Coherency in space of lake and river temperature and water quality records

Claire Miller, Marian Scott, Francesco Finazzi, Ruth Haggarty

Abstract Environmental time series observed over 100's of monitoring locations usually possess some spatial structure in terms of common patterns throughout time, commonly described as temporal coherence. This paper will apply, develop and compare two methods for clustering time series on the basis of their patterns over time. The first approach treats the time series as functional data and applies hierarchical clustering while the second uses a state-space model based clustering approach. Both methods are developed to incorporate spatial correlation and stopping criteria are investigated to identify an appropriate number of clusters. The methods are applied to Total Organic Carbon data from river sites across Scotland.

Key words: TOC, functional data, hierarchical clustering, state-space model, sensor networks

1 Introduction

Evaluation of water quality for regulatory purposes requires regular monitoring at multiple sites to assess changes in response to environmental forcing. However, many sites behave similarly throughout time and hence monitoring of a subset of sites with representative temporal patterns offers one possible solution to efficient

Claire Miller
University of Glasgow, Glasgow e-mail: claire.miller@glasgow.ac.uk

Marian Scott
University of Glasgow, Glasgow e-mail: marian.scott@glasgow.ac.uk

Francesco Finazzi
University of Bergamo, Italy e-mail: francesco.finazzi@unibg.it

Ruth Haggarty
University of Glasgow, Glasgow e-mail: ruth.haggarty@glasgow.ac.uk

and cost effective sampling. Formally, we say that sites that behave similarly over time possess temporal coherence. Understanding the spatial extent of temporal coherence for water quality parameters is a valuable tool to extrapolate from measured to unmeasured locations and to reduce monitoring effort. Spatially, monitoring may be undertaken at a river basin or catchment scale, at a national scale or even globally. As the temporal frequency and the spatial scale extends, then challenges in applying statistical models to high dimensional data in both space and time arise. For example, GloboLakes (Global Observatory of Lake Responses to Environmental Change, `www.globolakes.ac.uk`) is a five year research programme investigating the state of lakes and their response to climatic and other environmental drivers of change at a global scale. Data processing of archive satellite data will produce a 20-year time series, of observed ecological parameters and lake temperature. This will produce processed spatial images for over 1000 lakes at approximately fortnightly to monthly resolution. Novel and powerful statistical approaches are therefore required to assess temporal coherence within such data and to identify clusters of sites which illustrate temporal coherence.

Model-based functional clustering for river water quality data incorporating a stream based covariance structure is presented in [5], and [2, 3] apply model-based clustering using a state-space model to lake water temperature data to cluster lakes on the basis of both within and between lake variability respectively. This paper will compare and contrast two approaches to clustering such data on the basis of common patterns over time, one based on hierarchical clustering of functional data [5] and the other based on model-based clustering using a state-space model [1]. Both modelling approaches have been developed to incorporate spatial correlation. The comparison is illustrated using Total Organic Carbon data from 333 river monitoring sites across Scotland, supplied by the Scottish Environment Protection Agency. Monitoring the organic carbon levels in rivers is important, as rivers play an important role in transporting carbon from the land to the oceans, with constant feedbacks to and from the atmospheric carbon pool. The concentration of organic carbon in many Scottish rivers, has approximately doubled over the last twenty years, with soils being the most likely source [6]. Therefore, it is increasingly important to improve our understanding of the behaviour of organic carbon in aquatic systems.

2 Methodology

This paper compares two approaches for identifying spatial clusters of time series' that possess temporal coherence. The first approach is to represent each individual time series as a curve. These functional data are then clustered using hierarchical clustering. The second approach uses a state-space model to represent each time series in terms of latent variables. In both modelling approaches, spatial correlation has been incorporated based on euclidean distance between monitoring locations.

2.1 Hierarchical functional clustering (HFC)

Firstly, the curves corresponding to each monitoring location are estimated using p-splines. Secondly, the spatial covariance structure between locations is estimated and then finally a hierarchical clustering algorithm, which is modified to incorporate spatial distance between curves and correlation [7], is applied.

Since the curves have been estimated using splines, [5] illustrates that each curve can be expressed as the product of a set of spline coefficients ($\mathbf{c}_i$) and a vector of basis functions, $\Phi(t)$ as follows;

$$g_i(t) = \mathbf{c}_i^T \Phi(t),$$

where $i = 1,...,N$. Let the ith and jth curves, $g_i(t)$ and $g_j(t)$, be expressed as a linear combination of basis functions with coefficient vectors $\mathbf{c}_i$ and $\mathbf{c}_j$ respectively. The distance between the curves can then be written as

$$d_{ij} = (\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j), \tag{1}$$

where $W = \int \Phi(t)\Phi(t)^T dt$, which is a symmetric square matrix of order P, where P is the number of spline basis functions. For each set of basis functions, W can be evaluated using numerical integration, if necessary, and the functional distance matrix D with entries d_{ij} as defined above can be computed. Standard algorithms for hierarchical clustering can then be applied to the functional distance matrix.

Spatial covariance can be incorporated into hierarchical functional clustering by multiplying the functional distance matrix, defined in Equation 1, by a covariance matrix that has been estimated using the trace variogram [4]. Let $g_1(t),...,g_N(t)$, defined for $t \in [a,b] \subset \mathbf{R}$, be a set of curves which are realizations of a stationary, isotropic functional random process collected from N stations with corresponding location co-ordinates denoted by $x_1,\ldots,x_N$. Then, writing the distance between two locations i and j as h, the trace variogram can be defined as,

$$\gamma^*(h) = \frac{1}{2} E\left[(\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j)\right],$$

where $W = \int_{[a,b]} \Phi(t)\Phi(t)^T dt$. As with standard variograms, to obtain the empirical trace variogram, the trace variogram cloud can be computed by calculating the differences between all pairs of curves and plotting these differences against the corresponding distance between the locations. The points on this plot can then be 'binned' and averaged at a series of regular intervals. The estimated trace variogram can therefore be written as

$$\hat{\gamma}^*(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} (\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j),$$

where $|N(h)|$ is the number of curves separated by a distance of h units. After obtaining the empirical trace variogram from observed data, any standard variogram

model can be fitted as if it were a standard univariate variogram. The Matèrn function has been used here and parameters estimated based on the empirical variogram.

2.2 State-space model based clustering (SSMC)

Let $\mathbf{y}(t)$ be the $N \times 1$ observation vector at time $t = 1,...,T$, with N the number of monitoring locations and T the total number of time steps. In order to cluster the N time series with respect to their temporal coherence, the following state-space model is considered

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{K}\mathbf{z}(t) + \varepsilon(t) \\ \mathbf{z}(t) &= \mathbf{G}\mathbf{z}(t-1) + \eta(t), \end{aligned} \tag{2}$$

where $\mathbf{z}(t)$ is the $p \times 1$ latent state vector, $\mathbf{K}$ is an $N \times p$ matrix of coefficients, $\varepsilon(t) \sim N\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_N\right)$ is the $N \times 1$ measurement error vector, $\mathbf{G}$ is a $p \times p$ stable transition matrix and $\eta(t) \sim N(\mathbf{0}, \Sigma_\eta)$ is the $p \times 1$ innovation vector. Assuming $\mathbf{z}(0) \sim N(\mu_0, \Sigma_0)$ with Σ_0 a known variance-covariance matrix, the model parameter set is $\Psi = \left\{\mathbf{K}, \sigma_\varepsilon^2, \mathbf{G}, \Sigma_\eta, \mu_0\right\}$. In general, if restrictions are not imposed on Ψ, model (2) is not identifiable. Since the aim is to use model (2) for clustering, the coefficients of the matrix $\mathbf{K}$ are restricted to be only 0 or 1, with the additional constraint that each row $\mathbf{K}_i$ of $\mathbf{K}$, $i = 1,...,N$ must contain exactly one 1. These restrictions ensure model identifiability and allow the cluster membership to be determined directly from $\mathbf{K}$. In particular, the ith time series belongs to the kth cluster, $k = 1,...,p$, if the kth element of $\mathbf{K}_i$ is equal to 1. In order to incorporate spatial correlation, the elements of the matrix $\mathbf{K}$ are forced to be spatially correlated using the exponential function with parameter θ, where θ is the range of the spatial correlation which is assumed to be known. In particular, the generic element of the matrix $\mathbf{K}$ is given by

$$K_{ij} = \begin{cases} \left[\frac{1}{N}\sum_{i=1}^{N} w_i \text{corr}\left(y_i(t), \hat{z}_k(t)\right)\right]^m & \text{if } \frac{1}{N}\sum_{i=1}^{N} w_i \text{corr}\left(y_i(t), \hat{z}_k(t)\right) > 0 \\ 0 & \text{otherwise} \end{cases},$$

where $w_i = \exp\left(-\left\|\mathbf{s}_i - \mathbf{s}_j\right\| / \theta\right)$ and $\|\cdot\|$ is the distance between the spatial location $\mathbf{s}_i$ at which the time series $y_i(t)$ is observed and the spatial location $\mathbf{s}_j$ at which the time series $y_j(t)$ is observed while $\hat{z}_k(t)$ is the output of the Kalman smoother, $k = 1,...,p$.

Note that K_{ij} converges to either 0 or 1 when the EM iteration number m converges to infinity. In practice, apart from rounding error, convergence is obtained after a relatively small number of iterations ($m < 20$).

The model parameter set Ψ is estimated using a maximum likelihood approach by means of a modified version of the Expectation Maximization (EM) algorithm which is able to provide an estimated matrix $\hat{\mathbf{K}}$ subject to the above mentioned con-

straints. Since the EM algorithm is not guaranteed to converge to a global maximum of the likelihood function, the model is estimated M times randomizing the starting values of the model parameters each time. The parameter estimates corresponding to the highest observed-data likelihood are retained. In particular, the starting values for the elements of the matrix $\mathbf{K}$ are randomly generated from the uniform distribution $U(0,1)$ and they are normalized so that each row of $\mathbf{K}$ sums to one.

2.3 Selecting the optimal number of clusters

One of the main difficulties associated with cluster analysis is identifying how many clusters are most appropriate given the data. Stopping criteria are well developed for hierarchical clustering but less so for clustering based on state-space models.

For HFC, the L-curve and gap statistic [8] approaches have been used. Both the gap statistic and L-curve use the within cluster dispersion, W_k, to determine the number of clusters. For the L-curve approach a plot of W_k versus k is produced. The value of k at which W_k begins to flatten markedly indicates the number of clusters where there has been the largest increase in goodness of fit. 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. A number of reference data sets, say B, are calculated and HFC is then applied to each of them.

For the SSMC approach proposed here, model (2) is estimated starting from $k = 1$ (one cluster) and the number of clusters is incremented progressively until a stopping criterion is satisfied. In this work, the number of clusters is increased by one until an empty cluster is identified, that is, the matrix $\hat{\mathbf{K}}$ contains a column of zeros for cluster $k+1$. If the best solution (with respect to the smallest change in observe-data log-likelihood) includes an empty cluster, then the optimum number of clusters is given by k.

3 Applications

The above clustering methodology has been applied to Total Organic Carbon (TOC) data from several hundreds of monitoring locations across rivers in Scotland. This paper will illustrate the results for TOC from 333 monitoring locations on rivers across Scotland over 44 months, covering the period January 2007 - August 2010. Each of the 333 river time series have been standardised, individually, to have zero mean and unit variance.

The residuals, obtained after removing the fitted cluster means (from HFC and SSMC respectively) from each time series, display evidence of spatial correlation. For the HFC approach, spatial correlation was incorporated using the Matèrn function while an exponential function was used in the SSMC, both use an effective

range of 30km. Reducing the spatial correlation results in too many (similar) clusters being identified by the state space model and hence the spatial correlation acts as a penalty on the log-likelihood.

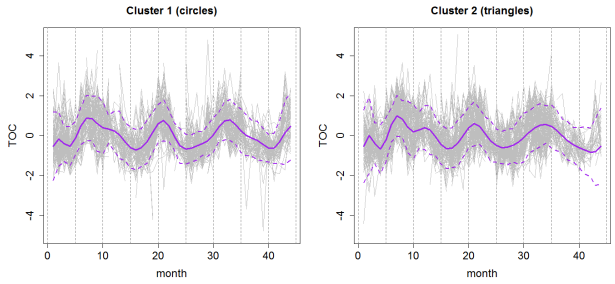

Fig. 1 Hierarchical functional clustering (HFC), mean cluster curves (thick line) with river time series for each cluster and $\pm$ 2 standard errors (dashed lines), gap statistic selects 2 clusters with spatial correlation incorporated using Matèrn weights and effective range of 30km

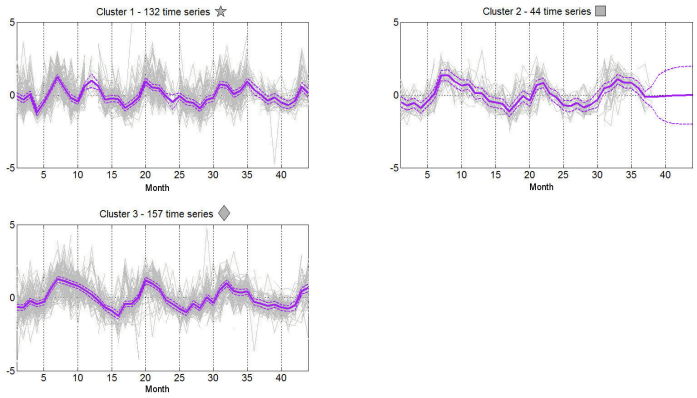

Fig. 2 Model-based clustering (SSMC), mean cluster curves (thick line) with river time series for each cluster and $\pm$ 2 standard errors (dashed lines), 3 clusters are selected with spatial correlation incorporated using exponential weights and range of 30km

Figures 1 and 2 display the cluster mean curves for the optimal number of clusters selected using each approach with the locations of clustered sites displayed on Figure 3. There is no apparent trend in the original series over this short period of time and hence the clustering is based on short term fluctuations, including seasonal patterns. It appears that one fewer cluster is required when the data are smoothed initially and treated as functional data (HFC) than when the raw data are used with the state-space model (SSMC). The figures also illustrate the larger standard errors

for the HFC approach and the missing data at the end of the time period for the time series in Figure 2, cluster 2.

In Figure 1 the main differences in the temporal pattern are the more pronounced 'bump' in cluster 2 around 10 months and a more pronounced peak in cluster 1 after 30 months. In Figure 2 these features are distributed amongst three clusters, with cluster 1 displaying the additional 'bump' around 10 months, cluster 2 & 3 are fairly smooth around this time point. However, cluster 2 contains a more pronounced peak after 30 months. Cluster 1 (HFC) and clusters 2 and 3 (SSMC) show a strong declining feature from months 7 to 15.

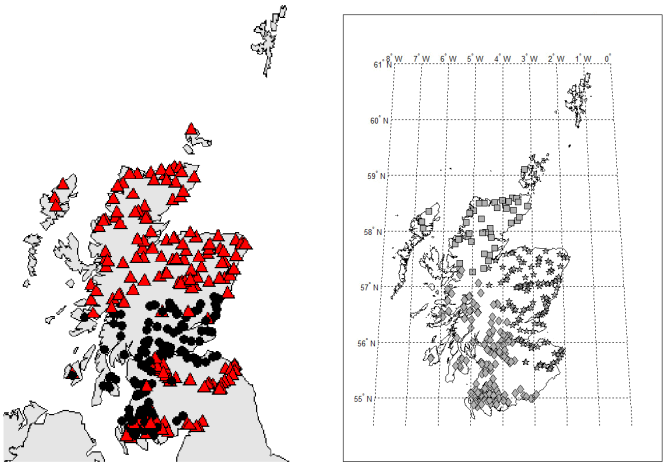

Fig. 3 Map of Scotland displaying clusters from hierarchical functional clustering (HFC) (left) and map of Scotland displaying clusters from model-based clustering (SSMC) (right).

The spatial location of each cluster is displayed in Figure 3 and while the cluster means are very similar, within each clustering approach, the distributed spatial patterns appear quite distinct. The predominant feature is the North/South separation, common in both methods, but the SSMC clusters also show an east/west split. Cluster 2 from both methods covers the sites in the far north of Scotland with those in central and low land Scotland split between clusters 1 and 3 for SSMC. Several explanations have been put forward to describe the temporal patterns in organic carbon in aquatic systems including changes in soil chemistry (decreasing mineral acidity), rising temperatures, deforestation and disturbance, and changes in hydrology, including diffuse pollutants. TOC will have a seasonal pattern, (generally lower in spring, higher in autumn) but this may also vary from site to site. The environmental covariates will also vary spatially, with predominantly peat in the far north, different rainfall patterns between east and west and more local variation in disturbances and diffuse pollutants. Therefore, the features in the cluster means highlight small but potentially ecologically important variations.

4 Discussion

The clustering approaches proposed enable a large number of time series to be clustered on the basis of patterns over time. HFC, being based on smoothed data, detects less local variation and has well-established stopping criteria. Using HFC, the computing time for the application in Section 3 is less than two minutes using a standard laptop. However, the stopping criteria are computationally demanding with the gap statistic taking in the order of hours. SSMC provides similar results, but identifies more local variation. It provides a computationally efficient algorithm based on the Kalman smoother, characterized by a computing time which is linear with respect to the number of time series, the number of time steps and the number of clusters. For instance, the computing time for the application of Section 3 is less than one minute on a standard laptop. Stopping criteria, however, are less well-developed and selection based on negligible difference in log-likelihoods and empty clusters is proposed here.

Acknowledgements Miller and Scott were partly funded for this work through the NERC GloboLakes project (NE/J022810/1). SEPA (Janet Moxlet and Mark Hallard) are thanked for the provision of the data.

References

1. Finazzi, F., Fassò, A.: D-STEM - A statistical software for multivariate space-time environmental data modeling. Proceedings of the International Workshop on Spatio-Temporal Modelling (METMA VI) (2012)
2. Finazzi, F., Miller, C., Scott, M.: A model-based clustering approach for the analysis of environmental time series. Proceedings of the 28th International Workshop on Statistical Modelling (2013)
3. Finazzi, F., Scott, M., Miller, C., Fassò, A.: A model-based clustering approach for the study of the temporal coherence of multivariate time series. Submitted to Annals of Applied Statistics (2013)
4. Giraldo, R., Delicado, P., Mateu, J.: Ordinary kriging for function-valued spatial data. Environmental and Ecological Statistics **18**, 411–426 (2011)
5. Haggarty, R.A., Miller, C.A., Scott, E.M.: Spatially Weighted Functional Clustering of River Network Data. Submitted to Journal of the Royal Statistical Society Series C (Applied Statistics) (2013)
6. Moxley. J.: Trends in organic carbon in Scottish rivers and lochs. Scottish Environment Protection Agency (2011)
7. Romano, E., Giraldo, R., Mateu, J.: Clustering Spatially Correlated Functional Data. Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics, F. Ferraty (ed), Physica-Verlag HD (2011)
8. Tibshirani, R., Walther, G., Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistic, Journal of the Royal Statistical Society. Series B (Statistical Methodology), **63**(2), 411-423 (2001)