Using co-clustering to analyze spatio-temporal patterns: a case study based on spring phenology

R. Zurita-Milla, X. Wu, M.J. Kraak

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente,
PO Box 217, 7500 AE Enschede, The Netherlands
Telephone: +31 53 4874367
Email: r.zurita-milla@utwente.nl

Abstract

Clustering is a widespread method to explore patterns in large spatio-temporal datasets. Most clustering studies are, however, performed either from a spatial or from a temporal point of view. This is sub-optimal because patterns explored from a spatial perspective cannot describe the time-varying behavior present in the dataset and vice versa. Here we illustrate a co-clustering-based analysis that enables the simultaneous analysis of spatial and temporal patterns. In particular we show that by combining the Bregman block average co-clustering with I-divergence and the k-means algorithms we can extract the main patterns of leaf onset over Europe, northern Africa, Turkey and the Middle East. Our results indicate that four main spatial patterns exist in the period under study (1950 to 2011). These patterns were visualized using maps and a timeline was used to indicate the years of occurrence of these patterns.

Keywords: Spatio-temporal analytics, Bregman block average co-clustering, information divergence, spring indices, phenoregions.

1. Introduction

Clustering is a widespread method to analyze patterns in large spatio-temporal dataset. It detects groups of similar data and empowers analysts, as clustering helps to understand the data at a higher level of abstraction (Andrienko et al., 2009). Until now, most clustering studies are performed either from a spatial or from a temporal point of view. Take time evolving values for a variable measured at several locations as an example. This data is typically arranged in a matrix where each row represents a geographic location and each column contains the value of the variable at each timestamp. From a spatial point of view, locations can be regarded as objects and each timestamp as an attribute in this matrix and clustering this data will identify location-clusters with similar values along all timestamps. Alternatively from a temporal point of view, one can (virtually) transpose the matrix and perform a temporal clustering. Now each timestamp can be regarded as an object and each location as an attribute and clustering will provide the timestamp-clusters with similar values along all locations. However, both types of clustering are sub-optimal in the sense that patterns explored by only using spatial clustering cannot describe the time-varying behavior present in the dataset and vice versa (Deng et al., 2013). Co-clustering enables the simultaneous analysis of spatial and temporal patterns.

The rest of the abstract is organized as follows: section 2 contains a brief description of co-clustering and more details on the specific co-clustering algorithm used in this study; section 3, demonstrates the use of co-clustering for a real world application: the study of long-term spring phenological patterns over Europe; finally, section 4 contains the conclusions of our work.

2. Co-clustering

Co-clustering regards locations and timestamps equally (Han et al., 2012). This means that similar locations are mapped to location-clusters and similar timestamps to timestamp-clusters at the same time. Co-clusters are located at the intersection of location and timestamp clusters. Co-clustering results are thus a series of non-overlapping subsets of rows and columns of the original data matrix that contain similar data.

In 2007, Banerjee and colleagues developed a generic co-clustering algorithm called Bregman co-clustering. In the same publication, they also proved the superiority of an information theory-based metric called the I-divergence for co-clustering complex datasets. Wu et al. (2015) recently applied the Bregman block average co-clustering algorithm with I-divergence (BBAC_I), a special case of the previously mentioned algorithm that preserves the co-cluster averages, to Dutch temperature data and successfully identified co-clusters containing similar temperatures along both the spatial and the temporal dimensions. Therefore, we adopted the BBAC_I algorithm for this paper too. This algorithm can be used to cluster positive data matrices with real-valued elements, which represent co-occurrences or joint probability between two random variables. It treats the co-clustering as an optimization problem in information theory. In the solution to this problem minimizes the loss of mutual information between the original and the co-clustered data matrix. Briefly, the algorithm starts with an initial random mapping from locations to location-clusters and years to year-clusters. This allows the calculation of the co-clustered matrix. Then the loss in mutual information is calculated as the I-divergence between the original and the co-clustered matrices. After that, the algorithm starts an iterative process to update the mapping from locations to location-clusters and years to year-clusters to minimize the loss function (see Banerjee et al, 2007 and Wu et al., 2015 for more details and the pseudo-code of the BBAC_I algorithm).

3. Case study

Phenology is the science that studies the timing of recurrent life cycle events in plants and animals, their inter-relations and the impact of environmental factors on them (Lieth, 1974). Phenological studies are an excellent proxy to study climate change as the timing of many life cycle events (e.g. appearance of first leave or flower in a plant) is strongly influenced by environmental factors. Thus, the identification of phenoregions and their changes in time provide value information on the impacts of climate change on the area of interest. Here we use the extended spring index (SI-x; Schwartz et al., 2013) models to characterize spring onset over Europe. More precisely, we predicted the phenological event “first leaf” from daily maximum and minimum E-OBS temperature records. These datasets cover the period 1950 to 2011 and have a spatial resolution of 0.25 degrees. Then we applied the BBAC_I algorithm to identify co-clusters that contain similar first leaf dates (FLD) along the spatial and temporal dimensions. We asked the algorithm to identify 45 spatial groups and 4 temporal groups (Figure 1). We used this relatively large number of clusters because we wanted to have a fine “segmentation” of the data that would allow a regrouping into more meaningful co-clusters. Such a re-grouping is needed because the BBAC_I co-clustering algorithm assigns full rows (spatial units) / columns (year) to the co-clusters whereas rows and columns can contain some heterogeneity. In this study we use k-means to refine the 45 by 4 co-clusters identified by the BBAC_I algorithm. The value of k was optimized using the Silhouette method, which identified five FLD groups that were named “very late”, “late”, “early”, “very early” and “abnormal”. After that, the re-grouped FLD data matrix was projected back to the geographic space (Figure 2; left) to generate four distinct FLD spatial patterns. These results show that the first years of the period under study had very late FLD (i.e. cold springs, especially in northern Russia, Scandinavian countries, Iceland and few areas of Western Europe where the Alps are located). Recent years, with the exception of 1996, display early FLD (figure 2; right). This is, recent years exhibit warm springs, particularly in most of the Iberian Peninsula, northern France and Ireland. Results also show that the E-OBS temperature dataset might have some quality issues at selected locations/years. For instance, southern Iceland has abnormal FLD values for the 62 years covered by the temperature datasets.

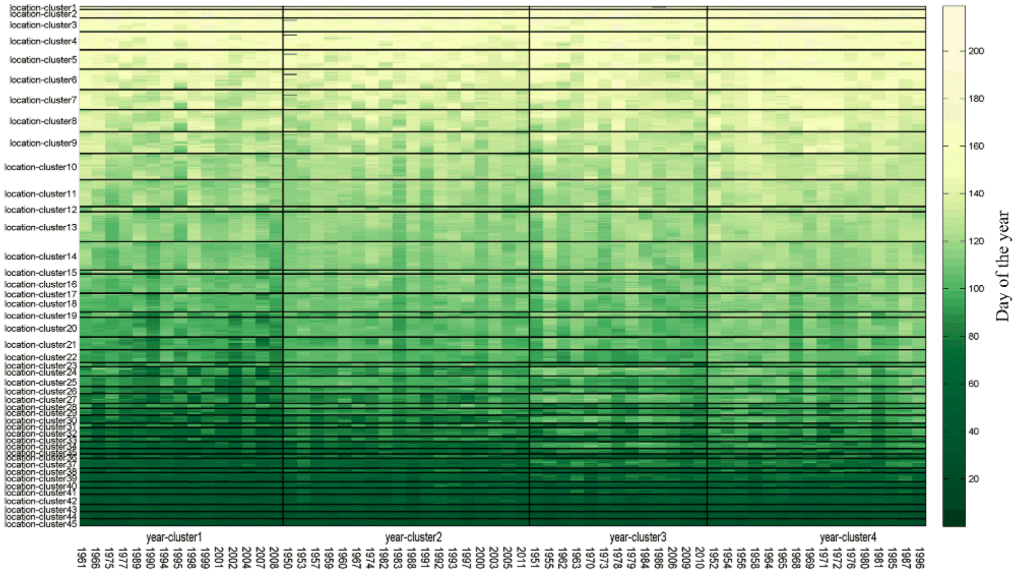

Figure 1. The output of the co-clustering: a reordered FLD matrix with 45 by 4 FLD co-clusters. The X-axis shows the 62 years of the dataset arranged from year-cluster1 to year-cluster4. The Y-axis shows all the grid cells from location-cluster45 to location-cluster1. The greener the color, the earlier the FLD value.

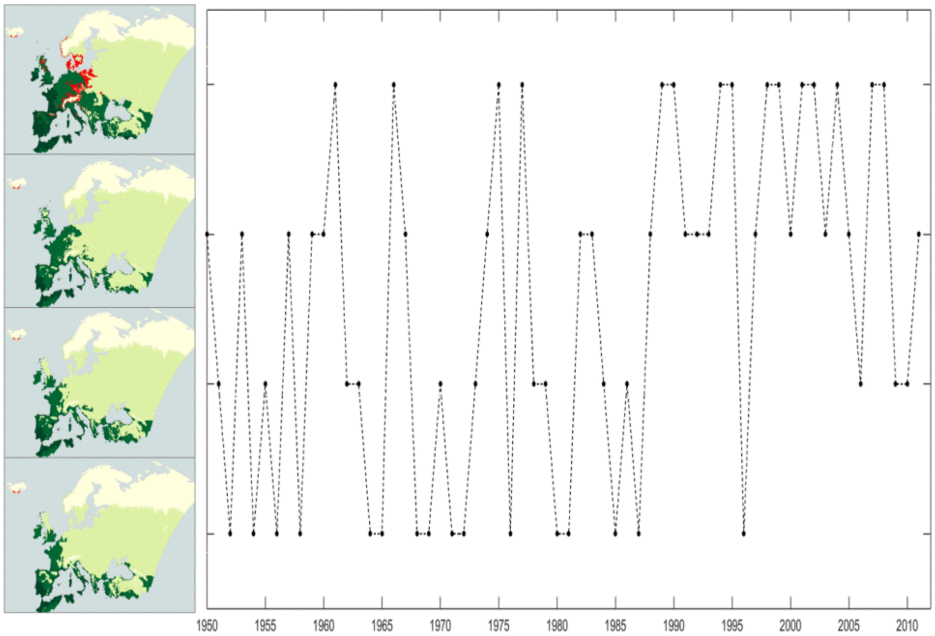

Figure 2. Left panels: the four FLD spatial patterns identified in this study. The greener the color the earlier the spring; red cells indicate cells with “abnormal” values. Right plot: a timeline with the temporal variability of the spatial patterns in the period 1950 to 2011.

4. Conclusions

In this abstract, we have presented a novel analytical approach to extract spatio-temporal patterns. The approach is based on the Bregman block average co-clustering algorithm with I-divergence (BBAC_I), which regards space (locations) and time (timestamps) as equally important dimensions. The BBAC_I algorithm optimizes the composition of the co-clusters by minimizing the loss of mutual information between the original data matrix and the co-cluster one. This optimization is done by moving full rows (space) and full columns (time), neglecting the heterogeneity of the rows/columns. Thus, here we used the BBAC_I to produce many co-clusters that were later regrouped into an optimal number of clusters using k-means and the Silhouette method,

The proposed approach was illustrated with a spring phenological dataset: 62 years of modelled first leaf dates (FLD) over Europe, northern Africa, Turkey and the Middle East. Our results indicate that co-clustering, coupled with k-means, can efficiently capture complex spatio-temporal patterns in large datasets. Five FLD groups arranged in four fundamental FLD spatial patterns were discovered in the dataset. This co-clustering based approach also allowed us to the study the temporal dynamics of the four spatial patterns.

5. Acknowledgements

We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu). We also thank T. R. Ault, M.D. Schwartz and J.L. Betancourt for the lively discussion on the Spring index models and David G Rossiter for his insightful feedbacks on earlier vesions of this work.

6. References

Andrienko G, Andrienko N, Rinzivillo S, Nanni M, Pedreschi D, Giannotti F (2009) Interactive visual clustering of large collections of trajectories, *Proceedings of the IEEE Symp. Visual Analytics Science and Technology (VAST)* pp: 3-10.

Banerjee A, Dhillon I, Ghosh J, Merugu S, Modha DS (2007) A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8, 1919-1986.

Deng M, Liu, QL, Wang, JQ, Shi, Y (2013). A general method of spatio-temporal clustering analysis. *Science China Information Sciences*, 56, 1-14

Han J, Kamber M, Pei J (2012) *Data Mining Concepts and Techniques*. Morgan Kaufman MIT press.

Lieth, H (1974) Phenology and seasonality modeling. Ecological Studies, 8.

Schwartz MD, Ault TR, Betancourt JL (2013) Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *International Journal of Climatology*, 33, 2917-2922.

Wu X, Zurita-Milla R, Kraak M-J (2015) Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science*, DOI:10.1080/13658816.2014.994520.