.1002/env.1123
- Finazzi F, Fassò A (2014) D-STEM - a Software for the Analysis and Mapping of Environmental Space-Time Variables. *J Stat Softw* (To appear)
- Franco-Villoria M, Scott E, Hoey T, Fischbacher-Smith D (2012) Temporal investigation of flow variability in scottish rivers using wavelet analysis. *J Environ Stat* 3(6). <http://eprints.gla.ac.uk/62946/>
- Grinsted A, Moore JC, Jevrejeva S (2004) Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes Geophys* 11(5/6):561–566. doi:10.5194/npg-11-561-2004
- Henderson B (2006) Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics* 17(1):65–80. doi:10.1002/env.750
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218
- Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. *Environmetrics* 19(7):672–686. doi:10.1002/env.946
- Labat D (2010) Cross wavelet analyses of annual continental freshwater discharge and selected climate indices. *J Hydrol* 385(1–4):269–278. doi:10.1016/j.jhydrol.2010.02.029
- Lansac-Tha F, Bini L, Velho L, Bonecker C, Takahashi E, Vieira L (2008) Temporal coherence of zooplankton abundance in a tropical reservoir. *Hydrobiologia* 614(1):387–399. doi:10.1007/s10750-008-9526-6
- Livingstone DM, Adrian R, Arvola L, Blenckner T, Dokulil MT, Hari RE, George G, Jankowski T, Jarvinen M, Jennings E, Noges P, Noges T, Straile D, Weyhenmeyer GA (2010) Regional and supra-regional coherence in limnological variables. In: G. George (ed) *The impact of climate change on European lakes*, no. 4 in *Aquatic Ecology Series*, Springer, pp. 311–337
- Lopes HF, Gamerman D, Salazar E (2011) Generalized spatial dynamic factor models. *Computat Stat Data Anal* 55(3):1319–1330. doi:10.1016/j.csda.2010.09.020
- MacCallum S, Merchant C (2013) Arc-lake v2.0, 1995–2011 [alidxxxx_p1rec9d_ts366lm]. University of Edinburgh, School of GeoSciences / European Space Agency, <http://hdl.handle.net/10283/88>
- Mardia KV, Goodall C, Redfern EJ, Alonso FJ (1998) The kriged kalman filter. *Test* 7(2):217–282
- Muoz-Carpena R, Ritter A, Li Y (2005) Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to everglades national park. *J Contam Hydrol* 80(1–2):49–70
- Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the kppen-geiger climate classification. *Hydrol Earth Syst Sci* 11(5): 1633–1644. doi:10.5194/hess-11-1633-2007. <http://www.hydrol-earth-syst-sci.net/11/1633/2007/>
- Salisbury J, Vandemark D, Campbell J, Hunt C, Wisser D, Reul N, Chapron B (2011) Spatial and temporal coherence between Amazon river discharge, salinity, and light absorption by colored organic carbon in western tropical atlantic surface waters. *J Geophys Res* 116(C7). doi:10.1029/2011JC006989
- Shumway R, Stoffer D (2006) *Time series analysis and ts applications, with R Examples*. Springer, New York
- Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J Royal Stat Soc* 63(2):411–423
- Zuur A, Ieno E, Smith G (2007) *Analysing Ecological Data*. Statistics for biology and health. Springer Science Business Media, LLC