Identifying homogenous regions for Chlorophyll in Lake Balaton

Lakes are considered as sensitive indicators of environmental change which are impacted by both natural and anthropogenic drivers. The potential impact of climate change on freshwater resources is critical, and improved understanding of the observed changes is key to ensure better management of aquatic resources. Large lakes may have several basins or areas within them that behave differently in terms of water quality indicators and hence these distinct areas may respond differently to drivers. It is therefore of interest to investigate and identify coherent regions within lakes which are similar in terms of trends, seasonal patterns and levels of determinands present.

In this context, we imagine a set of time series, each series corresponding to a pixel, observed over time. Coherence can then be defined as the synchrony between major fluctuations in a set of time series. Figure 1 shows an example set of time series curves where different colours indicate coherent sets of curves. Coherence in Figure 1 is primarily determined by the timing and amplitude of the seasonal pattern in the curves. In order to identify regions within lakes which are temporally coherent, we have taken a clustering approach to identify the statistically optimal number of clusters which we then map in space.

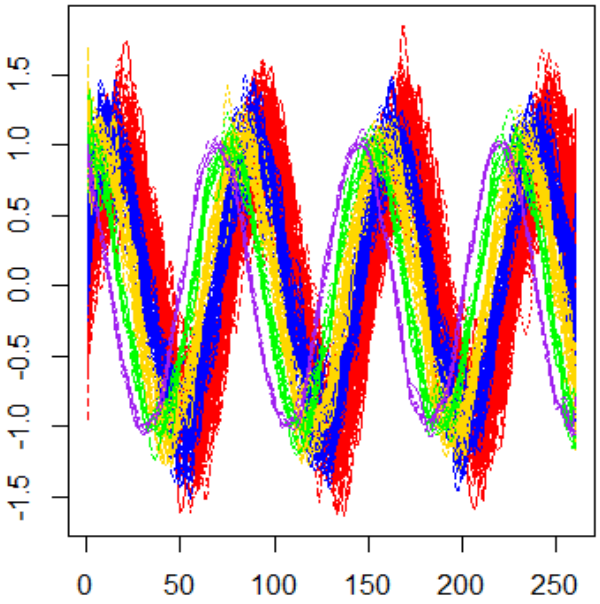

Figure 1: Simulated time series data where different colours represent temporally coherent clusters of curves.

The aim of this case study is to investigate statistical methodology that can be used to define coherent regions within lakes. An illustrative example will be presented which considers chlorophyll a at Lake Balaton in Hungary.

Case Study: Lake Balaton

Lake Balaton is situated in Hungary and is the largest freshwater lake in Central Europe. It is situated 104.8m above sea level with a surface area of 592km^2 and an average depth of 3.2m. For the case study presented, Earth Observation (EO) data for Chlorophyll a (mg/m^3, MPH product from the Diversity II project) will be used. Around 10 years of monthly observations were available, covering the time period from June 2002 to April 2012. EO data are available for 8219 pixels covering the lake surface at a resolution of approximately 300m. A two pixel border around the boundary of the lake has been removed due to the high likelihood of edge effects. Removing this boundary resulted in 6064 pixels being used in further analysis.

Methodology

A functional data analysis approach (Ramsay and Silverman, 1997) has been taken where the time series for each pixel is represented as a curve over time. The curve then becomes the observation of interest rather than the individual data values in subsequent analysis and is estimated via penalised regression spline smoothing (Eilers and Marx, 1996). Using such an approach provides a smooth estimate of Chlorophyll a for an individual pixel as a function of time, removing local variability. These smooth pixel curves then become the 'functional data'.

One approach to find clusters based on functional data is to first decompose the variation in the data by applying a functional principal component analysis (FPCA) and subsequently, to cluster the corresponding principal component scores. The application of FPCA enables the dimension of the functional data to be substantially reduced and hence provides a very computationally efficient way of exploring any underlying structure in the data.

FPCA can be used to identify the dominant modes of variation in a data set. In the functional case, both the data and the estimated functional principal components (FPCs) are curves. The FPCs can be thought of as a set of orthogonal basis functions constructed so as to account for as much variation as possible. Full details of functional principal components are provided in Ramsay and Silverman (1997).

Following the estimation of functional principal component scores, standard statistical clustering approaches can be applied such as k-means, hierarchical and model based clustering (Fraley and Raftery, 1998). The principal component scores have been weighted firstly to account for the proportion of variability each component explains and then also to account for spatial variability prior to the application of clustering procedures.

To obtain the weights which were used to account for spatial variability, a global Moran's I (Anselin, 1995) is calculated. This is a measure of the spatial homogeneity in the lake. If there is little evidence of any long term trend in the spatial pattern over time, this measure of spatial correlation can be based on the temporal mean lake surface. Moran's I is computed using correlations between each pixel and its neighbouring pixels to provide a measure of spatial correlation that summarises how similar neighbouring pixels are to one another. This value can be incorporated into the clustering by weighting the set of FPC scores by an appropriate neighbourhood matrix.

In order to choose the statistically optimal number of clusters there are well developed data-driven methods such as the L-curve, gap statistic (Tibshirani et al., 2001) and Dunn Index (Dunn, 1974) which can be utilised. Each of these methods was explored for Lake Balaton.

Application of Clustering at Lake Balaton

Figure 2 shows the spatial mean surface for Chlorophyll a (mg/m^3) at Lake Balaton over the time period. The two pixel boundary has been removed in this image. As can be seen, there is broad variation across the surface of the lake, with much lower values of Chl a in the upper basin of the lake. From these data Moran's I was calculated as 0.89. This relatively high value reinforces that there is a strong spatial correlation structure across the surface of the lake in terms of the temporal mean values for each pixel.

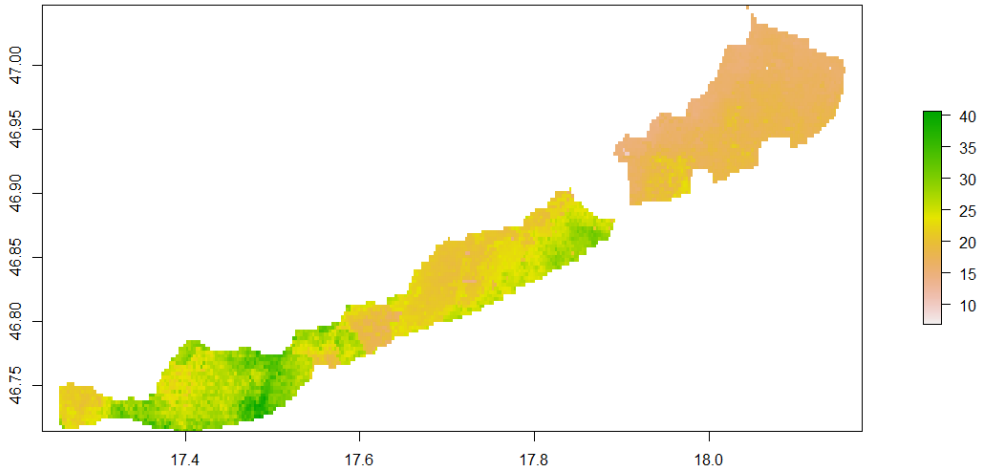

Figure 2: Map showing temporal mean Chlorophyll a (mg/m3, MPH product) over surface of Lake Balaton

Smooth functions were fitted to the curves corresponding to each pixel using a b-spline basis. There is some question as to how to select the flexibility of the smooth curve as clustering results can be sensitive to the degree of smoothing applied. Too much smoothing may mean important features in the data are missed, while too little smoothing will result in an estimated function that follows the observed data closely and has high variation in local areas.

Preliminary results are presented here for Lake Balaton where the flexibility of smoothing was selected by allowing one degree of freedom per season (3 months). This quantity of smoothing was selected as it enabled the estimated smooth curves to capture the key features of the time series without retaining excessive local variability. After estimating the smooth function for all pixels FPCA was applied to the full set of curves. The first two principal components were found to account for 90% of the variability in the data and so clustering was applied to the first two principal component scores only.

Figures 3 and 4 show clustering results after applying model based clustering. The L-curve and Dunn Index indicated that 4 clusters was the statistically optimal number for describing variability in the pixel curves. Figure 3 displays the spatial distribution of the clusters, with each cluster represented by a different colour, while Figure 4 shows the uncertainty in cluster classification associated with each pixel. In general, the pixel curves are well separated into the clusters, with increased uncertainty only at the boundaries of the clusters. The clustering has identified 4 distinct regions within the lakes.

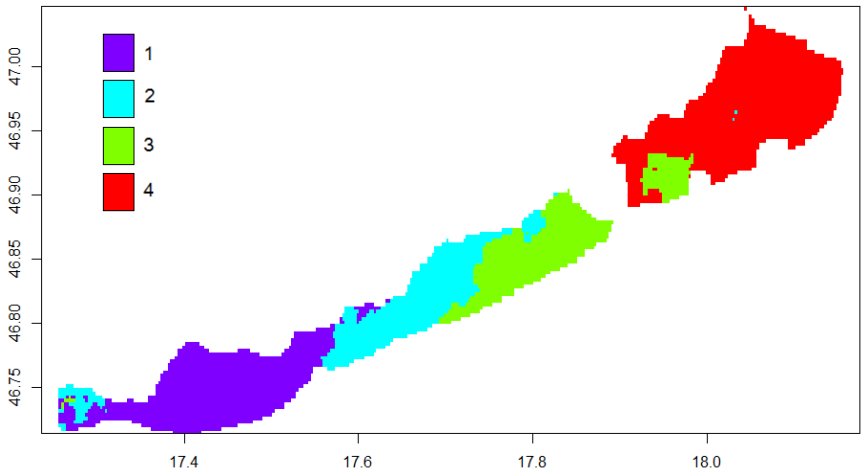

Figure 3 Map of Lake Balaton showing results of model based clustering for Chlorophyll a (mg/m3, MPH product). Cluster assignment for each pixel indicated by different colours.

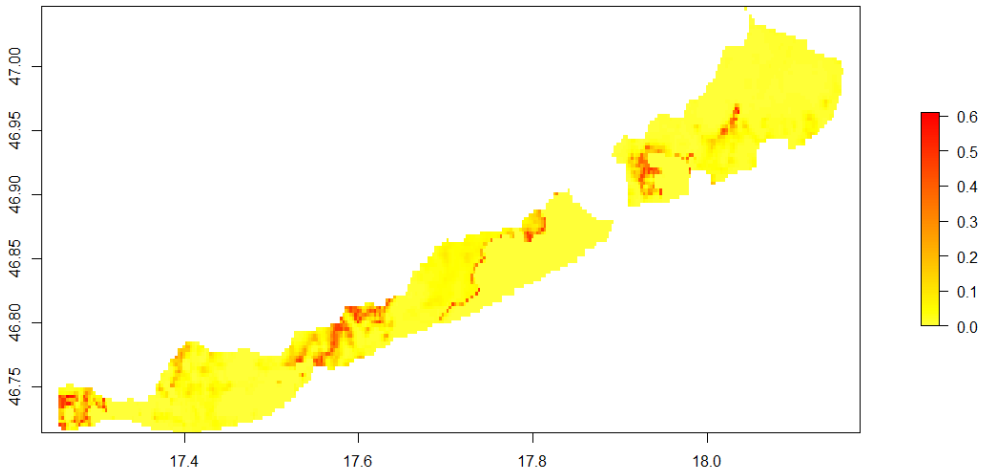

Figure 4 Map of Lake Balaton with cluster membership uncertainty for each pixel

It is reassuring to see the spatial homogeneity amongst the clusters, with only a small proportion of pixels appearing 'inconsistent' with their neighbours. These pixels are identified as having the greatest values of uncertainty associated with them.

Figure 5 shows the temporal mean Chlorophyll a curves for each cluster. Colours correspond to those in Figure 3. The most apparent feature is the peak at the start of 2003 for Cluster 1 (purple). As this dominates the scale of the mean curves, Figure 6 shows the same results but from mid 2003 onwards. Beyond 2003 it is clear that Cluster 3 displays most variability in terms of Chlorophyll and is generally higher than the other clusters. Cluster 4, shown in red and located in the top section of Lake Balaton tends to have lower values, particularly from 2008 onwards, where there is minimal variability in the Chlorophyll levels.

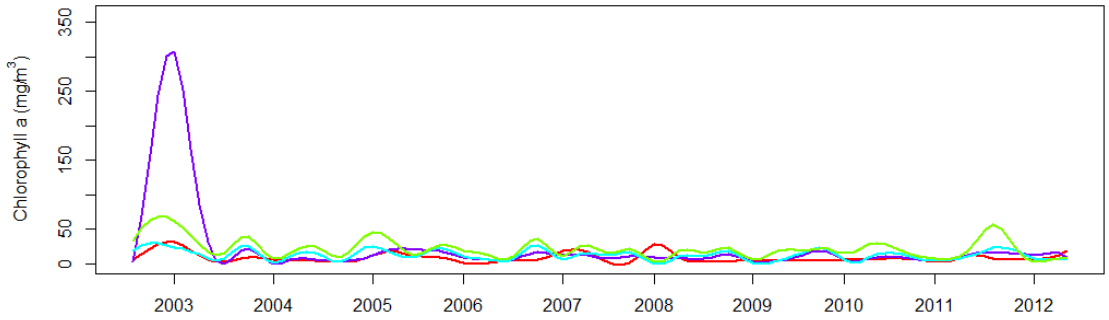

Figure 5 Temporal cluster mean curves for Chlorophyll a at Lake Balaton. Colours correspond to those shown in Figure 3.

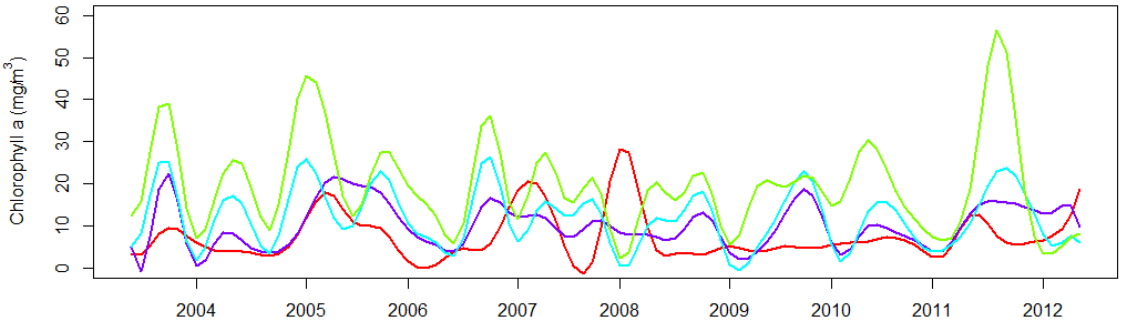

Figure 6 Temporal cluster mean curves for Chlorophyll a at Lake Balaton from 2003 onwards. Colours correspond to those shown in Figure 3.

Summary

It has been demonstrated that clustering approaches can be applied to EO data in order to identify coherent regions within lakes. Data driven approaches can be used to let the patterns in the data determine the number of clusters which is statistically optimal.

In contrast to standard clustering approaches, treating the pixel curves as functional data enables not only the mean value to inform the cluster membership, but also ensures temporal patterns such as trends and seasons are incorporated. Dimension reduction via spline smoothing and functional principal component analysis ensures clustering methods are computationally efficient.

References

Anselin, L., (1995), Local Indicators of Spatial Association—LISA, Geographical Analysis, 27(2), 9-1157.

Dunn, J. C., (1974), Well Separated Clusters and Fuzzy Partitions. Journal on Cybernetics, 4, 95-104.

Eilers, P. H. C. and Marx, B. D. (1996), Flexible smoothing with b-splines and penalties, Statistical Science 11(2), 89 -121.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41(8), 578–588.

Ramsay, J. and B. W. Silverman (1997). Functional Data Analysis (Springer Series in Statistics) (1st Ed.). Springer.

Tibshirani, R., G. Walther, and T. Hastie (2001). 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), pp. 411–423.