Triclustering Georeferenced Time Series for Analyzing Patterns of Intra-Annual Variability in Temperature

Xiaojing Wu,* Raul Zurita-Milla,† Emma Izquierdo Verdiguier,† and Menno-Jan Kraak †

*Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, and Humanities, Arts and Social Sciences, Singapore University of Technology and Design
†Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente

Clustering is often used to explore patterns in georeferenced time series (GTS). Most clustering studies, however, only analyze GTS from one or two dimension(s) and are not capable of the simultaneous analysis of the data from three dimensions: spatial, temporal, and any third (e.g., attribute) dimension. Here we develop a novel clustering algorithm called the Bregman cuboid average triclustering algorithm with I-divergence (BCAT_I), which enables the complete partitional analysis of 3D GTS. BCAT_I simultaneously groups the data along its dimensions to form regular triclusters. These triclusters are subsequently refined using *k* means to fully capture spatiotemporal patterns in the data. By applying BCAT_I to time series of daily average temperature in The Netherlands (twenty-eight weather stations from 1992 to 2011), we identified the refined triclusters with similar temperature values along the spatial dimension (weather stations that represent locations) and two nested temporal dimensions (year and day). Geovisualization techniques were then used to display the patterns of intra-annual variability in temperature. Our results show that in the last two thirds of the study period, there is an intense variability of spring and winter temperatures in the northeast and center of The Netherlands. For the same period, an intense variability of spring temperatures is also visible in the southeast of the country. Our results also show that summer temperatures are homogenous across the country for most of the study period. This particular application demonstrates that BCAT_I enables a complete analysis of 3D GTS and, as such, it contributes to a better understanding of complex patterns in spatiotemporal data. *Key Words: data mining, geovisualization, georeferenced time series, intraannual variability, triclustering.*

聚类经常被用来探讨标示地理坐标的时间序列 (GTS) 模式。多半的聚类研究, 却仅从单维或二维分析 GTS, 且无法从时间、空间与任何第三面向 (例如属性), 对数据进行三维共时分析。我们于此建立一个名叫布雷格曼立方平均三角聚类化演算、并具有I散度的崭新聚类演算法 (BCAT_I), 该演算法得以对三维 GTS 进行完整的分隔分析。BCAT_I 同时依照数据的维度进行分群, 已形成规律的三角聚类。这些三角聚类接着运用 k 平均算法进行精炼, 以全面捕捉数据中的时空模式。我们透过将 BCAT_I 应用至荷兰每日平均温度的时间序列 (1992 年至 2011 年的二十八座气象站), 随着空间面向 (呈现地点的气象站) 和两个套迭的时间面向 (年与日) 指认具有相似温度值的精炼后之三角聚类。我们接着运用地理可视化技术, 展现温度的年度变化模式。我们的研究结果显示, 研究时期的后三分之二期间, 荷兰东北部与中部的春季与冬季温度呈现剧烈的变异。在同一期间, 荷兰的东南方亦能见到春季温度的剧烈差异。我们的研究同时显示, 在大部份的研究时期中, 全国的夏季气温是均值的。此一特殊的应用, 证实 BCAT_I 能够对三维 GTS 进行完整分析, 从而对更佳理解时空数据中的复杂模式做出贡献。 *关键词： 数据挖掘, 地理可视化, 标示地理坐标的时间序列, 年度变异, 三角聚类化。*

A menudo el agrupamiento se usa para explorar patrones en series de tiempo georreferenciadas (GTS). No obstante, la mayoría de los estudios sobre agrupamiento solamente analizan las GTS desde una o dos dimensiones, y no son capaces del análisis simultáneo de datos desde tres dimensiones: la espacial, la temporal y cualquier tercera dimensión (por ejemplo, atributo). Aquí, nosotros desarrollamos un novedoso algoritmo de agrupamiento denominado algoritmo triagrupador cuboide promedio de Bregman con I–divergencia (BCAT_I), que permite el análisis particional completo de GTS en 3D. El BCAT_I agrupa simultáneamente los datos a lo largo de sus dimensiones para formar triagrupamientos regulares. Estos triagrupamientos se refinan subsiguientemente usando medios *k* para captar completamente en los datos los patrones espaciotemporales. Aplicando BCAT_I a las series de tiempo de la media de temperatura diaria en los Países Bajos (veintiocho estaciones meteorológicas de 1922 a 2011), identificamos los triagrupamientos refinados con valor de temperatura similares a lo largo de la dimensión espacial (estaciones meteorológicas que representan localizaciones)

Annals of the American Association of Geographers, 108(1) 2018, pp. 71–87 © 2018 by American Association of Geographers
Initial submission, November 2016; revised submission, January 2017; final acceptance, February 2017
Published by Taylor & Francis, LLC.

y dos dimensiones temporales anidadas (año y día). Luego se usaron técnicas de geovisualización para desplegar los patrones de la variabilidad intra-anual de la temperatura. Nuestros resultados indican que en los últimos dos tercios del período de estudio se presenta intensa variabilidad en las temperaturas de primavera e invierno en el nordeste y centro de los Países Bajos. Para el mismo período, también es visible una intensa variabilidad en las temperaturas de primavera en el sudeste del país. También muestran nuestros resultados que las temperaturas de verano son homogéneas a través del país durante la mayor parte del periodo de estudio. Esta particular aplicación demuestra que el BCAT_I permite un análisis completo de GTS en 3D y de por sí, contribuye a un mejor entendimiento de los complejos patrones en los datos espaciotemporales. *Palabras clave: extracción de datos, geovisualización, series de tiempo georreferenciadas, variabilidad intra-anual, triagrupamiento.*

Georeferenced time series (GTS) describe the time-evolving behavior of one or more attributes that are typically recorded at fixed locations and uniform time intervals (e.g., number of infected patients per administrative unit and month or daily temperature data collected by a network of meteorological stations). GTS are one type of spatiotemporal data and, as such, they "live" in the n-dimensional space formed by their spatial, temporal, and (multi)-attribute dimensions (Guo et al. 2006). In this study, we focus on the analysis of 3D GTS with one spatial, one temporal, and any third (e.g., attribute) dimension. Such data are naturally modeled and viewed as a data cuboid (Harinarayan, Rajaraman, and Ullman 1996; Han, Kamber, and Pei 2011), in which each cell stores the value of each attribute observed at one location in one time stamp. The analysis of such data cuboids reveals the patterns along all three dimensions. In particular, we present a novel triclustering algorithm that allows the analysis of this kind of GTS.

Clustering is an important task in geospatial analysis because it facilitates the extraction of patterns from large and complex data sets by assigning similar data elements to the same group (G. Andrienko et al. 2009). By this means, clustering provides an overview of the data distribution at a higher level of abstraction and also allows the extraction of insights by focusing on particular groups or clusters. Many studies have used one-way clustering to analyze patterns in the data sets (e.g., G. Andrienko et al. 2010; Hagenauer and Helbich 2013; Helbich et al. 2013; Grubesic, Wei, and Murray 2014). In these studies, the authors group data elements along a single data dimension (e.g., space) based on similar values along the other dimension. Recently, Wu, Zurita-Milla, and Kraak (2015, 2016) used coclustering to perform 2D clustering for pattern analysis in GTS. Coclustering, in this context, means to simultaneously group data elements along their spatial and temporal dimensions. Neither one-way clustering nor coclustering, however, is capable of analyzing 3D GTS; that is, to group the data elements along spatial, temporal, and the third dimensions. Hence we focus on the use of triclustering for such analysis.

Triclustering algorithms, which group data elements based on their similarity along three dimensions, have already been applied in other fields. For instance, Zhao and Zaki (2005) developed a triclustering algorithm called TRICLUSTER that identifies gene–sample–time clusters in 3D microarray data sets; Ji, Tan, and Tung (2006) proposed the CubeMiner algorithm to mine frequent cooccurrences of gene–sample–time in 3D microarrays, too; and Sim, Aung, and Gopalkrishnan (2010) presented a triclustering algorithm called MIC to mine correlated 3D subspace clusters from financial data sets. CubeMiner is only applicable to binary data sets, however. Even though TRICLUSTER and MIC can be applied to real-value data sets, they aim at searching for significant clusters, which are usually of small amounts and only represent the intrinsically outstanding information in the data set depending on specific tasks and clustering methods, for example, high values over a certain threshold (Sim et al. 2013). Instead of only significant ones, we aim at exhaustively identifying all clusters, which are expected to provide more information in the data set. In this situation, none of these existing triclustering algorithms are able to perform such analysis of 3D GTS, which requires the complete partition of the data set. The issue necessitates a new triclustering algorithm that is capable of identifying all clusters in 3D GTS. To this end, here we expand the previous works on coclustering (Wu, Zurita-Milla, and Kraak 2015, 2016) and present a triclustering algorithm specifically designed to perform a complete partitional analysis of GTS that fit in data cuboids.

The main objective of this study is therefore to develop a triclustering algorithm that allows the complete partitional analysis of GTS along its three dimensions at the same time. The possibilities of this algorithm are demonstrated by analyzing a GTS of daily temperatures. Such data naturally fit into a cuboid where each cell contains a temperature value

indexed by its location and time stamps (year and day) of measurement. This application of triclustering paves the way toward the analysis of spatiotemporal patterns of intra-annual variability in temperature records, thereby supporting the study of the ecological impacts of climate change (Walther et al. 2002).

Method

Following the principle of "divide and group" in exploratory data analysis (N. Andrienko and Andrienko 2006; Feng, Wang, and Chen 2014), our proposed approach contains two parts: the triclustering algorithm developed for dividing the whole data set and then k means used for regrouping the triclusters to refine them.

Bregman Cuboid Average Triclustering Algorithm with I-Divergence

This section describes the development of the Bregman cuboid average triclustering algorithm with I-divergence (BCAT_I) as a similarity metric. Without losing generality, the description is guided by a GTS of daily temperature data collected at m stations for n years so that the algorithm becomes less abstract.

The BCAT_I algorithm is an extension of Bregman block average coclustering algorithm with I-divergence (BBAC_I) used by Wu, Zurita-Milla, and Kraak (2015, 2016). BCAT_I enables the simultaneous clustering of the elements along all dimensions of a data cuboid filled with positive real-value data. Such a data cuboid can be regarded as cooccurrences among three random variables: the stations (S), the years (Y), and the days of the year (D). In this setup, the rows of the data cuboid refer to the stations that represent the fixed locations, the columns to the years, and the depths to the 365 days that belong to each year (for convenience, 29 February is removed in leap years for equal lengths). The elements of this cuboid are the daily average temperatures for each station, year, and day (Figure 1A). To assure having values equal to or larger than zero, the absolute value of the minimum temperature was added to all temperatures in the data set.

By concurrently grouping stations to station cluster, years to year clusters, and days to day clusters, the triclustering algorithm seeks triclusters that contain

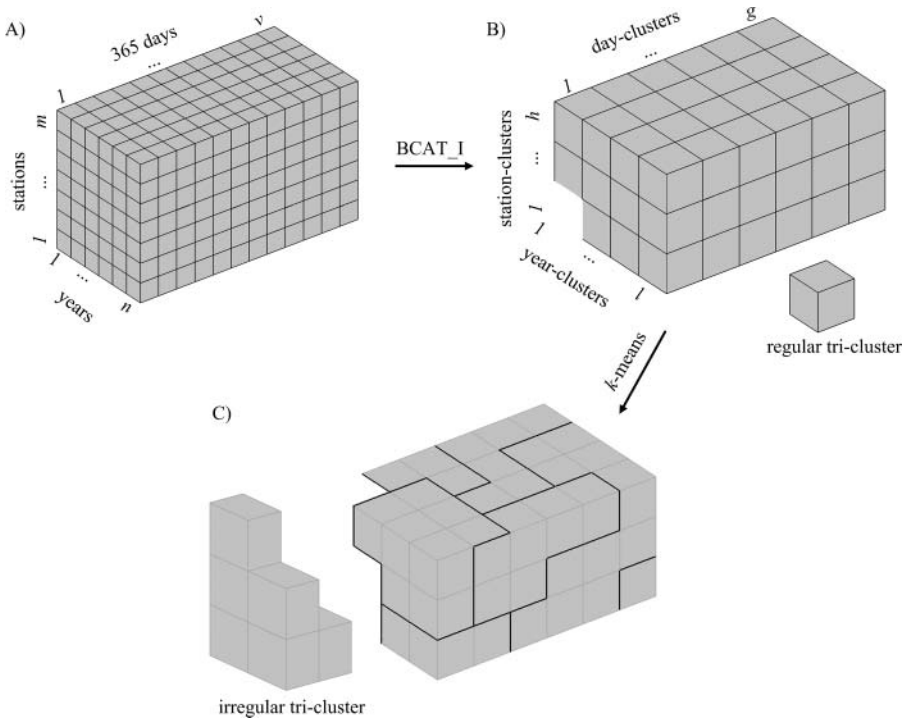

Figure 1. (A) The data cuboid with size $m \times n \times v$. The rows refer to stations, the columns to years, and depths to 365 days. (B) $h \times l \times g$ regular triclusters (subcuboid) after applying BCAT_I to the data cuboid. The rows refer to station clusters, the columns to year clusters, and depths to day clusters. (C) k irregular triclusters refined from regular ones with k means. The axes arrangement is the same as that for regular triclusters. BCAT_1 = Bregman cuboid average triclustering algorithm with I-divergence.

similar temperature values along the three dimensions of the input data cuboid. These triclusters are defined by the intersection of station, year, and day clusters (Figure 1B). Like the BBAC_I, the composition of the triclusters is optimized by using the I-divergence metric, the superiority of which over other metrics (e.g., Euclidean distance) has been empirically proved (Banerjee et al. 2007). As such, the triclustering problem can be regarded as an optimization problem where the optimal results minimize the loss of mutual information between the original data cuboid and the triclustered one.

Figure 2 shows the pseudocode of BCAT_I: The data cuboid containing the temperature values is represented by $\mathbf{O} \in \mathbb{R}^{m \times n \times v}$ where m represents the number of stations (S), n represents the number of years (Y), and v represents the number of days per year (D). The numbers of station, year, and day clusters, which are defined by the user as inputs, are represented by h, l, and g, respectively. BCAT_I starts by randomly initializing three binary matrices $\mathbf{R} \in \mathbb{R}^{m \times h}$, $\mathbf{C} \in \mathbb{R}^{n \times l}$, and $\mathbf{T} \in \mathbb{R}^{v \times g}$ that indicate the membership to clusters of each dimension. Next, an iterative process to optimize these memberships starts by updating $\mathbf{R}$, the station clustering. For this, the original data cuboid $\mathbf{O}$ is first reshaped to a matrix $O' \in \mathbb{R}^{m \times vn}$, of which the rows are the m stations and the columns are $v \times n$ days/years. This allows the definition of $\widehat{O'}$, which is calculated as the average of elements of O' that belong to the same cluster according to the current mapping. Then an approximate matrix of O', named $\mathbf{A} \in \mathbb{R}^{m \times vn}$, is created by expanding $\widehat{O'}$ based on the mapping and the values of the same cluster to be preserved. After that, the loss of the mutual information between O' and $\mathbf{A}$ is calculated. By minimizing the information loss, the optimal mapping from stations to station clusters is produced and $\mathbf{R}$ is updated. The optimization proceeds with the year and day clustering to update $\mathbf{C}$ and $\mathbf{T}$. Once the information loss is minimal, the update of the $\mathbf{R}$, $\mathbf{C}$, and $\mathbf{T}$ matrices stops and this yields the optimal triclustering results (the Appendix contains a detailed explanation of our triclustering algorithm). Finally, the original data cuboid $\mathbf{O}$ is reordered following the optimized binary

Algorithm 1 Tri-clustering algorithm

Require: $\mathbf{O} \in \mathbb{R}^{m \times n \times v}$: original data, h: num. of rows clusters, ℓ: num. of columns clusters, g: num. of vector clusters,
Ensure: $\mathbf{R}^* \in \mathbb{R}^{m \times h}$, $\mathbf{C}^* \in \mathbb{R}^{n \times \ell}$ and $\mathbf{T}^* \in \mathbb{R}^{v \times g}$
  Random initialization of $\mathbf{R}$, $\mathbf{C}$, $\mathbf{T}$
  $\mathbf{T1} \in \mathbb{R}^{nv \times g} \leftarrow$ vertically concatenate of $\mathbf{T}$ n times
  while untill the convergence **do**
    Updated row clustering:
    $\mathbf{O}' \in \mathbb{R}^{m \times nv} \leftarrow$ reshape $\mathbf{O}$
    $\mathbf{A} \leftarrow \mathbf{R}(\mathbf{R}^\top \mathbf{O}'\mathbf{T1}/\mathbf{R}^\top \mathbb{1}\mathbf{T1})'\mathbf{T1}^\top$
    $D_{I\,i,\cdot} \in \mathbb{R}^h \leftarrow D_I(\mathbf{O}'(i,\cdot)||\mathbf{A}(i,\cdot))$
    $\mathbf{R}^* \leftarrow$ binary encoding of $(\arg\min_{j \in [1,h]}\{D_{I\,i,j}\})$
    Updated column clustering:
    $\mathbf{R1} \in \mathbb{R}^{mv \times h} \leftarrow$ vertically concatenate of $\mathbf{R}^*$ v times
    $\mathbf{O}' \in \mathbb{R}^{n \times mv} \leftarrow$ reshape $\mathbf{O}$
    $\mathbf{A} \leftarrow \mathbf{C}(\mathbf{C}^\top \mathbf{O}'\mathbf{R1}/\mathbf{C}^\top \mathbb{1}\mathbf{R1})'\mathbf{R1}^\top$
    $D_{I\,p,\cdot} \in \mathbb{R}^\ell \leftarrow D_I(\mathbf{O}'(p,\cdot)||\mathbf{A}(p,\cdot))$
    $\mathbf{C}^* \leftarrow$ binary encoding of $(\arg\min_{q \in [1,\ell]}\{D_{I\,p,q}\})$
    Updated depth clustering:
    $\mathbf{C1} \in \mathbb{R}^{mn \times \ell} \leftarrow$ vertically concatenate of $\mathbf{C}^*$ m times
    $\mathbf{O}' \in \mathbb{R}^{v \times mn} \leftarrow$ reshape $\mathbf{O}$
    $\mathbf{A} \leftarrow \mathbf{T}(\mathbf{T}^\top \mathbf{O}'\mathbf{C1}/\mathbf{T}^\top \mathbb{1}\mathbf{C1})'\mathbf{C1}^\top$
    $D_{I\,w,\cdot} \in \mathbb{R}^g \leftarrow D_I(\mathbf{O}'(w,\cdot)||\mathbf{A}(w,\cdot))$
    $\mathbf{T}^* \leftarrow$ binary encoding of $(\arg\min_{e \in [1,g]}\{D_{I\,w,e}\})$
    $\mathbf{T1} \in \mathbb{R}^{nv \times g} \leftarrow$ vertically concatenate of $\mathbf{T}^*$ n times
  end while

Figure 2. The pseudocode of Bregman cuboid average triclustering algorithm with I-divergence.

matrices **R**, **C**, and **T**, to group together the elements that belong to the same tricluster. For our particular application, this reordering is such that station, year, and day clusters are ordered from bottom to top of rows, from left to right of columns, and from front to back of depths with increasing average temperatures along other dimensions, respectively. This arrangement means that the identified triclusters have increasing temperature values from the bottom left front corner to the top right back corner of the reordered cuboid. To simplify the analysis and visualization of the triclusters, their values are set to the average of the elements that belong to each tricluster.

Refinement of BCAT_I Result

The BCAT_I divides the data cuboid into $h \times l \times g$ triclusters that have cubical shapes and therefore are called regular triclusters. However, the need to predefine the numbers of clusters might result in various regular triclusters still having similar values (Wu, Zurita-Milla, and Kraak 2015, 2016). This prevents the complete identification of representative clusters in the data set. To mitigate this issue and to better capture the patterns, the regular triclusters are regrouped into noncubical but axis-parallel triclusters, known as irregular triclusters (Figure 1C). Considering the purpose of the triclustering analysis (Han, Lee, and Kamber 2009) and the fact that optimal and stable results can be achieved by using multiple runs (Dhillon, Mallela, and Modha 2003), a partitional clustering method is preferred for regrouping to obtain the irregular triclusters. For the partitional clustering method, we suggest using k means because it is the simplest partitional clustering algorithm and, more important, it has been proven to produce satisfactory results when used for the regrouping process (Wu, Zurita-Milla, and Kraak 2016). We also suggest using the mean and variance of data elements belonging to each of the regular triclusters as inputs for k means. These two variables are used because the former provides a representative value of elements in each tricluster and the latter considers the presence of possible outliers within each regular tricluster. The number of irregular triclusters (k in k means) is optimized by using the Silhouette method (Rousseeuw 1987). The Silhouette method was evaluated by Lewis, Ackerman, and De Sa 2012) and it was found to produce the highly correlated clustering results with expert judgment. Finally, like with the regular triclusters, the values of the irregular triclusters are set to the value of the k means centroids.

Using BCAT_I to Explore Spatiotemporal Patterns of Intra-Annual Temperature Variability

Intra-annual variability in weather records is often studied together with changes in annual averages, especially in studies that deal with the impact of climate change on ecosystems (Williams and Hero 2001; Walther et al. 2002; Doi, Gordo, and Katano 2008; Williams and Middleton 2008). As stated by Doi, Gordo, and Katano (2008), patterns of annual averages of weather variables do not properly capture those of intra-annual variability, whereas the latter have a stronger impact on ecosystems than the former. This motivates our experiment, which was set up to apply the BCAT_I to Dutch daily temperature records. By grouping locations and years with similar within-year (i.e., days) temperature values, this triclustering analysis allows the exploration of the spatiotemporal patterns of intra-annual temperature variability in The Netherlands.

Data

In this study, we use Dutch daily average temperature data to illustrate the triclustering analysis. The location of The Netherlands in Europe (bottom right of Figure 3) determines the moderate maritime climate found in its west, especially in those areas close to the coastline. Such a climate is characterized by cool summers and mild winters because of the influence of the North Sea and the North Atlantic Ocean. The east of The Netherlands, especially those areas that border Belgium and Germany, exhibit a more continental climate with somewhat warmer summers and colder winters. As a result of this location, and despite the relatively small area of the country, the within-year variations of temperature in the west of the country are smaller than those in the southeast (Lenderink et al. 2011).

Specifically, the temperature data are collected at twenty-eight Dutch meteorological stations over twenty years (from 1 January 1992 to 31 December 2011). These temperature data and the geographical coordinates of each station were obtained from the

Figure 3. Study area: The Thiessen polygon map of the Dutch meteorological stations.

Royal Netherlands Meteorological Institute (KNMI; http://www.knmi.nl/). Using the stations' coordinates and the boundary of The Netherlands, we generated a Thiessen polygon map that defines the area influenced by each station (Figure 3). In this map, each polygon is labeled with both the station ID given by the KNMI (e.g., 290) and the name of the station (e.g., Twente). It is worth mentioning that the Thiessen polygons are only used to represent the clustering results and that other spatial units (e.g., administrative units or grid cells) are possible in other case studies.

Experiment Design

As discussed earlier, the Dutch temperature data were first organized in a data cuboid of 28 (stations) by 20 (years) by 365 (days). BCAT_I was then used to tricluster this cuboid. The number of clusters in each dimension needs to be predefined. As with other clustering methods, these BCAT_I parameters are fixed by the user or analyst after considering the application and the case study data set. Usually, an exploratory analysis is conducted to fix clustering parameters where various combinations are tried. In our case, such analysis was done by Wu, Zurita-Milla, and Kraak (2015), who set four for the number of both station and year clusters when applying the coclustering analysis to the annual averages of the same temperature data set. The same values were chosen for this experiment to allow the comparison of the clustering results. The number of day clusters was fixed to eight because this is the optimum value found by using k means and the Silhouette method to cluster a representative Dutch daily temperature profile (1*365) made by averaging all the temperature records. This means that the original data cuboid was clustered by BCAT_I into four (station clusters) by four (year clusters) by eight (day clusters) regular triclusters. Other parameters for implementing BCAT_I were also empirically set for this experiment: The number of iterations for optimization was set to 100, the number of random initialization was set to 200, and the threshold for minimal information loss was set to 10^{-5} to assure the stable triclustering results. The running time with these

parameters was about one minute using a regular laptop (i54200U CPU@1.60 GHz, 8 GB RAM) with MATLAB (version 2015a, MathWorks, Natick, MA, USA). However, it is important to note that this set of parameters is not universally applicable and empirical experiments should be done to find a suitable set of parameters for different data sets.

The $4 \times 4 \times 8$ regular triclusters were subsequently refined into k irregular triclusters. After that, results were displayed using several (geo)visualization techniques. Both 3D and 2D heat maps were used to visualize both the regular and irregular triclusters. One set of small multiples and two sets of timelines were used to show the composition of the regular triclusters (i.e., the distribution of station, year, and day clusters), and another set of small multiples was used to show the spatial patterns of intra-annual variability in temperature. To reveal such spatial patterns from irregular triclusters, for each year cluster we examined the values of day clusters along station clusters. For instance, some day clusters have the same value for all station clusters, which shows that the spatial pattern for these day clusters is that the whole of The Netherlands exhibits the same variability. Some other day clusters have different values along station clusters, indicating that the spatial pattern for these day clusters is that The Netherlands is divided into more than one region, each one exhibiting different variabilities by corresponding station cluster(s). Finally, another four sets of timelines were used to show the temporal patterns of temperature variability within four year clusters. To reveal such temporal patterns, for each year cluster we combined day clusters with the same spatial pattern and chronologically visualized them in one timeline. The values of the irregular triclusters to which these day clusters belong were used to define the colors used in the timelines. These timelines were arranged in alignment with geographic maps in the second set of small multiples that indicate corresponding spatial patterns.

Results

Regular and Irregular Triclusters

The application of BCAT_I to the Dutch daily temperature data set yielded 128 ($4 \times 4 \times 8$) regular triclusters. The 3D heat map (center) and four 2D heat maps (side subplots) in Figure 4 display these triclusters in the reordered data cuboid. In the 3D heat map, rows indicate station clusters, columns indicate year clusters, depths indicate day clusters, and their intersections (subcuboids) indicate regular triclusters. As an example, the subcuboid marked by the thick line in Figure 4 shows the regular tricluster (4, 3, 1). This tricluster is formed by the intersection of station

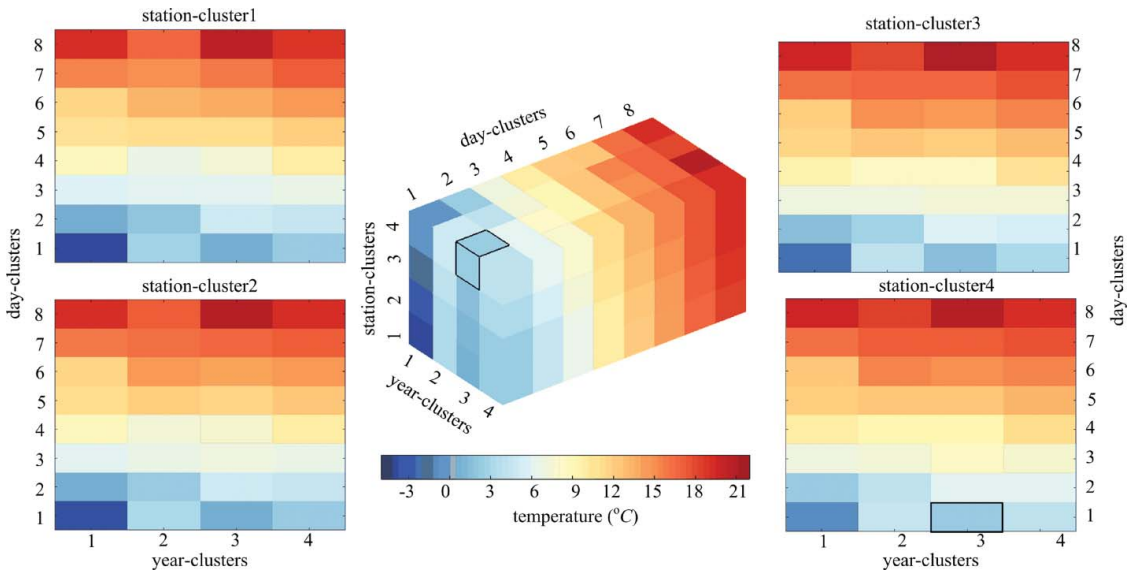

Figure 4. The resulting 128 ($4 \times 4 \times 8$) regular triclusters from BCAT_I in 3D heat map (middle) and 2D heat maps (side subplots) illustrating each station cluster. In the 3D heat map, the rows indicate station clusters, columns indicate year clusters, and depths indicate day clusters. In each 2D heat map, the x-axis indicates year clusters and the y-axis indicates day clusters. The regular tricluster (4, 3, 1), intersected by station cluster4, year cluster3 and day cluster1 is highlighted by the thin lines in both 3D and the 2D heat maps. BCAT_1 = Bregman cuboid average triclustering algorithm with I-divergence. (Color figure available online.)

cluster4, year cluster3, and day cluster1. The four 2D heat maps derived from the 3D one illustrate each of the station clusters. In each heat map, the *x*-axis indicates year clusters, the *y*-axis indicates day clusters, and their intersections (rectangles) indicate triclusters involving that station cluster. The rectangle marked by the thick line in the heat map of station cluster4 also shows the tricluster (4, 3, 1).

Both the 3D and 2D heat maps in Figure 4 show that many regular triclusters have similar temperature values. In addition, from heat maps we observe that the tricluster with the highest temperature, which is supposed to be (4, 4, 8) in this reordered data cuboid, is (4, 3, 8) instead. We suppose that is because the reorder is based on average values of whole station, year, and day clusters. This demands the refinement of these triclusters using *k* means.

After testing *k* values from four to thirty in steps of one, the Silhouette method identified twenty as the optimal number of irregular triclusters. These irregular triclusters are indexed and shown in Figure 5 using a 3D heat map (center) and four 2D heat maps (side subplots). The legend contains discrete values for each of twenty irregular triclusters, unlike the one in Figure 4 with continuous values to indicate representative temperatures assigned to regular triclusters. These discrete values of irregular triclusters suggest the usefulness of the refinement by *k* means.

The thick lines in the heat maps of Figure 5 show one example of an irregular tricluster (in this case number 5). By using the same axes arrangement in Figures 4 and 5, the composition of each irregular tricluster from regular ones can be observed. For example, the irregular tricluster number 5 is composed from the following regular triclusters: (1, 2, 1), (1, 2, 2), (2, 2, 1), (2, 2, 2), (3, 2, 1), (3, 2, 2), and (4, 3, 1). As such, this second clustering groups subcuboids with similar temperatures. This grouping completely identifies similar temperature values of the original data set along spatial, temporal, and day dimensions, which thus enables the full exploration of the complex patterns in the data.

Figure 6 shows the spatial distribution of station clusters using small multiples (Figure 6A) and the temporal distributions of year and day clusters using two sets of timelines (Figure 6B and Figure 6C, respectively). The small multiples show that temperature-wise, The Netherlands is divided into four regions: the northeast (station cluster1), the center-northeast (station cluster2), the center-southwest (station cluster3), and the southwest (station cluster4). Such a division confirms the fact that the within-year temperature variability in the west is different than that in the east of the country. This division is the same as that in the coclustering analysis by BBAC_I (Wu, Zurita-Milla, and Kraak 2015) that used annual averages of the same data set but different from that in the clustering analysis by Self-Organizing Map (SOM; Wu, Zurita-Milla, and Kraak 2013) using the same data set. We suppose that might be because the divided regions are only affected by the involvement of the temporal information in the

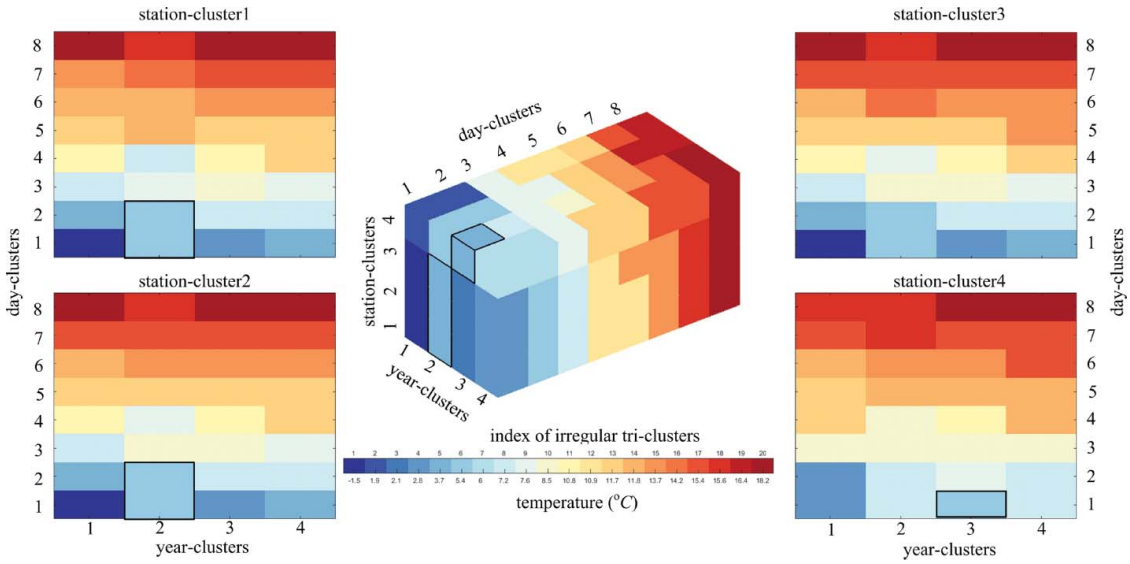

Figure 5. The resulting twenty irregular triclusters from *k* means in a 3D heat map (middle) and 2D heat maps (side subplots). The axes arrangement in all heat maps is the same as that in Figure 4. The same example of an irregular tricluster (number 5) is highlighted by the thick lines in both the 3D and the 2D heat maps. (Color figure available online.)

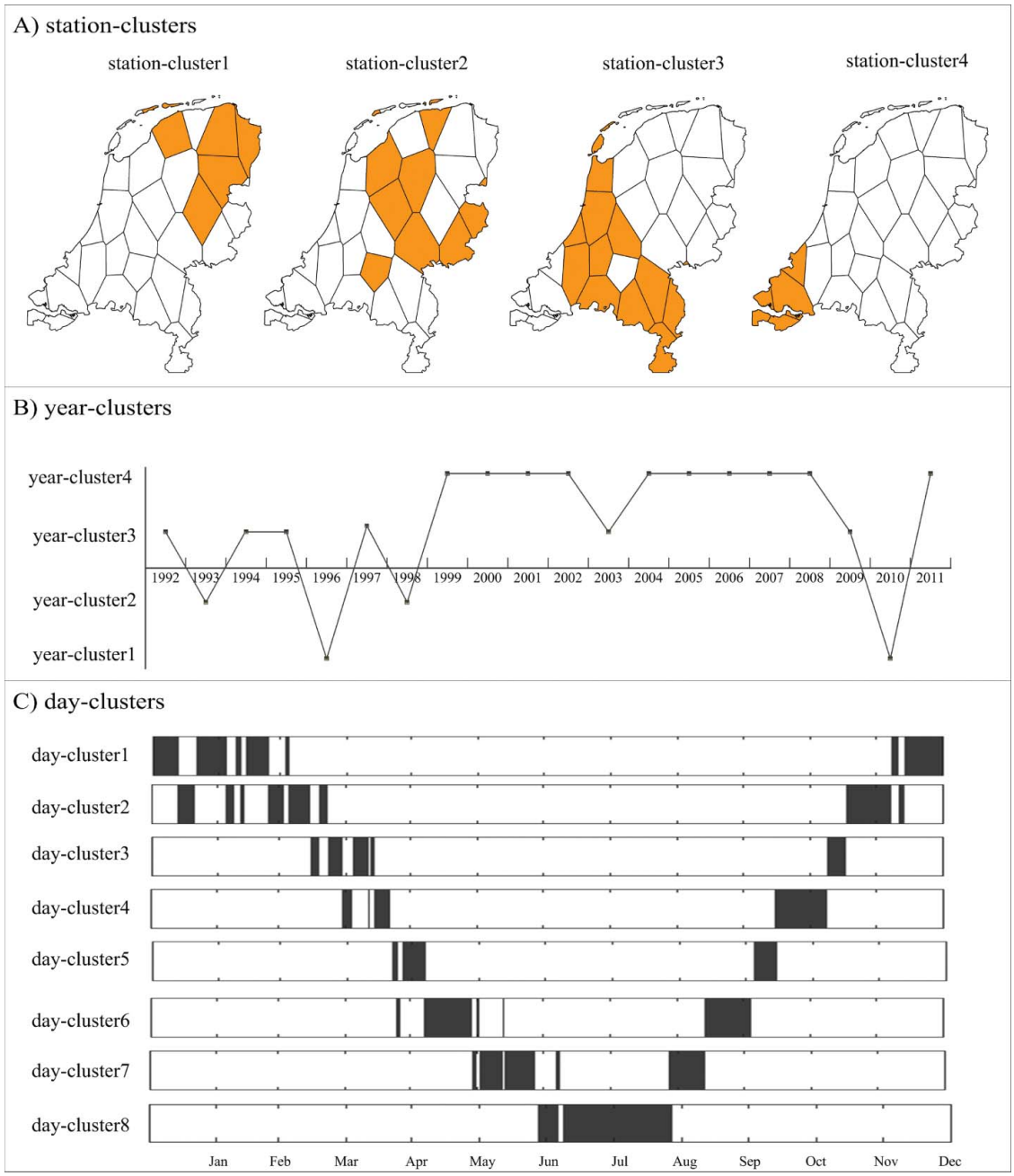

Figure 6. The composition of regular triclusters. (A) The small multiples to display the spatial distribution of station clusters 1 to 4. (B) A linear timeline to show temporal distribution of year clusters 1 to 4. (C) The timelines to show the temporal distribution of day clusters 1 to 8. (Color figure available online.)

data (i.e., year and day in triclustering analysis and year in coclustering analysis). The linear timeline of year clusters shows that 80 percent of the study period (sixteen out of twenty years), especially the period of 1999 to 2011, belongs to clusters with relatively high temperature values (i.e., year clusters 3 and 4). Only 10 percent of the study period (years 1996 and 2010) have very low temperature values, and the remaining 10 percent (years 1993 and 1998) have low temperature values. This distribution of year clusters over the study period is also supported by that in Wu, Zurita-Milla, and Kraak (2013, 2015). The timelines of day clusters show that a few clusters are compact in terms of the distribution of days assigned to them, for instance, day cluster8, which contains (quasi-) contiguous days from July to August. These timelines also show that other day clusters are loose. For instance, day cluster2 is formed by several winter and spring days. Compared

with those in other seasons, spring days are more loosely distributed in several day clusters. This indicates that temperature is much more changeable in the spring (Jaagus and Ahas 2000).

The combination of Figures 5 and 6 indicates that BCAT_I successfully identified regions and subsets of years and days that contain similar daily temperature values. It shows that the coldest temperatures occurred from the first days of December to the first days of March in 1996 and 2010 across the whole country, except in the southwest region. The warmest temperatures occurred in the summer period of all years except 1993 and 1998 across the whole country except the southwest region. The warmest temperatures were also experienced by the southwest region for most of the summers (i.e., not in 1993, 1996, 1998, and 2010). These results are supported by the findings in the coclustering analysis (Wu, Zurita-Milla, and Kraak 2015), except that more complete information about similar temperature values, especially from day to day, is discovered in this study.

Spatiotemporal Patterns of Intra-Annual Variability

Spatiotemporal patterns of intra-annual variability in Dutch daily average temperature from 1992 to 2011 were analyzed from the twenty irregular triclusters and displayed in the small multiples and timelines in Figure 7.

The small multiples in Figure 7A show the unique spatial patterns of intra-annual temperature variability. These spatial patterns were extracted by looking at the uniqueness of the patterns found in each year cluster. Take year cluster1, for example, as shown in the heat maps for irregular triclusters (Figure 5): The value of day clusters 1 to 5 and 8 is the same for station clusters 1 to 3 and different for station cluster4. This means that, for the days belonging to these day clusters in the year cluster1, station clusters 1 to 3 exhibit different temperature variability from station cluster4. Therefore, the spatial pattern for these days shows that The Netherlands is divided into two regions: northeast and center (station clusters 1 to 3) and southwest (station cluster4). As such, six spatial patterns were found after examining all day clusters in the four year clusters: (1) The Netherlands as one whole region (station clusters 1–4); (2) the northeast and center of the country as region1 (station clusters 1–3) and the southwest as region2 (station cluster4); (3) the northeast as region1 (station cluster1) and the center and southwest as region2 (station clusters 2–4); (4) the northeast as region1 (station cluster1), the center as region2 (station clusters 2 and 3), and the southwest

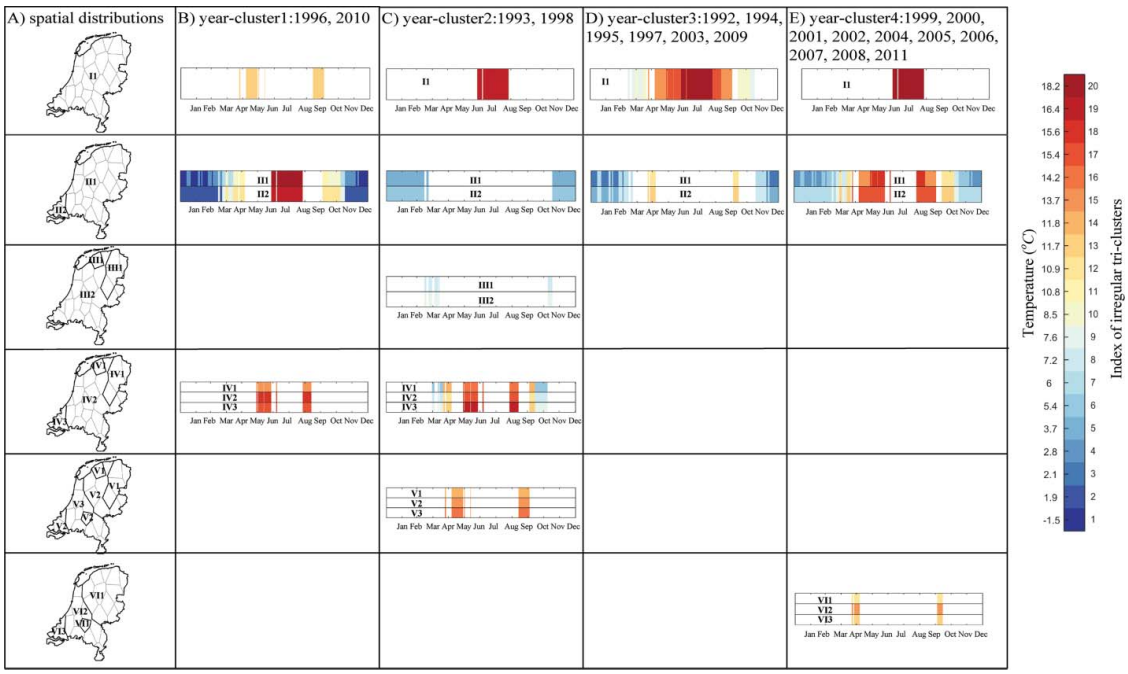

Figure 7. Spatiotemporal patterns of intra-annual variability in Dutch daily temperature from 1992 to 2011. (A) The small multiples show the unique spatial patterns of intra-annual variability. (B–E) The timelines show temporal patterns of temperature variability within each of the four year clusters. Each timeline is aligned with the corresponding spatial pattern. (Color figure available online.)

as region3 (station cluster4); (5) the northeast as region1 (station cluster1), the center-northeast and southwest as region2 (station clusters 2 and 4), and the center-southwest as region3 (station cluster3); and (6) the northeast and center-northeast as region1 (station clusters 1 and 2), the center-southwest as region2 (station cluster3), and southwest as region3 (station cluster4). These six spatial patterns were displayed in geographic maps in the small multiples (Figure 7A) from top to bottom.

In these spatial patterns, the northeast (station cluster1) and the southwest (station cluster4) of the country often exhibit different temperature variabilities from neighboring areas. We suppose that might be because these two areas are more directly influenced by continental and maritime climates, respectively. In addition, we observe that in the last two spatial patterns, station 356 (Herwijnen; see Figure 3) is isolated from the region it belongs to by another region with different variabilities. The possible reasons for this require further analysis with local environmental variables.

The timelines in Figures 7B to 7E show the temporal patterns of temperature variability within each of four year clusters. In terms of each year cluster, these temporal patterns were extracted from the irregular triclusters by combining day clusters with the same spatial patterns and visualizing them chronologically. The timelines are aligned with the corresponding spatial patterns (Figure 7A). As an example, let us discuss year cluster1 in detail. As already mentioned, day clusters 1 to 5 and day cluster8 have the same spatial pattern; thus, these day clusters are displayed in the timeline that is aligned with the corresponding spatial pattern: the northeast and center of the country as region1 and the southwest as region2 (i.e., the second spatial pattern). This timeline has two panels: The top one is for region1 and the bottom one is for region2. As such, the temporal patterns for all day clusters in four year clusters were extracted and displayed in timelines.

The combination of Figures 7A and 7B shows the spatiotemporal patterns of variability in temperature within year cluster1 (1996, 2010). In these two cold years, temperatures on most days exhibit different variabilities at the northeast and center of The Netherlands from those in the southwest. This could be because the continental climate is dominant in cold years and thus influences most of the country. It also shows that the northeast and center of the country experienced an intense variability in both winter and spring temperatures, whereas the southwest area only underwent such a variability in spring temperatures. Such changeable temperature in winter is worthy of special notice because such phenomena in spring and its effects are well known (Chmielewski and Rötzer 2001).

Figure 7C for year cluster2 (1993, 1998), together with Figure 7A, shows that these two years experienced the most complex temperature variability. Except for winter and the first half of summer, days in these two years are scattered with different spatial patterns, which indicates that temperatures in these days experienced intense variability across the whole of The Netherlands. Such a result is supported by the findings of Wu, Zurita-Milla, and Kraak (2015), who found that the days in 1993 had the most changeable temperatures. In addition, it is worth noting that the center-northeast and southwest of the country have the same temperature in May and September, which is lower than the center-southwest (the fifth row). This violates the general trend of the increasing temperature from the northeast to southwest of the country and needs further analysis.

As shown by the combination of Figures 7A and 7D, temperatures on most days from spring to autumn in year cluster3 underwent the same variability throughout The Netherlands. Temperatures on other days in the northeast and center and the southwest of the country experienced different but both mild variabilities. This could be because in years belonging to year cluster3, only one of maritime and continental climates completely influences the whole country on most days from spring to autumn, whereas the latter becomes dominant on other days.

The combination of Figures 7A and 7E shows the spatiotemporal patterns of temperature variability within year cluster4, in which most elements are recent and hot years. In these years, temperatures at the northeast and center of The Netherlands experienced more variabilities than the southwest for most days except summer.

Thus, The Netherlands has complex spatiotemporal patterns of intra-annual variability in temperature, despite its relatively small area. For most days in the whole study period, the variability in temperature defines two regions in the country: the northeast and center and the southwest. In both cold years (i.e., 1996, 2010) and hot years (i.e., years belonging to year cluster4), the northeast and center of the country experienced an intense variability in spring and winter temperatures, whereas the southwest only experienced such variability in spring temperatures. For most of the study period, summer temperatures are homogeneous across the whole country. This could be because in summer, either continental or maritime climate is

much more influential than the other, whereas in other seasons, especially the spring, both climates become influential.

Discussion

Unlike one-way and coclustering algorithms, BCAT_I is able to analyze 3D GTS by simultaneously grouping data elements along three dimensions. Like this, BCAT_I allows us to involve more information in the clustering process and produce more informative results. As displayed in Figure 4, the example of a regular tricluster contains one more dimension than the cocluster (Wu, Zurita-Milla, and Kraak 2015) in terms of days. By this means, BCAT_I allows the analysis of the original data along the day dimension without a loss of details.

Unlike previous triclustering algorithms, BCAT_I is able to provide the complete partitional analysis of 3D GTS, thereby revealing more information hidden in the data set. As displayed in Figure 4, both the 3D and the 2D heat maps clearly show that our newly developed triclustering algorithm exhaustively identifies all triclusters in the data set. This is an added value with respect to other triclustering algorithms that only focus on identifying a subset of clusters (Zhao and Zaki 2005; Sim, Aung, and Gopalkrishnan 2010). By this means, BCAT_I enables the extraction of complete information about temperature from low to high values in the data set.

Like any clustering algorithms, the preselection of the numbers of clusters is necessary for BCAT_I. In addition, BCAT_I only directly produces regular triclusters because it assigns complete rows, columns, and depths to corresponding clusters in the clustering process. These two aspects cause the potential issue that similar values exist in different triclusters, which always needs refinement to fully capture patterns in the data set.

Because there are no known labels in the case study data set, it is not easy to validate the patterns extracted by the clustering algorithms (Hagenauer and Helbich 2013). In this case, we think that it is reasonable to compare the extracted patterns using BCAT_I with the knowledge obtained from previous clustering works. To this end, we compared the detected patterns with those in Wu, Zurita-Milla, and Kraak (2013, 2015), who applied clustering analysis to the same data set.

Regarding the similarity metric used in the clustering algorithms, I-divergence is chosen for BCAT_I following the work of Wu, Zurita-Milla, and Kraak 2015). It is worth mentioning, however, that the choice of appropriate metric for other applications should depend on specific case studies if the triclustering analysis is to be used.

With respect to visualizing the triclustering results, it is worth pointing out that it would be more enlightening to have the interactive 3D heat maps and multiple linked views for these figures in the results. For instance, the 3D heat maps in Figures 4 and 5 can be rotated and edited as transparent to allow the highlighting of any selected regular or irregular tricluster. At the same time, the corresponding irregular or regular tricluster that makes it up in 3D and 2D heat maps and corresponding distributions in Figure 6 can be also highlighted.

Finally we elaborate on several challenges of this work that we have encountered or expect in future. The first relates to the optimization of our proposed approach. Because both BCAT_I and k means are local optimization clustering algorithms, multiple runs for each clustering algorithm are essential to maximize the likelihood of achieving a global minimum (i.e., an optimal and stable clustering result). The second relates to how BCAT_I is developed to analyze the cuboid by considering the three dimensions at the same time. As mentioned previously, the clustering procedure of BCAT_I is an iterative process to update the mapping of three dimensions. In each of three steps within one iteration, the cuboid is first reshaped into a matrix, the rows of which contain information about one dimension and the columns of which contain information about two others. By this means, the matrix keeps the information of all three dimensions and consequently, the update of mapping of each dimension in BCAT_I considers the mapping of two others. Finally, it is important to notice that the proposed triclustering approach might be computationally demanding when applied to large data sets because it first exhaustively identifies all of the clusters in the cuboid and then refines them. Hence, we recommend increasing the computational power by, for instance, using a powerful server or even migrating the approach to a cloud computing platform.

Conclusions

In this study, we present a newly developed triclustering algorithm named BCAT_I. This algorithm allows the analysis of GTS that fit into a data cuboid with one spatial, one temporal, and any third (e.g.,

attribute or nested spatial or temporal units) dimension. Unlike one-way clustering or coclustering, BCAT_I is capable of simultaneously clustering the data along all three dimensions of the data cuboid. Following the principle of "divide and group," the resulting triclusters are subsequently refined using k means to identify an optimal number of irregular triclusters that enables the full exploration of spatiotemporal patterns hidden in the 3D GTS.

In this article, BCAT_I was used to analyze time series of Dutch daily average temperatures collected from 1992 to 2011. In this particular application, the GTS has one spatial (weather stations that represent fixed locations) and two nested temporal dimensions (year and day), and the proposed triclustering analysis identified groups of stations and years that have similar within-year variability in average daily temperatures. Displayed using various geovisualization techniques, our results show that The Netherlands has six unique spatial patterns of intra-annual temperature variability associated with four groups of years. A detailed analysis of these patterns revealed that in most of the years from 1996, there is an intense variability in spring and winter temperatures at the northeast and center of The Netherlands. Such a variability is also found in the spring temperatures in the southeast of the country. In addition, we found that temperatures for most days of 1993 and 1998 experienced an intense variability across the whole country. We also found, however, that summer temperatures are homogeneous throughout the country for most of the study period.

These explored patterns of intra-annual variability are important to facilitate the understanding of climate change impacts on, for instance, phenology, as the variability in temperature, especially the breaking points, has critical impacts on the phenophases of plants (Verbesselt et al. 2010; Jong et al. 2013). In addition, the results in this study point out areas and time periods that have similar variability in temperature, which will facilitate the exploration of the driving forces and also the further buildup of prediction models.

All of these results indicate the possibilities of this newly developed triclustering algorithm to effectively analyze the complex patterns in daily temperature series. Nevertheless, it is important to note that the proposed BCAT_I algorithm and subsequent refinement of the triclusters are generic. As such, they can be applied to any 3D GTS. For instance, they could be used to generate environmental zones by applying them to a data cuboid formed by combining multiple climatic and environmental variables, or they could be used to analyze the expansion patterns of chain stores around the world by triclustering a data cuboid formed by counting the number of stores opening each year in every province (or appropriate administrative unit) and country. Therefore, BCAT_I contributes to a better understanding of complex patterns in spatiotemporal data.

ORCID

Menno-Jan Kraak http://orcid.org/0000-0002-8605-0484

References

Andrienko, G., N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, and D. Keim. 2010. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum* 29:913–22.

Andrienko, G., N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. 2009. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, 3–10. http://geoanalytics.net/and/papers/vast09.pdf (last accessed 8 June 2017).

Andrienko, N., and G. Andrienko. 2006. *Exploratory analysis of spatial and temporal data—A systematic approach*. Berlin: Springer-Verlag.

Banerjee, A., I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. 2007. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research* 8:1919–86.

Chmielewski, F.-M., and T. Rötzer. 2001. Response of tree phenology to climate change across Europe. *Agricultural and Forest Meteorology* 108:101–12.

Dhillon, I. S., S. Mallela, and D. S. Modha. 2003. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*, 89–98. Washington, DC: ACM. http://dl.acm.org/citation.cfm?id=956764 (last accesseed 8 June 2017).

Doi, H., O. Gordo, and I. Katano. 2008. Heterogeneous intra-annual climatic changes drive different phenological responses at two trophic levels. *Climate Research* 36 (3): 181–90.

Feng, C.-C., Y.-C. Wang, and C.-Y. Chen. 2014. Combining geo-SOM and hierarchical clustering to explore geospatial data. *Transactions in GIS* 18:125–46.

Grubesic, T. H., R. Wei, and A. T. Murray. 2014. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers* 104:1134–56.

Guo, D., J. Chen, A. M. Maceachren, and K. Liao. 2006. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics* 12:1461–74.

Hagenauer, J., and M. Helbich. 2013. Hierarchical self-organizing maps for clustering spatiotemporal data.

International Journal of Geographical Information Science 27:2026–42.
Han, J., M. Kamber, and J. Pei. 2011. *Data mining: Concepts and techniques*. Waltham, MA: Morgan Kaufmann.
Han, J., J.-G. Lee, and M. Kamber. 2009. An overview of clustering methods in geographic data analysis. In *Geographic data mining and knowledge discovery*, ed. H. J. Miller and J. Han, 150–87. London and New York: Taylor & Francis.
Harinarayan, V., A. Rajaraman, and J. D. Ullman. 1996. Implementing data cubes efficiently. *ACM SIGMOD Record* 25:205–16.
Helbich, M., W. Brunauer, J. Hagenauer, and M. Leitner. 2013. Data-driven regionalization of housing markets. *Annals of the Association of American Geographers* 103:871–89.
Jaagus, J., and R. Ahas. 2000. Space–time variations of climatic seasons and their correlation with the phenological development of nature in Estonia. *Climate Research* 15:207–19.
Ji, L., K.-L. Tan, and A. K. Tung. 2006. Mining frequent closed cubes in 3D datasets. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, 811–22. VLDB Endowment. http://dl.acm.org/citation.cfm?id=1164197 (last accessed 8 June 2017).
Jong, R. D., J. Verbesselt, A. Zeileis, and M. E. Schaepman. 2013. Shifts in global vegetation activity trends. *Remote Sensing* 5:1117–33.
Lenderink, G., H. Mok, T. Lee, and G. Van Oldenborgh. 2011. Scaling and trends of hourly precipitation extremes in two different climate zones—Hong Kong and The Netherlands. *Hydrology and Earth System Sciences* 15:3033–41.
Lewis, J. M., M. Ackerman, and V. De Sa. 2012. Human cluster evaluation and formal quality measures: A comparative study. In *Proceedings of the 34th Conference of the Cognitive Science Society (CogSci)*, 1870–75. https://www.researchgate.net/profile/Virginia_De_Sa/publication/267364511_Human_Cluster_Evaluation_and_Formal_Quality_Measures_A_Comparative_Study/links/54d476ba0cf2970e4e633d7c.pdf (last accessed 8 June 2017).
Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65.
Sim, K., Z. Aung, and V. Gopalkrishnan. 2010. Discovering correlated subspace clusters in 3D continuous-valued data. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, 471–80. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5694001 (last accessed 8 June 2017).
Sim, K., V. Gopalkrishnan, A. Zimek, and G. Cong. 2013. A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery* 26:332–97.
Verbesselt, J., R. Hyndman, A. Zeileis, and D. Culvenor. 2010. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment* 114:2970–80.
Walther, G.-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Barilein. 2002. Ecological responses to recent climate change. *Nature* 416:389–95.
Williams, S. E., and J.-M. Hero. 2001. Multiple determinants of Australian tropical frog biodiversity. *Biological Conservation* 98:1–10.
Williams, S. E., and J. Middleton. 2008. Climatic seasonality, resource bottlenecks, and abundance of rainforest birds: Implications for global climate change. *Diversity and Distributions* 14:69–77.
Wu, X., R. Zurita-Milla, and M.-J. Kraak. 2013. Visual discovery of synchronization in weather data at multiple temporal resolutions. *The Cartographic Journal* 50:247–56.
———. 2015. Co-clustering geo-referenced time series: Exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science* 29:624–42.
———. 2016. A novel analysis of spring phenological patterns over Europe based on co-clustering. *Journal of Geophysical Research: Biogeosciences* 121:1434–48.
Zhao, L., and M. J. Zaki. 2005. TRICLUSTER: An effective algorithm for mining coherent clusters in 3D microarray data. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, 694–705. http://dl.acm.org/citation.cfm?id=1164197 (last accessed 8 June 2017).

XIAOJING WU is a Postdoctoral Research Fellow in the Humanities, Arts, and Social Sciences at Singapore University of Science and Technology, 487372 Singapore. E-mail: xiaojing_wu@sutd.edu.sg. Her research interests focus on spatiotemporal pattern exploration using data mining methods and geovisualization and visualization and analysis of spatial networks in the urban context.

RAUL ZURITA-MILLA is an Associate Professor of Spatio-Temporal Analytics in the Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE, Enschede, The Netherlands. E-mail: r.zurita-milla@utwente.nl. His research interests focus on designing geocomputational approaches for the analysis and modeling of geographical phenomena using data mining and machine learning methods.

EMMA IZQUIERDO VERDIGUIER is a Postdoctoral Researcher in the Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE, Enschede, The Netherlands. E-mail: e.izquierdoverdiguier@utwente.nl. Her research interests focus on feature extraction and image classification based on kernel methods and phenology. Previously she worked on automatic identification and classification of multispectral images.

MENNO-JAN KRAAK is a Full Professor of Geovisual Analytics and Cartography in the Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE, Enschede, The Netherlands. E-mail: m.j.kraak@utwente.nl. His research interests focus on mapping time, trying to find the most suitable visual representations to show changes in geospatial-temporal data. He is currently president of the International Cartographic Association.

Appendix

Bregman Cuboid Average Triclustering Algorithm with I-Divergence

BCAT_I starts by randomly mapping m stations to h station clusters, n years to l year clusters, and v days to g day clusters. It yields $R \in \mathbb{R}^{m \times h}$, $C \in \mathbb{R}^{n \times l}$, and $T \in \mathbb{R}^{v \times g}$, which are binary matrices to indicate the membership of station clusters, year clusters, and day clusters. The loss function is formulated to measure the loss of mutual information before and after implementing triclustering:

$$f_{loss} = I(S;Y;D) - I(\hat{S};\hat{Y};\hat{D}), \tag{A.1}$$

where $I(;)$ indicates the mutual information between variables.

After that, the new station clustering is updated. To optimize the mapping from stations to station clusters, first the original temperature cuboid is reshaped as a data matrix $O' \in \mathbb{R}^{m \times vn}$ with m stations as rows and $v \times n$ days/years, indicating all days in each year for all years, as columns. By this means, the reshaped data matrix with $o'_{s,dy}$ ($dy \in \{1, \ldots, v \times n\}$) as elements focuses on each station while retaining information in all days and years. Correspondingly, the loss function in Equation A.1 can be reformulated as the loss of mutual information before and after mapping the reshaped data matrix:

$$f_{loss} = I(S;DY) - I(\hat{S};\widehat{DY}). \tag{A.2}$$

BCAT_I measures the loss of mutual information using I-divergence. Therefore, the loss function in Equation A.2 can be represented as I-divergence between this reshaped data matrix and a matrix that approximates it:

$$I(S;DY) - I(\hat{S};\widehat{DY}) = D_I(O' \,||\, A), \tag{A.3}$$

where $D_I(\cdot\,||\,\cdot)$ is the I-divergence between two elements (i.e., data matrices). $A \in \mathbb{R}^{m \times vn}$ is the approximation matrix of O' with $a_{s,dy}$ as elements. The calculation of A is determined by O', the current mapping, and the statistics of O' to be preserved in the approximation. Due to the reshaping, the current mapping for columns from O' to $\widehat{O'}$ is the repetition of T for n times because of the data arrangement of columns in O' and named $T1 \in \mathbb{R}^{vn \times g}$. Because the statistics of O to be preserved during triclustering are tricluster averages to consider variations along stations, years, and days, the averages of elements within the same station and day/year clusters are preserved in A. A is calculated as

$$A = R\,\widehat{O'}\,T1^T, \tag{A.4}$$

where $\widehat{O'}$ is the clustered matrix of O' calculated as averages of elements that are intersected by each station and day/year cluster. $T1^T$ denotes the transposition of T1.

Then the I-divergence between the reshaped data matrix and its approximation matrix in Equation A.3 can be further represented as

$$D_I(O' \,||\, A) = \sum_{\hat{s}} \sum_{\hat{dy}} \sum_{s \in \hat{s}} \sum_{dy \in \hat{dy}} o'_{s,dy} \log \frac{o_{s,dy'}}{a_{s,dy}}. \tag{A.5}$$

According to the illustration in Banerjee et al. (2007), Equation A.5 can be decomposed into the I-divergence according to the rows (stations) and columns (days/years). The decomposed loss function calculating I-divergence of mapping from stations to station clusters is

$$D_I(O' \,||\, A) = \sum_{\hat{s}} \sum_{s \in \hat{s}} o'_{s,\cdot} \log \frac{o'_{s,\cdot}}{a_{s,\cdot}}. \tag{A.6}$$

The assignment of each station to different station clusters leads to different values of the loss function and the optimal clustering result is to minimize the loss function for each cluster assignment. Therefore, the new mapping from stations to station clusters is updated by minimizing Equation A.6, which yields the optimal station clustering result encoded in R^* in a binary way.

$$j^*(\cdot) = argmin_{j \in [1,h]} D_{I i,j}\ [i]_1^m. \tag{A.7}$$

After that, the mapping from years to year clusters is optimized. First O is reshaped as a data matrix $O' \in \mathbb{R}^{n \times mv}$ with $o'_{y,sd}$ ($sd \in \{1, \ldots, m \times v\}$) as elements to focus on each year with information in all stations and days retained. This reshaped data matrix is with n years as rows and $m \times v$ stations/days as columns, arranged as all stations in each day for all days. Accordingly, the loss function in Equation A.1 is reexpressed as the loss of mutual information before

and after clustering the reshaped data matrix:

$$f_{loss} = I(Y; SD) - I(\hat{Y}; \widehat{SD}). \quad (A.8)$$

Measured with I-divergence by BCAT_I, the loss function in Equation A.8 can be reexpressed as I-divergence between this O′ and its approximation matrix:

$$I(Y; SD) - I(\hat{Y}; \widehat{SD}) = D_I(O' \mid\mid A), \quad (A.9)$$

where $A \in \mathbb{R}^{n \times mv}$ is the approximation matrix of O' with $a_{y,sd}$ as elements. The calculation of A is

$$A = C\,\widehat{O'}\,R1^T, \quad (A.10)$$

where $\widehat{O'}$ is the clustered matrix of O′ and calculated as averages of elements intersected by each year and station/day clusters. The current column mapping indicated by $R1 \in \mathbb{R}^{mv \times h}$ is the repetition of R, updated from row clustering, for v times due to column arrangement in the reshaped data matrix. $R1^T$ is the transposition of R1. Just as in row clustering optimization, the averages of elements within the same year and day/station clusters are to be preserved.

Then Equation A.9 can be further expressed as

$$D_I(O' \mid\mid A) = \sum_{\hat{y}} \sum_{s\hat{d}} \sum_{y \in \hat{y}} \sum_{sd \in \hat{sd}} o'_{y,sd} log \frac{o'_{y,sd}}{a_{y,sd}}. \quad (A.11)$$

Equation A.11 can be decomposed regarding rows (years) and columns (stations/days). The decomposed loss function measuring I-divergence of mapping from years to year clusters is

$$D_I(O' \mid\mid A) = \sum_{\hat{y}} \sum_{y \in \hat{y}} o'_{y,\cdot} log \frac{o'_{y,\cdot}}{a_{y,\cdot}}. \quad (A.12)$$

Because the mapping from each year to different year clusters results in different values of the loss function in Equation A.12, the optimization is achieved by minimizing the loss function for the mapping of each year. Such optimization is done according to Equation A.13. Thus, the mapping from years to year clusters is updated and the optimal year clustering result is produced and saved in C^* with binary encoding:

$$q^*(\cdot) = argmin_{q \in [1,l]} D_{Ip,q}\ [p]_1^n. \quad (A.13)$$

The last step is to optimize the mapping from v days to g day clusters. First, O is reshaped as a data matrix $O' \in \mathbb{R}^{v \times nm}$ with $o'_{d,ys}$ ($ys \in \{1, \ldots, n \times m\}$) as elements to focus on each day while keeping information in all years and stations. The rows of the reshaped data matrix are days and columns are years/stations as columns arranged as all years in each station for all stations. Accordingly, the loss function in Equation A.1 is reformulated in this step as the loss of mutual information before and after clustering:

$$f_{loss} = I(D; YS) - I(\widehat{D}; \widehat{YS}). \quad (A.14)$$

Due to the I-divergence measure used by BCAT_I, Equation A.14 can be reformulated as

$$I(D; YS) - I(\widehat{D}; \widehat{YS}) = D_I(O' \mid\mid A), \quad (A.15)$$

where $A \in \mathbb{R}^{v \times nm}$ is the approximation matrix of O′ with $a_{d,ys}$ as elements. Here A is calculated as

$$A = T\,\widehat{O'}\,C1^T, \quad (A.16)$$

where $\widehat{O'}$ is the clustered matrix of O′ and calculated as average values of elements that are intersected by each day and year/station clusters. The current column mapping from O' to the clustered matrix is indicated by C1, which is the repetition of C yielded from updating year clustering for m times because of column arrangement in the reshaped matrix O'. $C1^T$ is the transposition of C1. A preserves the averages of elements within the same day and year/station clusters in the process.

Then the I-divergence between the reshaped days × years/stations matrix and its approximation matrix can be further expressed as

$$D_I(O' \mid\mid A) = \sum_{\hat{d}} \sum_{\hat{ys}} \sum_{d \in \hat{d}} \sum_{ys \in \hat{ys}} o'_{d,ys} log \frac{o'_{d,ys}}{a_{d,ys}}. \quad (A.17)$$

Then Equation A.17 is decomposed into the loss function in terms of the rows (days) and columns (years/stations). The decomposed loss function regarding the mapping from days to day clusters is

$$D_I(O' \mid\mid A) = \sum_{\hat{d}} \sum_{d \in \hat{d}} o'_{d,\cdot} log \frac{o'_{d,\cdot}}{a_{d,\cdot}}. \quad (A.18)$$

The last step of optimization is minimizing the loss function in Equation A.18 for each day cluster assignment. Therefore, the mapping from days to day clusters is updated according to Equation A.19, which yields the optimal day clustering result and encoded in T^* in the binary way.

$$e^*(\cdot) = argmin_{e\in[1,g]} D_{Iw,e}\ [w]_1^v. \tag{A.19}$$

Finally, the loss in mutual information is recalculated with the updated mapping R^*, C^*, T^* according to the loss function in Equation A.1 until convergence (i.e., the loss achieves a local minimum).