Tri-clustering: a novel approach to explore spatio-temporal data cubes

R. Zurita-Milla*, X. Wu, E. Izquierdo-Verdiguier, M.J. Kraak.

Faculty of Geo Information Science and Earth Observation (ITC), University of Twente,
PO Box 217, 7500 AE Enschede, the Netherlands
*Email: r.zurita-milla@utwente.nl

Abstract

Clustering is a popular technique to explore patterns buried in large and complex spatio-temporal datasets. However, most clustering studies only analyse the data from either the spatial or the temporal dimensions. Here we present a novel clustering approach based on Bregman cuboid average tri-clustering (BCAT), which enables the complete analysis of data cubes. We demonstrate the potential of our clustering approach by analysing a georeferenced time series of daily average temperature in the Netherlands. BCAT was used to identify tri-clusters with similar temperature values along the spatial (weather stations) and two nested temporal dimensions (year and day). Then, k-means is used to extract an optimum number of irregular tri-clusters. This application demonstrates that BCAT allows a complete analysis of data cubes and, as such, the proposed approach contributes to a better understanding of complex spatio-temporal patterns.

Keywords: clustering, data mining, geovisualization, geo-referenced time series

1. Introduction

Clustering is a key task in various types of geospatial analysis because it supports the extraction of patterns from large and complex datasets (Andrienko et al. 2009). Clustering provides an overview of the data at a higher level of abstraction and allows the extraction of important information cues by focusing on particular data clusters. Several studies have used clustering to analyse patterns in spatio-temporal datasets (e.g. Hagenauer and Helbich 2013, Grubesic et al. 2014). In these studies, clustering is used to group the data elements along a single data dimension (e.g. space or time) based on similar values along the other dimension. In recent times, Wu et al. (2015) and Wu et al. (2016) used co-clustering to analyse patterns in georeferenced time series. Co-clustering allows to concurrently group data elements along their spatial and temporal dimensions. However, neither clustering nor co-clustering are capable of analysing three dimensional datasets.

In this paper we present a novel approach to analyse complex spatio-temporal data cubes by using tri-clustering so that the data can be fully analysed across its three dimensions. This study is illustrated by a analysing a georeferenced time series of daily temperatures in the Netherlands. This dataset fits into a cube where each cell contains a temperature value indexed by its location and timestamps (year and day) of measurement.

2. Methods

Here we describe the development of the Bregman cuboid average tri-clustering algorithm (BCAT). The development is guided by the use case (daily temperature data collected at m stations for n years) so that the explanation becomes less abstract.

BCAT is an extension of Bregman block average co-clustering algorithm with I-divergence used by Wu et al. 2015 and Wu et al. 2016. BCAT allows clustering cubes that contain positive real-values data. This cube can be viewed as co-occurrences among three random variables: the stations, the years and the days of the year. In our case, the stations represent fixed geographical locations, the columns represent years in which daily average temperatures were measured and the depth of the cube contains the 365 days of a year (February, 29th). To avoid negative temperatures, the absolute value of the minimum temperature was added to all the average temperatures in the data cube.

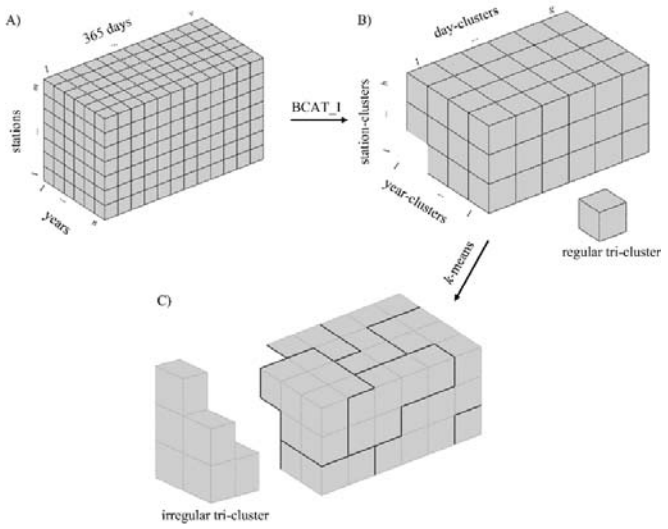

Figure 1: tri-clustering data cubes A) the cube for the temperature data;
B) regular tri-clusters; C) irregular tri-clusters (Wu et al., 2017)

BCAT concurrently groups stations to station-cluster, years to year-clusters and days to day-clusters, to identify tri-clusters that contain similar daily average temperatures (Figure 1B). Like BBAC_I, the composition of the tri-clusters is optimized by using the I-divergence similarity metric (Banerjee et al. 2007). The tri-clustering problem is then an optimization problem where the optimal results minimize the loss of mutual information between the original and the tri-clustered data cube created by providing the number of expected clusters in each dimension.

Because the number of tri-clusters is set a priori, these might still contain very similar values (Wu et al. 2015, Wu et al. 2016). To overcome this problem we propose re-grouping them into non-cubical but axis-parallel tri-clusters (Figure 1C). For a detailed explanation of BCAT, please refer to Wu et al., 2017.

3. Case study: exploring inter-annual temperature variability

Time series of Dutch daily average temperatures are used to illustrate the potential of the proposed tri-clustering approach. More explicitly, we used temperature data collected at 28 Dutch meteorological stations over 20 years (from 1st of January 1992 to 31st of December 2011). To "spatialise" this dataset we collected the coordinates of the weather stations and generated a Thiessen polygon map that defines the area of influence of each weather station.

As briefly explained in section 2, BCAT requires the definition of the number of expected clusters for each the dimensions of the data cube. Based on experimentation and past research with the same dataset (Wu et al.) we decided to use 4 station-clusters, 4 year-clusters and 8 day-clusters. These regular tri-clusters were subsequently refined into *k* irregular tri-clusters, where the optimal value of *k* was found by using the Silhouette method. In our case that meant that the 128 (4 × 4 × 8) regular tri-clusters were re-grouped into 20 irregular tri-clusters.

The 20 irregular tri-clusters were used to analyse the spatio-temporal patterns of intra-annual temperature variability. Six unique spatial patterns of intra-annual variability were found after examining all day-clusters in the four year-clusters (figure 2A). Timelines were used to represent the temporal patterns of temperature variability within each of four year-clusters (Figures 2B to 2E). These patterns were extracted from the irregular tri-clusters by combining day-clusters with the same spatial patterns.

As we can see in Figure 2, the Netherlands has intricate patterns of intra-annual variability in temperature. For most of the study period, the variability in temperature defines two regions in the country: the northeast and centre, and the southwest. The northeast and centre of the country experiences an intense variability in spring and winter temperatures while the southwest only experiences such a variability in spring. Summer temperatures are much more homogeneous across the whole country.

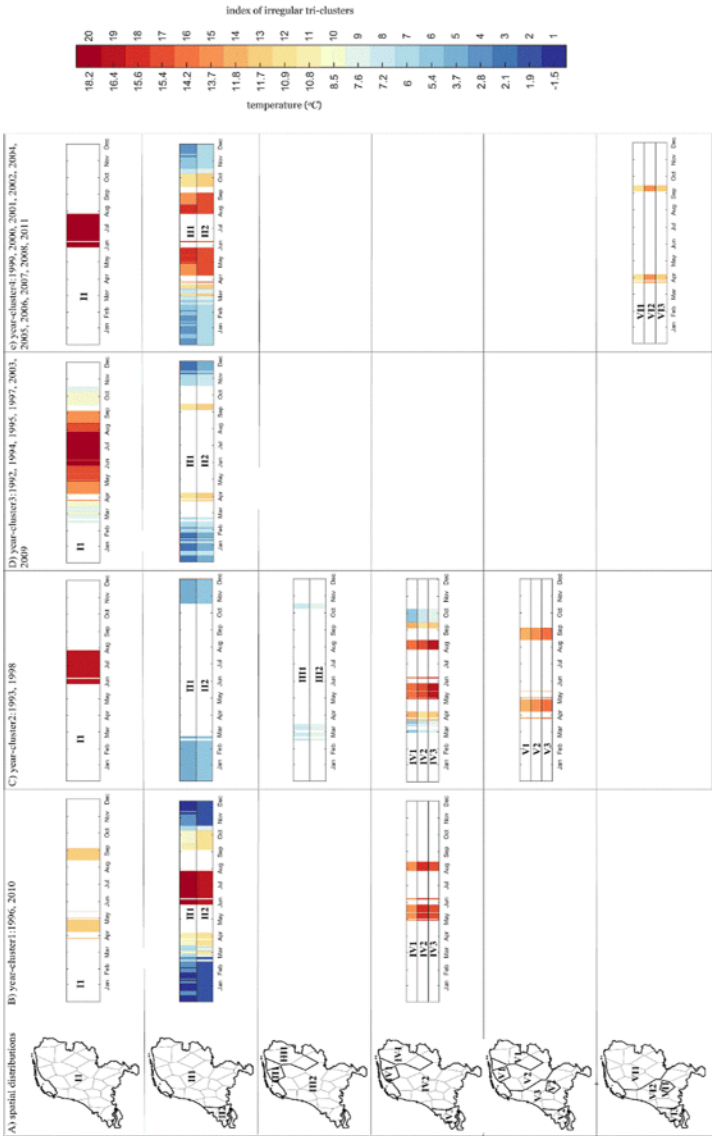

Figure 2: Spatio-temporal patterns of intra-annual variability in average daily temperatures A) the unique spatial patterns of intra-annual variability. B-E) timelines (aligned with the corresponding spatial patterns) of the temporal patterns of temperature variability within each of four year-clusters.

4. Conclusions

In this paper we briefly present a newly developed tri-clustering algorithm: Bregman cuboid average tri-clustering (BCAT). This algorithm allows the analysis of geospatial data cubes that have three dimensions (e.g. space, time and a nested spatial or temporal dimension). Unlike clustering or co-clustering, BCAT can concurrently analyse the three dimensions of the data. The resulting tri-clusters can subsequently be refined using *k*-means to identify an optimal number of clusters in the data cube.

The usefulness of BCAT was demonstrated by analysing time series of Dutch daily average temperature collected from 1992 to 2011. In this application, the data cube has one spatial (weather stations) and two nested temporal dimensions (year and day). Our tri-clustering analysis identified groups of stations and years that have similar within-year temperature variability. Identifying and mapping spatio-temporal patterns of intra-annual variability is important to facilitate the understanding of climate change impacts on our planet. For instance to study the impact of climate change on plant phenology.

Although the use case focus on analysing temperature data, the proposed BCAT-based tri-clustering approach is generic and that, as such, it can be applied to any other data cube (e.g. to remote sensing images with spatial, spectral and temporal dimensions). We hope that BCAT can contribute to a better understanding of the patterns buried in the large and complex spatio-temporal data cubes. A challenging task considering the huge amount of data cubes that are being produced or released at the moment.

5. References

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

Grubesic, T. H., Wei R., and Murray, A. T. 2014. Spatial clustering overview and comparison: accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers*, 104, 1134-1156.

Hagenauer, J. and Helbich M. 2013. Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science*, 27, 2026-2042.

Wu, X., Zurita-Milla, R. and Kraak, M.J. 2015. Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science*, 29, 624-642.

Wu, X., Zurita-Milla, R. and Kraak, M.J. 2016. A novel analysis of spring phenological patterns over Europe based on co-clustering. *Journal of Geophysical Research: Biogeosciences*, 121, 1434-1448.

Wu, X., Zurita-Milla, R., Izquierdo-Verdiguier E. and Kraak, M.J. 2017. Tri-clustering geo-referenced time series for analyzing patterns of intra-annual variability in temperature. *Annals of the Association of American Geographers*, submitted.