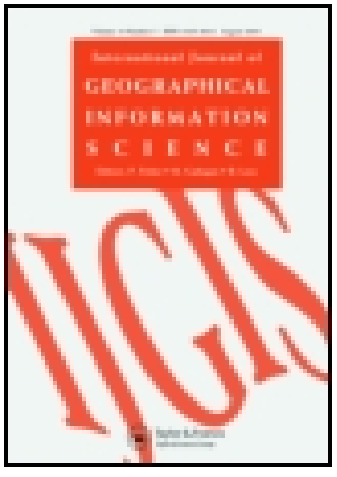

International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data

Xiaojing Wu[a], Raul Zurita-Milla[a] & Menno-Jan Kraak[a]
[a] Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
Published online: 10 Feb 2015.

To cite this article: Xiaojing Wu, Raul Zurita-Milla & Menno-Jan Kraak (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

To link to this article: http://dx.doi.org/10.1080/13658816.2014.994520

International Journal of Geographical Information Science, 2015
http://dx.doi.org/10.1080/13658816.2014.994520

Taylor & Francis
Taylor & Francis Group

Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data

Xiaojing Wu*, Raul Zurita-Milla and Menno-Jan Kraak

Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

(*Received 30 June 2014; final version received 30 November 2014*)

Clustering allows considering groups of similar data elements at a higher level of abstraction. This facilitates the extraction of patterns and useful information from large amounts of spatio-temporal data. Till now, most studies have focused on the extraction of patterns from a spatial or a temporal aspect. Here we use the Bregman block average co-clustering algorithm with I-divergence (BBAC_I) to enable the simultaneous analysis of spatial and temporal patterns in geo-referenced time series (time evolving values of a property observed at fixed geographical locations). In addition, we present three geovisualization techniques to fully explore the co-clustering results: heatmaps offer a straightforward overview of the results; small multiples display the spatial and temporal patterns in geographic maps; ringmaps illustrate the temporal patterns associated to cyclic timestamps. To illustrate this study, we used Dutch daily average temperature data collected at 28 weather stations from 1992 to 2011. The co-clustering algorithm was applied hierarchically to understand the spatio-temporal patterns found in the data at the yearly, monthly and daily resolutions. Results pointed out that there is a transition in temperature patterns from northeast to southwest and from 'cold' to 'hot' years/months/days with only 3 years belonging to 'cool' or 'cold' years. Because of its characteristics, this newly introduced algorithm can concurrently analyse spatial and temporal patterns by identifying location-timestamp co-clusters that contain values that are similar along both the spatial and the temporal dimensions.

Keywords: co-clustering; geo-referenced time series; geovisualization; spatio-temporal pattern; temperature

1. Introduction

Large amounts of spatio-temporal data are becoming available both to the scientific community and to the general public due to advances in data collection and data sharing techniques. Extracting patterns from these data is essential for improving decision-making in many application areas (Li *et al.* 2012, Zurita-Milla *et al.* 2013). However, this task is not trivial, and it has been compared with searching for a needle in a haystack when the data come in large volumes (Keim and Kriegel 1996). In this case, spatio-temporal data mining is necessary to reveal hidden information in the data and to transform it into useful knowledge for a variety of users (Hagenauer and Helbich 2013).

Clustering is an important task in spatio-temporal data mining that aims at identifying groups of data elements that are similar among them but dissimilar to the elements present in other groups (Han *et al.* 2009). Clustering allows considering groups of

*Corresponding author. Email: x.wu-1@utwente.nl

similar data elements at a higher level of abstraction and, therefore, facilitates the extraction of patterns and useful information (Andrienko *et al.* 2009, Hagenauer and Helbich 2013). Hence, the clustering of spatio-temporal data is especially useful when dealing with large amounts of data. In this regard, there are several types of data according to the extension and combination of the spatial and temporal dimensions (Kisilevich *et al.* 2010): (1) spatio-temporal events; (2) geo-referenced variables; (3) geo-referenced time series (GTS); (4) moving objects and (5) trajectories. Consequently, there are different spatio-temporal clustering methods. In this study only spatio-temporal clustering of GTS is considered. GTS contain time evolving values for an observed property that is recorded at fixed locations typically, but not necessarily, at uniform intervals, for instance, average daily temperature measured by a network of weather stations distributed over an area.

Many studies on pattern analysis in GTS using clustering can be found in literature (Crane and Hewitson 2003, Zhang *et al.* 2003, Guo *et al.* 2006, Wu *et al.* 2008, 2013, Andrienko *et al.* 2010a, Hagenauer and Helbich 2013). Crane and Hewitson (2003) clustered precipitation records of individual stations into regional data sets using self-organizing maps (SOMs) to analyse regional patterns. Zhang *et al.* (2003) proposed a cone-based filter-and-refine algorithm to detect correlations and therefore form clusters in Earth science data. Wu *et al.* (2008) suggested a range-based searching nearest neighbours (RSNN) spatial clustering to mine patterns in climate data. Guo *et al.* (2006), Andrienko *et al.* (2010a) and Wu *et al.* (2013) all analysed spatio-temporal pattern in GTS using SOMs in a dual way: identify spatial clusters of similar temporal distributions and temporal clusters of similar spatial situations. Hagenauer and Helbich (2013) analysed spatio-temporal patterns in socio-economic data using SOMs in a hierarchical structure where spatial and temporal patterns are analysed independently in the upper layer and then merged in the lower layer. However, in all these studies clustering is used to analyse spatial or temporal patterns. Such analysis violates the inseparability of space and time stated by Hagerstand (1970) since 'there is nothing spatial that is not temporal' (Andrienko *et al.* 2010b). Also, patterns extracted only relying on spatial clustering cannot fully describe the time-varying behaviour present in the data and vice versa (Deng *et al.* 2011). This deficiency necessitates of a clustering method capable of mining spatial and temporal patterns in a concurrent fashion.

Another important issue of spatio-temporal patterns analysis in GTS lies with temporal resolution of the data. As stated by Li and Chou (2000), the results of pattern analysis could be different when the temporal resolution of input data changes. This issue has attracted attentions recently and is part of a broad research problem known as the Modifiable Temporal Unit Problem (MTUP; Coltekin *et al.* 2011, Jong and Bruin 2012, Wu *et al.* 2013).

To fully explore the spatio-temporal patterns in GTS, appropriate geovisualization techniques are needed. Geovisualization is an integrated approach from cartography, scientific visualization, exploratory data analysis and GIScience (Dykes *et al.* 2005, Miller and Han 2009). Geovisualization techniques are to stimulate visual thinking and help exploit spatial and temporal patterns in the data.

Considering the aforementioned problems, this study introduces to the geo-community a so-called co-clustering algorithm, which enables the concurrent analysis of spatial and temporal patterns in GTS. In addition, we present three geovisualization techniques: heatmaps, small multiples and ringmaps that support the exploration of the results of this co-clustering algorithm.

2. Data and methods

This section first introduces the study area and the data used in this study. Then background information about clustering and co-clustering is given. After that, the specific co-clustering algorithm used in this study – the Bregman block average co-clustering algorithm with I-divergence (BBAC_I) – is presented. Finally, four geovisualization techniques that facilitate the exploration of the co-clustering results are described.

2.1. Study area and data

Daily average temperatures collected at 28 Dutch weather stations from 1 January 1992 to 31 December 2011 (i.e. over 20 years) are used to illustrate this study. This freely available data were downloaded from the website of the Royal Netherlands Meteorological Institute, KNMI (https://data.knmi.nl/portal/KNMI-DataCentre.html).

The Netherlands is located in the northwest of Europe (bottom right of Figure 1). The west and north are bordered by the North Sea, whereas the south and the east border with Belgium and Germany, respectively. Thus, even though the Netherlands only covers about 41,500 km^2, Dutch weather is influenced by both oceanic (in the southwest) and continental (in the northeast) climates. As a result, temperatures in the southwest and northeast are different.

To illustrate the spatio-temporal patterns present in this data set, a Thiessen polygon map was generated based on geographic coordinates of all weather stations (also available on the KNMI website) to define the area influenced by each station. The Thiessen polygon map is shown in Figure 1 where each polygon is labelled by the station ID (e.g. 290) and its name (e.g. Twente).

2.2. Clustering and co-clustering algorithms

GTS, such as the temperature series available for this study, can be organized in a data matrix where rows indicate space (in our case, weather stations) and columns are used to store the values at different timestamps (e.g. years). Until now this kind of matrices have been clustered using one-way algorithms. That is, algorithms that regard stations as objects and timestamps as attributes and the clustering consists on partitioning stations into groups based on similarities along all timestamps or the other way around. The most popular one-way clustering algorithm is k-means. Take clustering from spatial perspective as an example. After the initialization of randomly chosen k station-cluster centroids, the sum of squared errors between each station and its corresponding station-cluster centroids is calculated. This sum constitutes the objective function that k-means minimizes iteratively by assigning each station to the closest cluster centroid and re-computing new station-cluster centroids. These iterations cease when a convergence is met (e.g. the sum decreases to a predefined threshold) and result in the optimal k station-clusters that best represent all stations.

Co-clustering algorithms, however, treat objects and attributes equally (Han *et al.* 2012) by mapping stations to station-clusters and timestamps to timestamp-clusters at the same time. Suppose that the stations are to be grouped into k disjoint clusters and the timestamps into l disjoint clusters. After the initialization of the k station-clusters and l timestamp-clusters, a co-clustered data matrix with size $k \times l$ is determined. The difference between the original temperature data matrix and the co-clustered one is calculated as the distortion function (Anagnostopoulos *et al.* 2008). Then co-clustering algorithms

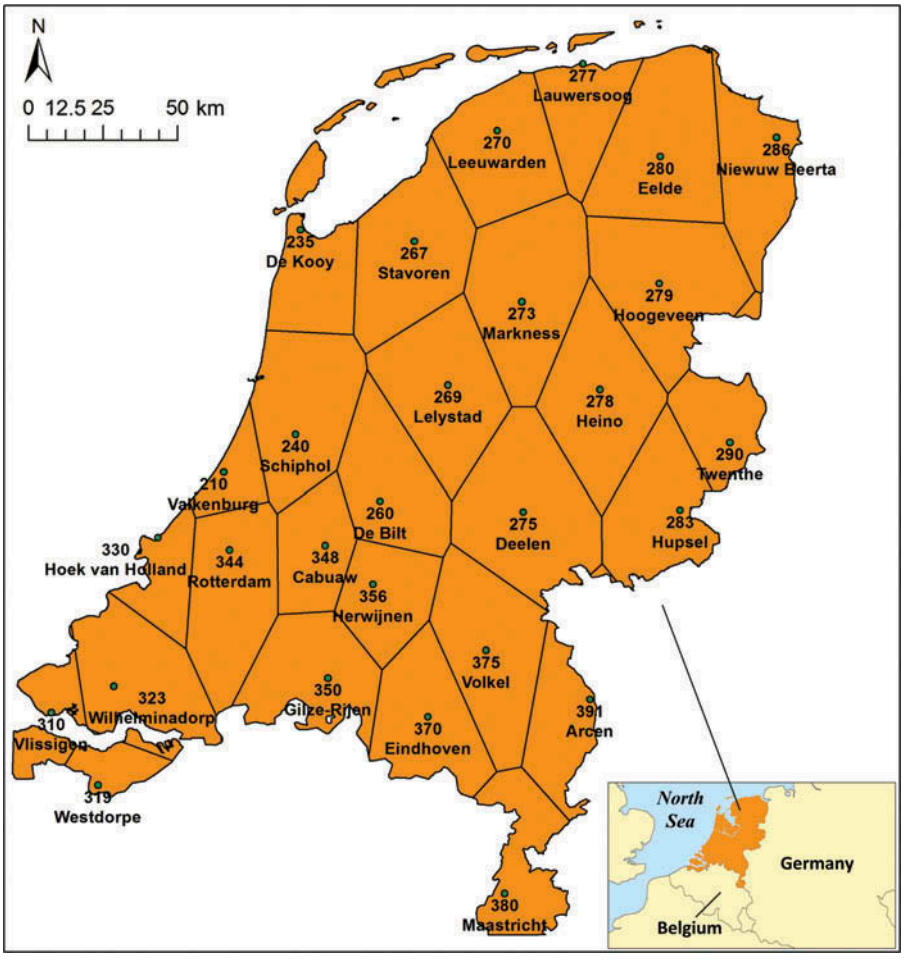

Figure 1. Location of the Netherlands in Europe (bottom right) and the Thiessen polygon map of the Netherlands to define influenced area of each station.

minimize the distortion function by iteratively assigning each station to the nearest station-cluster and each timestamp to the nearest timestamp-clusters. This process stops when a convergence is met and yields the optimal co-clustering as results. Then a re-ordered data matrix is generated according to the co-clustering results. That is, all stations/timestamps that belong to the same co-cluster are put together by swapping rows and columns of the original matrix. A co-cluster is defined in the re-ordered matrix as the intersection of a station/timestamp-cluster and a timestamp/station-cluster. In this sense, co-clustering enables the analysis of spatial patterns with the consideration of time-varying behaviour of timestamps (temporal patterns) and vice versa.

Several works have studied co-clustering algorithms (Dhillon *et al.* 2003, Cho *et al.* 2004, Banerjee *et al.* 2007). Dhillon *et al.* (2003) introduced the information theoretic co-clustering algorithm, which uses I-divergence as the distortion function and also preserve a set of linear summary statistics (e.g. row/column and co-cluster averages) of the original data matrix in the iterative process of co-clustering. Cho *et al.* (2004)

introduced the minimum sum-squared residue co-clustering algorithm, which uses squared Euclidean distance in the distortion function and preserves the row and column averages of each co-cluster in the process. Banerjee *et al.* (2007) generalized above works as Bregman co-clustering algorithm that allows several distortion functions and enables various co-clustering schemes that preserve different sets of summary statistics. In this study, we chose the I-divergence because Banerjee *et al.* (2007) empirically proved its superiority. Besides, we chose the co-clustering scheme that preserves co-cluster averages because it considers the variations among temperature values within each co-cluster along both spatial (stations) and temporal (timestamps) dimensions. This algorithm is termed as Bregman block average co-clustering (BBAC) in (Banerjee *et al.* 2007). Here, we call it BBAC_I as we use the I-divergence. The next section explains this co-clustering algorithm in more detail.

2.3. *Bregman block average co-clustering algorithm with I-divergence (BBAC_I)*

The BBAC_I algorithm enables the co-clustering of any data matrix with positive and real valued elements that typically represent a joint probability distribution or co-occurrences between two random variables. Our temperature matrix can be regarded as a co-occurrence matrix between the stations (S) and the timestamps (T). In more formal terms, the temperature matrix can be represented as *O(S, T)* where S takes values in the station sets $\{s_1, \ldots, s_m\}$ and T in the timestamp sets $\{t_1, \ldots, t_n\}$. Accordingly, the co-clustered data matrix is $O(\hat{S}, \hat{T})$, where $\hat{S}$ take values in the station-cluster sets $\{\hat{s}_1, \ldots, \hat{s}_k\}$ $(k \leq m)$ and $\hat{T}$ in the timestamp-cluster sets $\{\hat{t}_1, \ldots, \hat{t}_l\}(l \leq n)$.

Being part of the information theoretical co-clustering family, the BBAC_I algorithm regards the co-clustering from *O(S,T)* to $O(\hat{S}, \hat{T})$ as an optimization problem. In information theory, the amount of information shared between two variables is called mutual information and the optimal co-clustering minimizes the loss in mutual information before and after performing the co-clustering. Therefore, the optimal co-clustering minimizes $I(S, T) - I(\hat{S}, \hat{T})$, where $I(,)$ denotes the mutual information between the variables at hand. In the following we present the summarized five steps of this co-clustering algorithm (Figure 2):

Step 1: to randomly initialize the mapping of stations to station-clusters and timestamps to timestamp-clusters. This is a prerequisite step to calculate the loss in mutual information in the next step.

Step 2: to calculate the loss in mutual information before and after mapping. BBAC_I measures the loss in mutual information with I-divergence between the original data matrix and a matrix that approximates it:

$$I(S, T) - I(\hat{S}, \hat{T}) = D_I(O(S, T) \,||\, A(S, T)) \tag{1}$$

where $D_I(\cdot||\cdot)$ denotes I-divergence between two matrices; $O(S, T)$ denotes the original data matrix with $o(s, t)$ as elements; $A(S, T)$ is the matrix approximation of $O(S, T)$ with $a(s, t)$ as elements.

The matrix approximation is determined by the original data matrix, the current mapping and the co-clustering scheme. According to the co-clustering scheme of BBAC_I, the matrix approximation is calculated as:

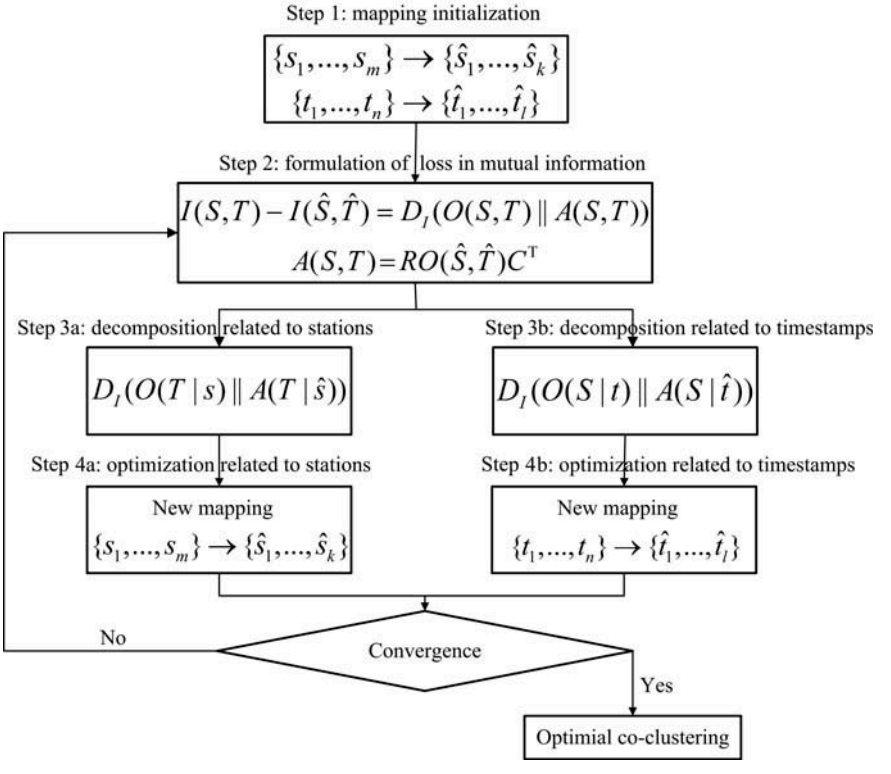

Figure 2. Summarized five steps of the Bregman block average co-clustering algorithm with I-divergence (BBAC_I).

$$A(S,T) = RO(\hat{S},\hat{T})C^{\mathrm{T}} \tag{2}$$

where R and C are binary matrices with size $m \times k$ and $n \times l$to indicate the station- and timestamp-cluster membership separately; C^{T} denotes the transpose of the matrix C.

Then Equation (1) of loss in mutual information can be further represented as

$$D_I(O(S,T) \,||\, A(S,T)) = \sum_{\hat{s}} \sum_{\hat{t}} \sum_{s \in \hat{s}} \sum_{t \in \hat{t}} o(s,t) \log \frac{o(s,t)}{a(s,t)} \tag{3}$$

Step 3: to decompose Equation (3) according to the rows (stations) and columns (timestamps) of the matrix helps to find the new mapping. As demonstrated by Banerjee *et al.* (2007), Equation (3) can be decomposed into the I-divergence of mapping from stations to station-clusters (Step 3a):

$$D_I(O(T \mid s) \,||\, A(T \mid \hat{s})) = \sum_{\hat{s}} \sum_{s \in \hat{s}} o(T \mid s) \log \frac{o(T \mid s)}{a(T \mid \hat{s})} \tag{4}$$

or the I-divergence of mapping from timestamps to timestamp-clusters (Step 3b):

$$D_I(O(S\,|\,t)\,||\,A(S\,|\,\hat{t})) = \sum_{\hat{t}} \sum_{t\in\hat{t}} o(S\,|\,t) \log \frac{o(S\,|\,t)}{a(S\,|\,\hat{t})} \tag{5}$$

Step 4: As aforementioned, the optimal co-clustering is to minimize the loss in mutual information. Now since Equation (3) is decomposed to I-divergences in terms of stations clustering and timestamps clustering separately, then Step 4 is to find the new mapping of each station to station-clusters that minimizes Equation (4) (Step 4a):

$$i = \arg \min_{i\in\{1,\ldots,k\}} D_I(O(T\,|\,s)\,||\,A(T\,|\,\hat{s}_i)) \tag{6}$$

and similarly find the new mapping of each timestamp to timestamp-clusters that minimizes Equation (5) (Step 4b):

$$j = \arg \min_{j\in\{1,\ldots,l\}} D_I(O(S\,|\,t)\,||\,A(S\,|\,\hat{t}_j)) \tag{7}$$

Step 5: to re-calculate the loss of mutual information using the new mapping according to Equation (3). If the change of the loss in mutual information is smaller than a predefined threshold ε (say 10^{-6}), then the new mapping obtained in Step 4 is the optimal co-clustering result; else go to Step 2 to start a new loop.

It has been guaranteed by Banerjee *et al.* (2007) that Equation (3) in Step 2 monotonically decreases the loss in mutual information. Therefore, BACC_I always converges to a local minimum. In practice, various random mappings are used in Step 1 to select the smallest local minimum as the optimum co-clustering of the input matrix.

Let us illustrate how BBAC_I yields the optimal co-clustering with an example. Suppose that we want to co-cluster the co-occurrence matrix between seven stations and five timestamps:

$$O(S,T) = \begin{bmatrix} 2.5 & 3.0 & 3.2 & 5.1 & 5.3 \\ 2.5 & 3.2 & 3.0 & 5.0 & 5.5 \\ 5.2 & 5.2 & 5.0 & 3.1 & 3.2 \\ 5.3 & 5.0 & 5.4 & 3.2 & 3.1 \\ 5.0 & 5.2 & 5.1 & 3.3 & 3.1 \\ 7.8 & 8.0 & 8.0 & 5.2 & 5.3 \\ 7.5 & 7.8 & 7.6 & 5.1 & 5.2 \end{bmatrix} \tag{8}$$

Given the row distribution of the temperature matrix, it is natural to decide that $\hat{S}$ takes values in $\{\hat{s}_1, \hat{s}_2, \hat{s}_3\}$ where $\hat{s}_1 = \{s_1, s_2\}$, $\hat{s}_2 = \{s_3, s_4, s_5\}$, $\hat{s}_3 = \{s_6, s_7\}$. Similarly, $\hat{T}$ takes values in $\{\hat{t}_1, \hat{t}_2\}$ where $\hat{t}_1 = \{t_1, t_2, t_3\}$, $\hat{t}_2 = \{t_4, t_5\}$. Then the resulting co-clustered data matrix is

$$O(\hat{S},\hat{T}) = \begin{bmatrix} 2.9000 & 5.2250 \\ 5.1556 & 3.1667 \\ 7.7833 & 5.2000 \end{bmatrix} \tag{9}$$

It can be verified that above co-clustering is the optimal one since no other co-clustering produces a smaller loss in mutual information.

Such optimal co-clustering can be achieved by BBAC_I. Table 1 shows how BBAC_I iteratively yields the local optimal co-clustering for the example matrix *O(S,T)* illustrated in Equation (8). Each iteration in Table 1 exhibits the steps of the BBAC_I, the resulting co-clustered matrix $O(\hat{S},\hat{T})$ and the matrix approximation $A(S,T)$. The matrices are surrounded by station-cluster and timestamp-cluster number to indicate the mapping. At the end of iterations, BBAC_I precisely achieves the natural mapping from stations to station-clusters and timestamps to timestamp-clusters. Also, it recovers the co-clustered data matrix $O(\hat{S},\hat{T})$ shown in Equation (9) and minimizes the loss in mutual information.

Table 1. BBAC_I in Figure 2 iteratively yields the optimal co-clustering for the example *O(S,T)* in Equation (8).

$\downarrow O(S,T)$			Step 1 & 2 of Figure 2					
$A(S,T)$	$\hat{t}_2$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_2$	5.1167	5.4000	5.4000	5.1167	5.1167	$\hat{s}_1$	6.6000	4.9833
$\hat{s}_3$	3.9889	4.4500	4.4500	3.9889	3.9889	$\hat{s}_2$	5.4000	5.1167
$\hat{s}_3$	3.9889	4.4500	4.4500	3.9889	3.9889	$\hat{s}_3$	4.45000	3.9889
$\hat{s}_1$	4.9833	6.6000	6.6000	4.9833	4.9833			
$\hat{s}_3$	3.9889	4.4500	4.4500	3.9889	3.9889			
$\hat{s}_1$	4.9833	6.6000	6.6000	4.9833	4.9833			
$\hat{s}_2$	5.1167	5.4000	5.4000	5.1167	5.1167			
$\downarrow$			Step 3 & 4 of Figure 2					
$A(S,T)$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_1$	2.9000	2.9000	2.9000	5.2250	5.2250	$\hat{s}_1$	2.9000	5.2250
$\hat{s}_1$	2.9000	2.9000	2.9000	5.2250	5.2250	$\hat{s}_2$	5.1167	3.1750
$\hat{s}_2$	5.1167	5.1167	5.1167	3.1750	3.1750	$\hat{s}_3$	6.9333	4.5167
$\hat{s}_3$	6.9333	6.9333	6.9333	4.5167	4.5167			
$\hat{s}_2$	5.1167	5.1167	5.1167	3.1750	3.1750			
$\hat{s}_3$	6.9333	6.9333	6.9333	4.5167	4.5167			
$\hat{s}_3$	6.9333	6.9333	6.9333	4.5167	4.5167			
$\downarrow$			Step 2 & 3 & 4 of Figure 2					
$A(S,T)$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_1$	2.9000	2.9000	2.9000	5.2250	5.2250	$\hat{s}_1$	2.9000	5.2250
$\hat{s}_1$	2.9000	2.9000	2.9000	5.2250	5.2250	$\hat{s}_2$	5.1556	3.1667
$\hat{s}_2$	5.1556	5.1556	5.1556	3.1667	3.1667	$\hat{s}_3$	7.7833	5.2000
$\hat{s}_2$	5.1556	5.1556	5.1556	3.1667	3.1667			
$\hat{s}_2$	5.1556	5.1556	5.1556	3.1667	3.1667			
$\hat{s}_3$	7.7833	7.7833	7.7833	5.2000	5.2000			
$\hat{s}_3$	7.7833	7.7833	7.7833	5.2000	5.2000			

2.2. Geovisualization techniques

In order to fully exploit co-clustering results, appropriate geovisualization techniques must be used. Here we rely on three techniques: heatmaps, small multiples and ringmaps. Heatmaps offer a straightforward view of the station-clusters, timestamp-clusters as well as station-timestamp co-clusters; small multiples display the spatial distribution of station-clusters and co-clusters in geographic maps and also the temporal distribution of timestamp-clusters associated to linear timestamps; ringmaps illustrate the temporal distribution of timestamp-clusters associated to cyclic timestamps separately. Among them, only small multiples allows visualizing the co-clustering results in geographic space.

2.2.1. Heatmap

A heatmap is a graphical representation of a data matrix where the individual values in the matrix are visualized using a range of colours. As aforementioned, the rows and columns of the original matrix are re-ordered according to the optimal co-clustering results after co-clustering. The re-ordered matrix has the following properties: (1) stations belonging the same station-cluster are arranged together and so are the timestamps; (2) station-clusters are arranged from lowest (bottom) to highest (up) average temperatures along timestamps and (3) timestamps-clusters are arranged from lowest (left) to highest (right) average temperatures along the stations. By this means, all stations and timestamps mapped into the same co-clusters are arranged together in the re-ordered matrix, and using a heatmap to display the re-ordered matrix is therefore a straightforward way to visualize the station-clusters, timestamp-clusters as well as station-timestamp co-clusters. Since the case study deals with temperature values, the selection of an appropriate heat–colour scale provides a rapid view of all the data.

2.2.2. Small multiples

Small multiples are a set of adjacent graphics to support the understanding of multivariate information (Maceachren *et al.* 2003). Map-based small multiples are one kind to use adjacent geographic maps.

The small multiples enable the visualization of spatial patterns of station-clusters and also co-clusters for each of timestamp-clusters. Within a set of small multiples, each map is used to visualize station-clusters or co-clusters for each timestamp-cluster in the re-ordered matrix. Station-clusters or co-clusters are displayed in the map, each station-cluster or co-cluster with one colour interpreted from average temperature values in that station-cluster or co-cluster. Each station-cluster shows up as a region that is not necessarily composed of adjacent stations.

Besides, the small multiples enable the visualization of the temporal distribution of timestamp-clusters associated to linear timestamps. Since the re-ordered matrix arranges timestamps according to timestamp-clusters, it is impossible to see the temporal patterns in temperature in the linear chronological order. In this case, the small multiples arrange one map for each of all timestamps. The colour of each map interpreted from average temperature values in that timestamp-cluster indicates which timestamp-cluster the timestamp belongs to.

2.2.3. Ringmap

Ringmaps, first proposed by Zhao *et al.* (2008), are an extension of the circular timeline. They are composed of multiple concentric rings where each ring illustrates values for an entity related to cyclic time. Therefore, it not only enables the visualization of values related to cyclic time but also the comparison of values between entities.

Ringmap arranges cyclic timestamps in circles and each timestamp indicates which timestamp-cluster it belongs to with colours. The colours are interpreted from the average temperature value of timestamps in each timestamp-cluster for all stations.

3. Case study: co-clustering temperature data at different temporal resolutions

The BBAC_I algorithm described in Section 2.3 was used to hierarchically analyse the spatio-temporal patterns in the Dutch temperature series at yearly, monthly and daily resolutions.

This was done using the workflow illustrated in Figure 3. First, the daily temperature data were averaged to create a yearly temperature data set, which was organized as a 28 × 20 data matrix (stations × years). The proposed co-clustering algorithm was applied to this matrix to obtain the station-year co-clusters at yearly resolution. A heatmap was used to visualize these results, and based on the visualization, the labels ‘cold’, ‘cool’, ‘warm’ and ‘hot’ years were assigned to the year-clusters. After that, small multiples were used to display and explore the resulting spatio-temporal patterns. Then, for each year-cluster a representative monthly temperature data set was created by calculating the, per station, average monthly temperature using the temperature records of all the years

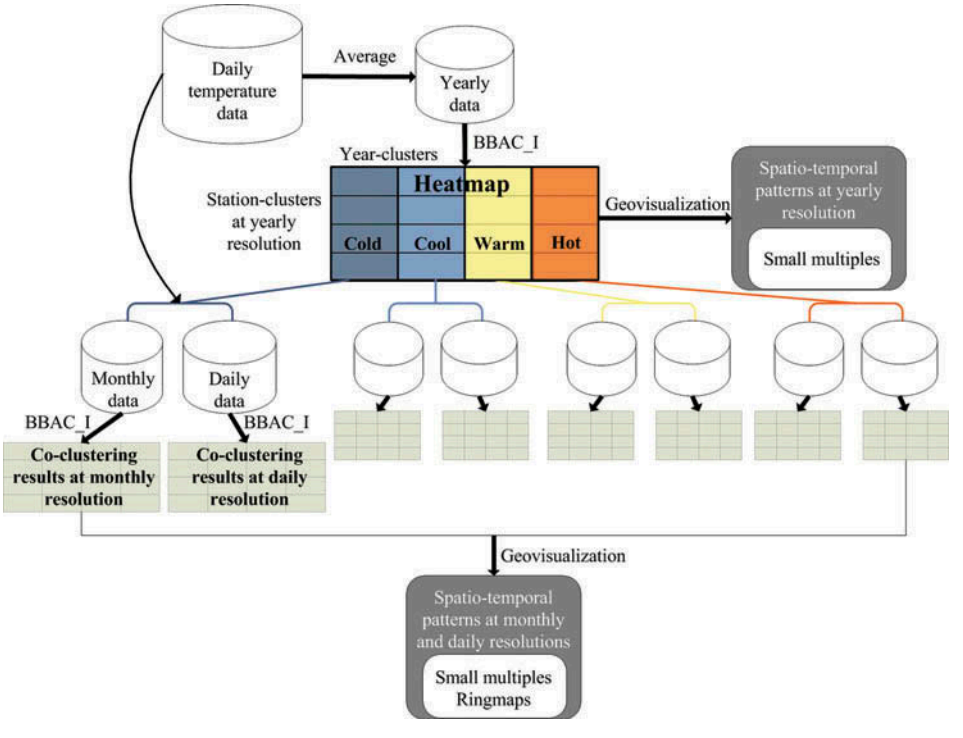

Figure 3. Workflow of the experiment at different temporal resolutions.

belonging to that year-cluster. This resulted in four monthly temperature matrices of 28 × 12 (stations × months). Each of these matrices, associated to 'cold', 'cool', 'warm' and 'hot' years, was processed with BBAC_I to identify station-month co-clusters for each year-cluster. Similarly, four representative daily temperature matrices of 28 × 365 (stations × days) were created for the four year-clusters by calculating the, per station, average daily temperature using the temperature records of all the years belonging to each year-cluster. These four matrices were also subjected to BBAC_I to identify station-day co-clusters for each year-cluster. After that, small multiples combined with ringmaps were used to illustrate the co-clustering results as well as to compare the spatio-temporal patterns obtained both at the monthly and the daily temporal resolutions.

The number of station-clusters and timestamp-clusters should be specified before the co-clustering analysis. Like any other clustering algorithms, the selection, optimization and evaluation of cluster numbers in co-clustering remains challenging and depends on specific applications. In this study, the number of station-clusters was empirically chosen as four based on previous clustering results of the same data presented by Wu *et al.* (2013); the number of timestamp-clusters was set to 4 to categorize 'cold', 'cool', 'warm' and 'hot' years, months and days. Other parameters in the analysis were set as follows: the threshold of change in loss of mutual information (ε) was empirically set to 10^{-6}, which is small enough to guarantee the quality of the optimal co-clustering results for each loop in this study; the number of loops was set to 1000 within which the convergence of BBAC_I with the data in this case study can be assured. Besides, 100 times of random mapping were used for initialization in Step 1 to explore various local minima and select the smallest one. Also, from a computational point of view, it is worth mentioning that both the co-clustering analysis and the geovisualization techniques were implemented in MATLAB version 2014a and that the BBAC_I algorithm used in this study was downloaded from http://www.ideal.ece.utexas.edu/software/bregcc_code.tar.

4. Results and discussions

4.1. Spatio-temporal patterns at yearly resolution

The BBAC_I was applied to the aggregated yearly temperature data matrix to study the spatio-temporal patterns at this temporal resolution. After co-clustering, the 28 stations were mapped to four station-clusters and the 20 years were grouped into four year-clusters.

The heatmap in Figure 4 straightforwardly displays the station-clusters, year-clusters as well as station-year co-clusters in the results. Small multiples in Figure 5 show the spatial distribution of station-clusters (Figure 5(a)) and temporal distribution of year-clusters (Figure 5(b)). Besides, the small multiples in Figure 6 illustrate the spatial patterns of station-year co-clusters.

The re-ordered yearly temperature matrix is showed in Figure 4 using a heatmap where yellow indicates low temperature and red means high temperature. The values of the *x*-axis show the years belonging to each year-clusters. These clusters have been re-ordered from low to high temperatures and thus moving from left to right in the *x*-axis we have the 'cold', 'cool', 'warm' and 'hot' years. From the bottom to the top of the *y*-axis, the heatmap shows the stations IDs (as used in Figure 1) arranged in the order from station-cluster4 to station-cluster1 with increasing temperature values for all years. Each map of the small multiples in Figure 5(a) shows the spatial distribution from station-cluster1 to station-cluster4: the coloured area indicates the region of the station-cluster and the colour symbolizes the average temperature values in that station-cluster. The higher

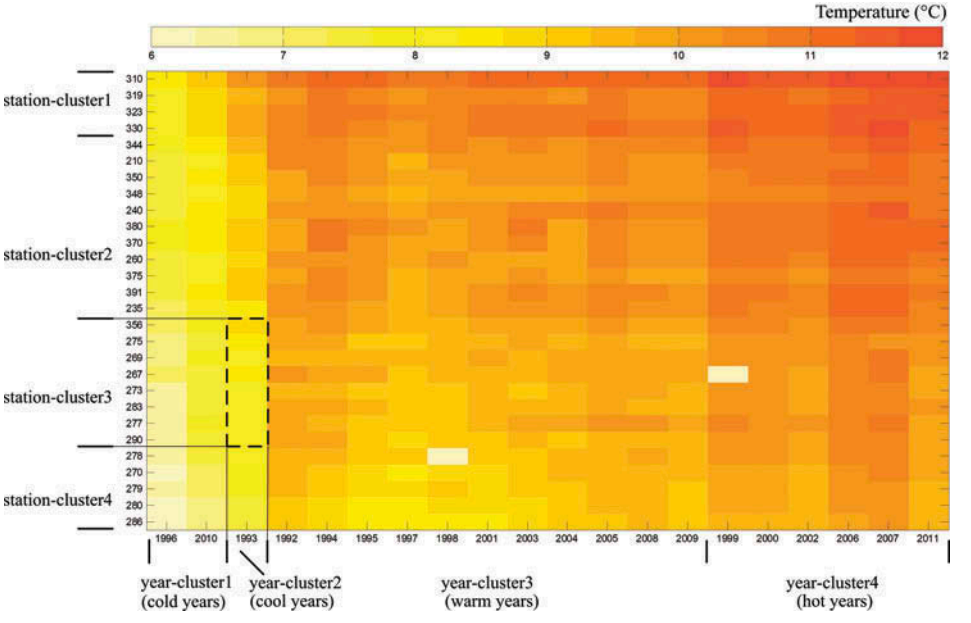

Figure 4. Heatmap to visualize the co-clustering results at yearly resolution, where light yellow indicates low temperature and red means high temperature; *x*-axis: years arranged according to year-clusters with increasing temperature values; *y*-axis: stations ordered according to station-clusters with increasing temperature values. An example of a co-cluster: the intersection between station-cluster3 and year-cluster2 is indicated by means of a dashed rectangle.

the temperature, the darker the colour. Figure 5(a) clearly shows that the Netherlands is divided into four regions from left to right: southwest, centre-southwest, centre-northeast and northeast. These four regions reveal a decreasing temperature pattern across the country, which confirms the expected temperature patterns: oceanic (milder) in the southwest and more continental in the northeast. Figure 5(a) also shows almost all stations in each station-cluster are geographically adjacent. Each map of the small multiples in Figure 5(b) shows the year-cluster to which that particular year belongs to. This time, the colour represents the average temperature values in that year-cluster. Figure 5(b) shows that yearly temperature in the Netherlands from 1992 to 2011 are mostly 'warm' or 'hot', with only 3 years belonging to 'cool' (1993) or 'cold' (1996, 2010) years. Such results agree with the clustering results produced with SOMs (Wu *et al.* 2013). It is remarkable that the temperature is increasing in recent years since before 1998 there was no 'hot' year and the variations of temperature in were between 'warm' and 'cool'/'cold', whereas variations in recent years occur between 'warm' and 'hot' years. Such an increase in temperature might be associated to global warming.

Since the elements mapped to each station- and year-cluster are clearly illustrated in the heatmap, Figure 4 also displays the station-year co-clusters straightforwardly. For instance, the dashed rectangle in the bottom left of Figure 4 shows one of the 16 co-clusters. This co-cluster indicates that the temperatures at the stations (356, 275, 269, 267, 273 283, 277, 290: station-cluster3) in the year (1993) are similar. Based on the order of station-clusters and year-clusters, the co-cluster with the lowest temperature is in the bottom-left corner, intersected by station-cluster4 and 'cold' years while the one with the highest temperature is in the top-right corner, intersected by station-cluster1 and 'hot'

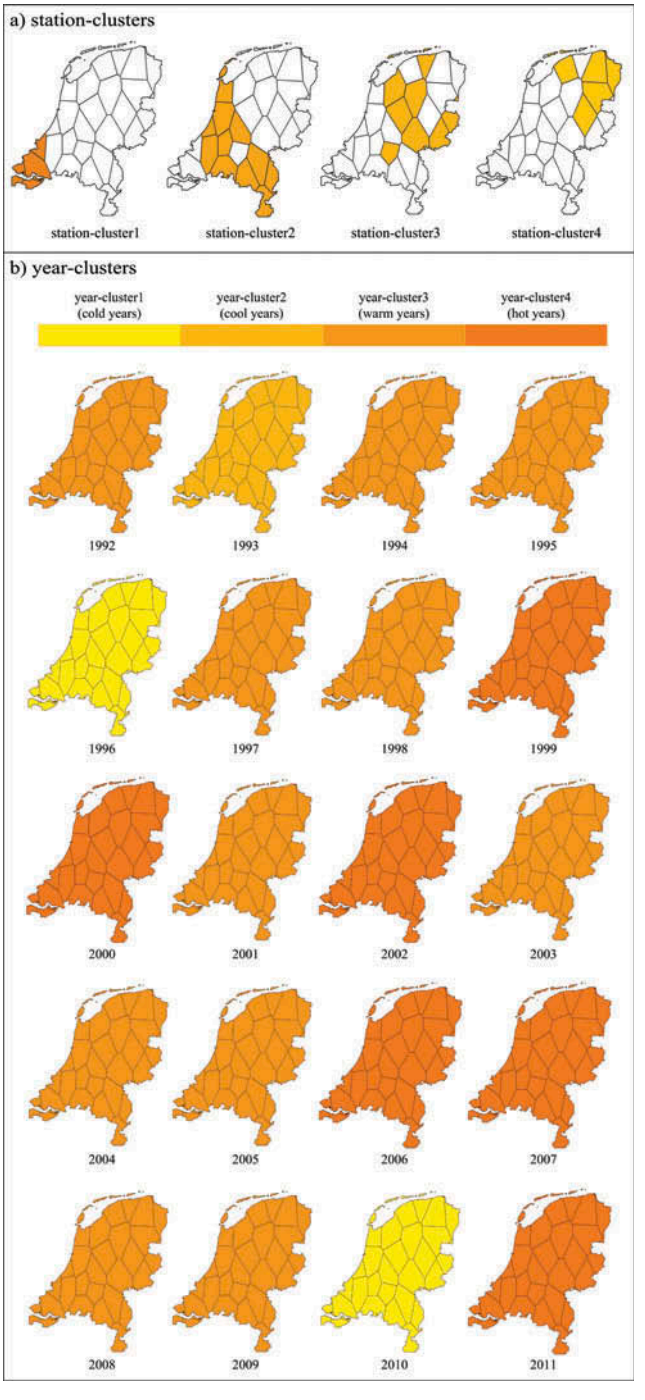

Figure 5. Small multiples to display spatial distribution of station-clusters and temporal distribution of year-clusters. (a) Small multiples to show the spatial distribution of four station-clusters: within each map the coloured area indicates the region of the station-cluster. The higher is the average temperature values of the station-cluster, the darker the colour is. (b) Small multiples to show the temporal distribution of the four year-clusters over 20 years: one map for each year indicating its year-cluster. The higher is the average temperature values of the year-cluster, the darker the colour is.

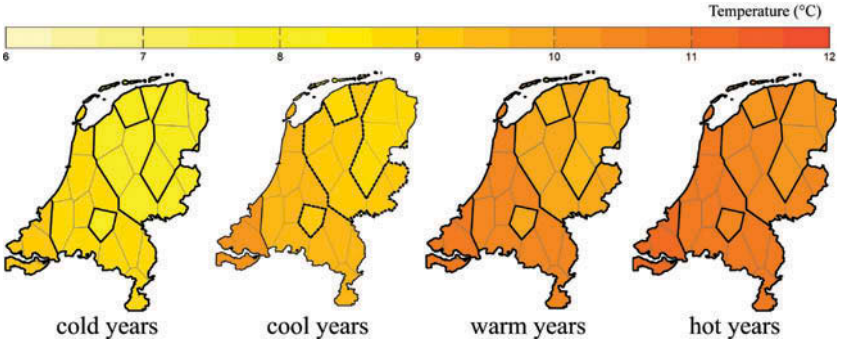

Figure 6. Small multiples to display spatial patterns in station-year co-clusters for each year-cluster, where light yellow indicates low temperature and red means high temperature. Region within dashed boundary lines in the second map corresponds with the co-cluster within the dashed rectangle in Figure 4.

years. Therefore, from left to right and from bottom to top the temperature values of the co-clusters become increasingly high. Each map of the small multiples in Figure 6 displays the spatial patterns of station-year co-clusters shown in Figure 4: four station-year co-clusters for each year-cluster. These maps show the typical spatial patterns of station-year co-clusters in 'cold', 'cool', 'warm' and 'hot' years from left to right. In each map, the four regions mentioned earlier: southwest, centre-southwest, centre-northeast and northeast are indicated by thick lines. Each of regions corresponds to each of the station-year co-clusters shown in Figure 4 and the value is the average temperatures in the co-cluster. The spatial composition of these four regions is the same for all year-clusters because the BBCA_I assigns complete rows of the data matrix to station-clusters. From southwest to northeast of the Netherlands and from 'cold' to 'hot' years, those station-year co-clusters reveal the same decreasing temperature pattern as that in station-clusters. Therefore, the northeast region in 'cold' years has the lowest temperature values while the southwest region in 'hot' years has the highest temperature values. The temperature values at other regions in year-clusters are between the two extremes and become increasingly high from 'cold' to 'hot' years and from northeast to southwest. The centre-northeast region bordered with a dashed line in the 'cool' year corresponds with the dashed rectangle displayed in the heatmap in Figure 4.

Moreover, the heatmap in Figure 4 can be used to detect anomalies, For instance, the two low temperature values at station 278 in 1998 and 267 in 1999 are because some missing values in the original daily data are replaced with zeros, which causes the abnormal low yearly temperature.

4.2. *Spatio-temporal patterns at monthly and daily resolutions*

The four representative monthly and four daily temperature matrices that were calculated by averaging the data belonging to 'cold', 'cool', 'warm' and 'hot' years were subjected to the BBAC_I algorithm too. After co-clustering, each of monthly (28 × 12) temperature matrices were mapped to one set of 4 × 4 station-month co-clusters and in all there were four sets associated to 'cold', 'cool', 'warm' and 'hot' years. Similarly, each of the daily temperature matrices (28 × 365) were mapped to one set of 4 × 4 station-day co-clusters and in all there were also four sets associated to 'cold', 'cool', 'warm' and 'hot' years.

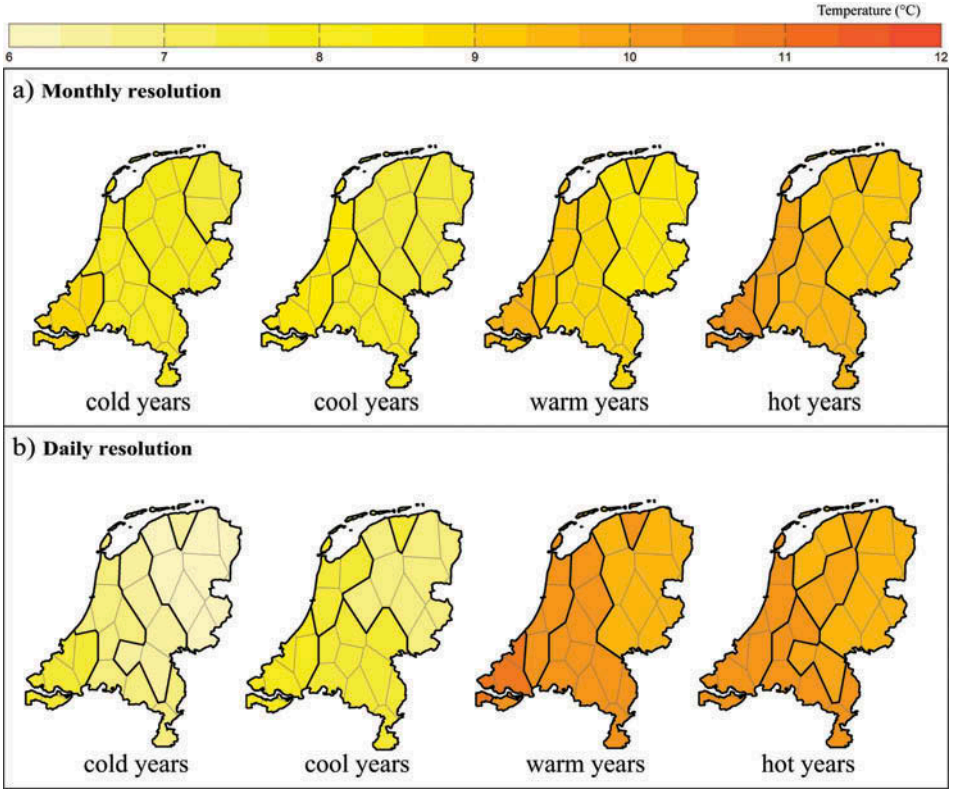

Figure 7. Small multiples to display spatial patterns for station-month and station-day co-clusters. (a) Small multiples display spatial patterns in station-month co-clusters for year-clusters. (b) small multiples from display spatial patterns in station-day co-clusters for year-clusters. Light yellow indicates low temperature and red means high temperature.

Figure 7 displays the spatial patterns via small multiples for the four sets of station-month co-clusters (Figure 7(a)) and four sets of station-day co-clusters (Figure 7(b)) and with the same colour scheme used for the yearly resolution (Figure 6).

In Figure 7(a) each map displays one set of station-month co-clusters and each region represents the spatial distribution of each station-cluster in that set. The value of each region is the average temperature in each station-cluster along all month-clusters. Figure 7(a) shows the decreasing temperature patterns from southwest to northeast for each year-cluster.

In Figure 7(b) each map displays one set of station-day co-clusters and each region represents the spatial distribution of each station-cluster in that set. The value of each region is the average temperature in each station-cluster along all day-clusters. Figure 7(b) shows the same decreasing temperature patterns from southwest to northeast. It can be noticed that the temperature in 'warm' and 'hot' years are similar. We suppose this is because the representative daily data for 'hot' years have more variations than those for 'warm' years.

Figure 8 shows the temporal patterns using the ringmap for the four sets of station-month co-clusters (inner four circles) and the four sets of station-day co-clusters (outer four circles). The colour scheme of this figure extends the one used previously with blue

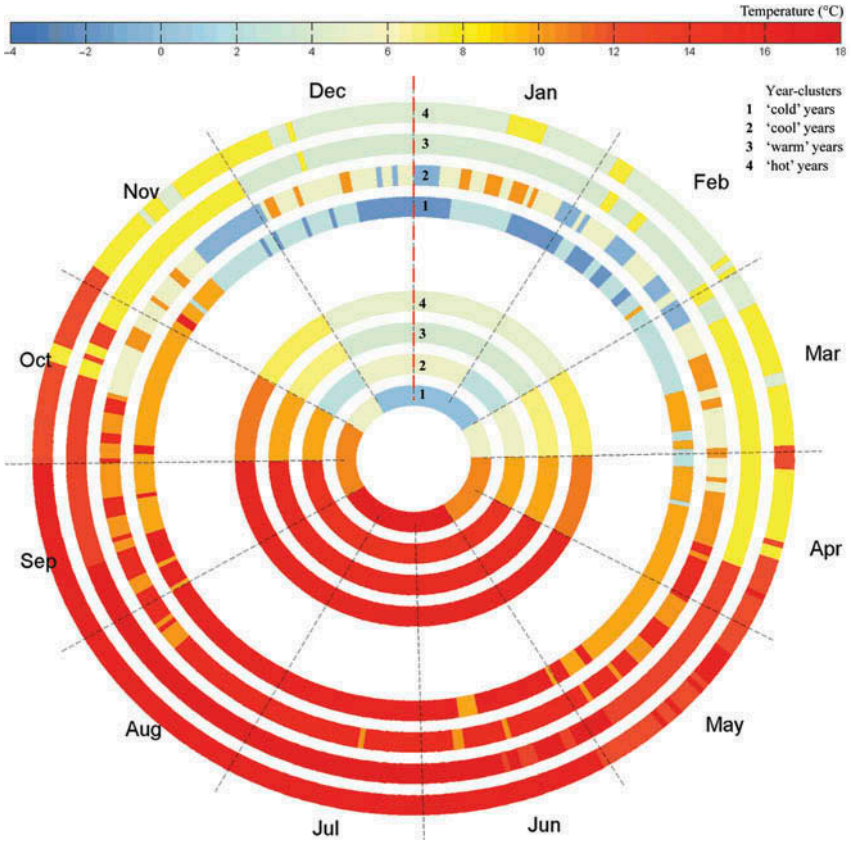

Figure 8. Ringmap to show temporal patterns for station-month co-clusters and station-day co-clusters. The inner four circles display temporal patterns for the four sets of station-month co-clusters for year-clusters. The outer four circles displays temporal patterns for the four sets of station-day co-clusters for year-clusters. Blue means very low temperature; light yellow indicates low temperature and red means high temperature.

meaning very low temperature, light yellow indicating low temperature and red meaning high temperature.

The inner four circles show the temporal variations in the four sets of station-month co-clusters for 'cold', 'cool', 'warm' and 'hot' years from inside out. Each circle contains 12 months clockwise from January to December, and the colour of each month indicates the month-cluster it belongs to. The value of each month-cluster is the average temperature in each month-cluster. The four month-clusters are named relative 'cold', 'cool', 'warm' and 'hot' months in the order of increasing temperature values. The term 'relative' is used because the values of 'cold'/'cool'/'warm'/'hot' months at different circles are not the same. There is an increasing temperature pattern across the four circles from inside out. Combining the ringmap with the small multiples in Figure 7(a), we can see that the northeast region in January, February and December in 'cold' years has the lowest temperature values while the southwest region from May to September in 'hot' years has the highest temperature values. The temperature values at other regions in other months are between the two extremes.

The outer four circles in Figure 8 show the temporal patterns in the four sets of station-day co-clusters for 'cold', 'cool', 'warm' and 'hot' years from inside out. Each circle contains 365 days clockwise from 1 January to 30 December, and the colour of each day indicates the day-cluster it belongs to. The value of each day-cluster is the average temperature in each day-cluster. Similar to the inner circles, the four day-clusters are named as relative 'cold', 'cool', 'warm' and 'hot' days. The outer four circles also show an increasing temperature pattern. The circles for 'warm' and 'hot' years have less shifts among day-clusters than the circle for 'cold' years. The circle for 'cool' year has the most shifts. From these circles we can further notice several anomalous days, for instance, 'hot' days in early November for 'cold' years, 'warm' days in January and December for 'cool' year and 'hot' days at the end of March for 'hot' years. Those anomalous days might be of interest for other disciplines such as climatology or phenology. Besides, with the combination of the small multiples for station-day co-clusters in Figure 7(b), we can observe that the northeast region in January and late December in 'cold' years has the lowest temperature values while the southwest region in most days from June to August in 'hot' and 'warm' years has the highest values.

Finally, the spatio-temporal patterns at monthly and daily resolutions were compared to examine the inconsistencies caused by the MTUP. From a spatial point of view, the two sets of small multiples illustrated in Figure7(a) and 7(b) allow to straightforwardly view the differences of station-month and station-day regions. Results show that the coarser the temporal resolution, the less variations in the temperature of the clusters. That is because the averaging of temperature from daily to monthly results in the loss of details in the data. Also most of the compositions in regions changed at the two resolutions for all year-clusters. That is reasonable because patterns at daily resolution reveal the variations among daily temperature data only and so do the patterns at monthly. Two stations with the same yearly average temperature might exhibit very different variability when studied at finer temporal resolutions. One exception was observed for 'warm' years where the composition is exactly the same at monthly and daily resolutions. We suggest that is because in 'warm' years the representative daily temperature have little within-month variations, and therefore spatial patterns remain the same at monthly resolution. Such suggestion was confirmed by observing the circle in 'warm' years for daily resolution in Figure 8. With respect to the temporal patterns, the inner four and outer four circles of Figure 8 explicitly illustrate the differences between station-month and station-day co-clusters. The temporal patterns in station-day co-clusters are more sophisticated than those in station-month co-clusters and provide more information.

5. Conclusions

In this study we introduced the use of the BBAC_I to analyse GTS. In contrast to common clustering methods currently used by the geo-community, the BBAC_I allows the simultaneous study of spatial and temporal patterns. By mapping locations to location-clusters and timestamps to timestamp-clusters, this newly introduced algorithm enables the identification of co-clusters which contain elements that are similar along the spatial (location) and temporal (timestamp) dimensions. Thus, the co-clustering results are able to capture the concurrent space-time varying behaviour present in GTS. In addition, we used three geovisualization techniques (heatmap, small multiples and ringmaps) to facilitate the exploration and comparison of the co-clustering results at different temporal resolutions.

The BBAC_I and associated geovisualization techniques were used to study patterns in 20 years of temperature data collected over 28 stations in the Netherlands. The analysis

was done at the yearly, monthly and daily temporal resolutions to study the MTUP. Results showed that BBAC_I can be used to effectively point out regions as well as subsets of years/months/days that have similar temperature values. It was shown that there is the increasing temperature pattern from northeast to southwest and from 'cold' to 'hot' years/months/days with only three years belonging to 'cool' or 'cold' years. Regarding the MTUP, the finer the temporal resolution is, the more complex the patterns become as they are based on.

Since this is the first time that the BBAC_I is used to analyse GTS, we anticipate that further work in the following areas will be required: (1) presently BBAC_I only produces rectangular co-clusters. This may not be enough to explore all spatio-temporal pattern in GTS. For example in Figure 4, the co-cluster of station-cluster4/year-cluster2 is similar with the co-cluster of station-cluster3/year-cluster1. This indicates that the number of station-clusters and of timestamps-clusters should be optimized in future studies. For example, mean squared residue (MSR) (Cheng and Church 2000, Zhou and Ashfaq 2006) is an index that can be used to evaluate clustering results and thereby optimize cluster numbers. (2) The current BBAC_I algorithm can only deal with a single numeric value for each element in the data matrix and consequently analyses only one observed property – yearly, monthly and daily averaged temperatures in this study. Future work is needed to adapt the BBAC_I to allow the co-clustering of more variables. (3) The BBAC_I has only been tested in an area that exhibits relatively small range of temperature values. Future work should deal with the co-clustering in areas with larger variability.

Acknowledgements

The authors would like to thank David G Rossiter (Cornell University) and Emma Izquierdo Verdiguier (University of Twente) for their insightful comments about this work.

References

Anagnostopoulos, A., Dasgupta, A., and Kumar, R., 2008. Approximation algorithms for co-clustering. *In*: *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems*. New York: ACM, 201–210.

Andrienko, G., *et al*. 2009. Interactive visual clustering of large collections of trajectories. *In*: *IEEE symposium on visual analytics science and technology (VAST)*, 12–13 October, Atlantic City, NJ. New York: IEEE, 3–10.

Andrienko, G., *et al*., 2010a. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29 (3), 913–922. doi:10.1111/j.1467-8659.2009.01664.x

Andrienko, G., *et al*., 2010b. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24 (10), 1577–1600. doi:10.1080/13658816.2010.508043

Banerjee, A., *et al*., 2007. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8, 1919–1986.

Cheng, Y. and Church, G.M., 2000. Biclustering of expression data. *Proceedings ISMB*, 93–103.

Cho, H., *et al*., 2004. Minimum sum-squared residue co-clustering of gene expression data. *In*: *Fourth SIAM international conference on data mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 114–125.

Coltekin, A., *et al*., 2011. Modifiable temporal unit problem. *In: ISPRS/ICA workshop "Persistent problems in geographic visualization" (ICC2011)*, Paris. Berlin: Springer-Verlag.

Crane, R.G. and Hewitson, B.C., 2003. Clustering and upscaling of station precipitation records to regional patterns using self-organizing maps (SOMs). *Climate Research*, 25, 95–107. doi:10.3354/cr025095

Deng, M., *et al*., 2011. A general method of spatio-temporal clustering analysis. *Science China Information Sciences*, 56 (10), 1–14. doi:10.1007/s11432-011-4391-8

Dhillon, I.S., Mallela, S., and Modha, D.S., 2003. Information-theoretic co-clustering. *In*: *The 9th international conference on knowledge discovery and data mining (KDD)*, 24–27 August, Washington, DC. New York: ACM, 89–98.

Dykes, J., MacEachren, A.M., and Kraak, M.-J., 2005. *Exploring geovisualization*. Amsterdam: Elsevier.

Guo, D., *et al*., 2006. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12 (6), 1461–1474. doi:10.1109/TVCG.2006.84

Hagenauer, J. and Helbich, M., 2013. Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science*, 27 (10), 2026–2042. doi:10.1080/13658816.2013.788249

Hagerstand, T., 1970. What about people in regional science. *Papers of the Regional Science Association*, 14, 7–21.

Han, J., Kamber, M., and Pei, J., 2012. *Data mining concepts and techniques*. 3rd ed. Waltham, MA: Morgan Kaufman, MIT Press.

Han, J., Lee, J.-G., and Kamber, M., 2009. An overview of clustering methods in geographic data analysis. *In*: H.J. Miller and J. Han, eds. *Geographic data mining and knowledge discovery*. 2nd ed. New York: Taylor & Francis Group, 150–187.

Jong, R.D. and Bruin, S.D., 2012. Linear trends in seasonal vegetation time series and the modifiable temporal unit problem. *Biogeosciences*, 9, 71–77. doi:10.5194/bg-9-71-2012

Keim, D.A. and Kriegel, H.-P., 1996. Visualization techniques for mining large databases: a comparison. *IEEE Transcations on Knowledge and Data Engineering*, 8 (6), 923–938.

Kisilevich, S., Mansmann, F., and Nanni, M., 2010. Spatio-temporal clustering. *In*: *Data mining and knowledge discovery handbook*. 2nd ed. New York: Springer Press, 855–874.

Li, S.-T. and Chou, S.-W., 2000. Multi-resolution spatio-temporal data mining for the study of air pollutant regionalization. *In*: *The 33rd Hawaii international conference on system sciences*, Maui, HI. Los Alamitos, CA: IEEE Computer Society, p. 7.

Li, X., *et al*., 2012. Explore multivariable spatio-temporal data with the time wave: case study on meteorological data. *In*: A.G.O. Yeh *et al.*, ed. *Advances in spatial data handling and GIS. Lecture notes in geoinformation and cartography, part 3*. Berlin: Springer, 79–92. doi:10.1007/978-3-642-25926-5_7

Maceachren, A., *et al*., 2003. Exploring high-D spaces with multiform matrices and small multiples. *In*: *IEEE Symposium on Information Visualization, INFOVIS 2003*. New York: IEEE, 31–38.

Miller, H.J. and Han, J., 2009. Geographic data mining and knowledge discovery: an overview. *In*: H.J. Miller and J. Han, eds. *Geographic data mining and knowledge discovery*. 2nd ed. London: Taylor & Francis Group, 1–26.

Wu, T., *et al*., 2008. Mining geographic episode association patterns of abnormal events in global earth science data. *Science in China Series E: Technological Sciences*, 51 (S1), 155–164. doi:10.1007/s11431-008-5008-3

Wu, X., Zurita-Milla, R., and Kraak, M.-J., 2013. Visual discovery of synchronisation in weather data at multiple temporal resolutions. *The Cartographic Journal*, 50 (3), 247–256. doi:10.1179/1743277413Y.0000000067

Zhang, P., *et al*., 2003. Correlation analysis of spatial time series datasets: a filter-and-refine approach. *In*: K.-Y. Whang *et al*., eds. *Advances in knowledge discovery and data mining*. Berlin: Springer, 532–544.

Zhao, J., Forer, P., and Harvey, A.S., 2008. Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization*, 7 (3–4), 198–209. doi:10.1057/palgrave.ivs.9500184

Zhou, J. and Ashfaq, K., 2006. ParRescue: scalable parallel algorithm and implementation for biclustering over large distributed datasets. *In*: *26th IEEE international conference on distributed computing system (ICDCS 2006)*. New York: IEEE, p. 21.

Zurita-Milla, R., *et al*., 2013. Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 51 (4), 1914–1921. doi:10.1109/TGRS.2012.2223218