CLUSTERING-BASED APPROACHES TO THE EXPLORATION OF GEO-REFERENCED TIME SERIES

Xiaojing Wu

Examining committee:

Prof. Dr. A.A. Voinov	University of Twente
Prof. Dr. V.G. Jetten	University of Twente
Prof. Dr. T. Cheng	University College London
Prof. Dr. N. Andrienko	Fraunhofer Institute IAIS/City University London

ITC dissertation number 286
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4161-9
DOI 10.3990/1.9789036541619

Cover designed by Benno Masselink
Printed by ITC Printing Department

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

CLUSTERING-BASED APPROACHES TO THE EXPLORATION OF GEO-REFERENCED TIME SERIES

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday 07 July 2016 at 14:45 hrs

by

Xiaojing Wu

born on 9 December 1986

in Shandong Province, China

This dissertation has been approved by

Prof. Dr. M.J Kraak (Promoter)
Dr. R. Zurita-Milla (Co-promoter)

Acknowledgements

The long PhD trip finally comes to the end. This trip cannot be completed without the help of many people. By this chance, I would like to express my sincere gratitude to all of them for their guide, support and company.

First of all, I would like to express my deep gratitude to my promoter Prof. Menno-Jan Kraak. Thank you for giving me the opportunity to do my PhD research at ITC. I appreciate all the discussions we had, which have guided me to the right way to become a researcher, especially in the first two years because I rarely knew how to do research then. I also appreciate your encouragement and support when I met problems in my research.

My deep gratitude also goes to Dr. Raul Zurita-Milla, my daily supervisor, for your strict and patient supervision on each detail of my research. From the first to last chapter of this thesis, from the preliminary ideas to the published journal papers, from the programming at the beginning to the writing and revision of the manuscripts, you have given a lot of suggestions and put a lot of time and efforts to help me with my research. I feel lucky to work with you as my supervisor because I could not finish my PhD research at this time without your supervision. Besides, your strict attitude towards research and your eagerness for the knowledge exemplifies a great researcher, which inspires me in my own research and will benefit me afterwards.

I would also like to express my thanks to my MSc supervisor Prof. Wenxiu Gao, who recommended me to pursue my PhD research in ITC. I also own my thanks to EMECW program, which provided me with the scholarship for the financial support. Also my thanks go to Prof. Guofeng Wu and Dr. Tiejun Wang. As the contacts of the EMECW program, they patiently answered my questions about the scholarship.

I would also like to express my thanks to colleagues in ITC. The colleagues in GIP department are very helpful and give good feedbacks at the research meetings. Thanks to Prof. Alexey Voinov, Dr. Connie Blok, Dr. Corné van Elzakker, Dr. Otto Huisman, Ellen-Wien Augustijn, Bas Retsios, Jolanda Kuipers, Manuel Garcia Alvarez, Ahmed Ibrahim, Oliver Macapinlac, Rehmat Ullah, Peng Wang, and Xiaoling Wang for their suggestions and help. Also thanks to Hamed Mehdi Poor, Irene Garcia Marti and Norhakim Yusof for the discussions and sharing. Special thanks go to Emma Izquierdo-Verdiguier, for her encouragements, perspectives and excellent programming to contribute to my last individual paper. Also thanks to Loes for her patient answers whenever I have a question and appear in her office. I wish all of you all the best in your life.

I also own thanks to my friends in ITC for their company and help. We have shared the happiness that will stay forever in my memory. Many thanks to, Zhihui Wang, Linlin Li, Zhenwen He, Ying Zhang, Yiwen Sun, Mengmeng Li, Sudan Xu, Fengfan Yang, Ya Ma, Xinping Ye, Yijian Zeng and Liang Rong, Xuelong Chen and Zhangbo, Shaoning Lv, Zhuo La, Cesar, Binbin Wang, and Novi. Special thanks to Irma Kveladze and Qiuju Zhang for sharing my concerns and giving their friendly help. Also special thanks to Myri, my best friend, for the friendship and happiness we have shared.

Last but not least, I express my deep gratitude to my parents and parents-in-law for their endless support and love, which make it possible for me to study abroad without worries. My foremost thanks go to my husband, Donghai Zheng. Thank you for coming into my life, which makes this PhD trip more joyful. Also thank you for being so understanding and supportive all the time.

Table of Contents

x

Chapter 1 Introduction

1.1 Background

Due to the technological advancements in data collection and sharing, unprecedented volumes of spatio-temporal data with various scopes and coverages are becoming accessible (Guo 2003, Miller and Han 2009). Remote sensing systems and geosensor networks gather vast amounts of data automatically while crowdsourcing approaches allow collecting massive amounts of both ad-hoc and planned volunteered geographic information (Goodchild 2009, Miller and Han 2009). The extraction of useful information and actionable knowledge from these huge volumes of available data becomes the overarching challenge of spatio-temporal analytics. In this context, data mining is especially necessary because it distils information and knowledge from data and reveals patterns hidden in large datasets (Han et al. 2011, Hagenauer and Helbich 2013). These patterns contribute to an improved decision-making in many application areas, such as climatology (Crane and Hewitson 2003), hydrology (Lenderink et al. 2011), urban planning (Pan et al. 2013), transportation (Giannotti et al. 2007), disease prevention (Carrel et al. 2009), etc. This PhD thesis focuses on methods to mine and explore patterns from one important type of spatio-temporal data, namely geo-referenced time series (GTS from now on).

Spatio-temporal data contains values for one or more attributes of any geographical phenomenon that are recorded at specific locations and timestamps. According to the extension of the spatial and temporal components, several types of spatio-temporal data can be classified (Kisilevich et al. 2010). GTS, the focus of this PhD thesis, contain time-evolving sequences for one or more attributes recorded at fixed locations, and typically at uniform temporal intervals. GTS are quite common: for instance, daily precipitation sequences measured at weather stations, or crime incidents recorded at administrative units, etc. Moreover, sequences of gridded images such as satellite image time series are also GTS, where the regularly distributed grids are the fixed locations (Kisilevich et al. 2010). Useful information such as spatio-temporal patterns of precipitation can be obtained by mining GTS (Hsu and Li 2010).

As one important task of data mining, clustering identifies data elements that are similar in terms of their attribute(s) and groups them together. As a result, the data elements in each group, or cluster, are similar to each other and dissimilar to those assigned to other groups (Berkhin 2006, Miller and Han 2009). Therefore, clustering analysis provides an overview of the data at a higher level of abstraction, and also an observation of details when looking into each cluster (Andrienko et al. 2009, Han et al. 2009). By this means, it effectively reveals the

patterns buried in the data. Nevertheless, these patterns are often not easy to understand because clustering methods are incapable of attaching meaning to the identified clusters (Guo 2009). Moreover, clusters extracted from GTS contain spatial, temporal and attribute information, and this makes their interpretation a complex task.

Graphic representations provided by (geo)visualization can be used to facilitate the understanding of the revealed (geographical) patterns from GTS (Dykes et al. 2005). By providing visual forms to examine clusters in their original dimensions (i.e. geographic space, time and attribute), geovisualization allows for a better understanding and also further interpretation of complex patterns (Miller and Han 2009).

Therefore, this research focuses on combining clustering methods and geovisualization techniques to fully explore spatio-temporal patterns from GTS. More precisely, clustering methods are used to identify similar groups of data elements, thereby uncovering patterns buried in the data, and geovisualization techniques are used to represent the patterns.

1.2 Geo-referenced time series

Being a type of spatio-temporal data, GTS are intrinsically structured in three dimensions as space, time and attribute (Guo et al. 2006). As a result of being collected at fixed locations and consistent temporal intervals, GTS can be organized as a data table or cuboid under different situations (Figure 1.1). If they refer to only one attribute, GTS can be organized as a data table where rows indicate the locations, columns indicate the timestamps at which this attribute was recorded, and the elements of the table are the values of the attribute (Figure 1.1a). For example, daily precipitation series that are recorded at weather stations in one year and fit into a data table. If the GTS refer to two or more attributes, they can be organized in a data cuboid defined by rows, columns and depths as its three dimensions. In this cuboid, rows refer to locations, columns to the timestamps, depths to attributes and the elements of the cuboid are the values of the attributes (Figure 1.1b). For instance, daily precipitation and humidity series recorded at stations that fit into a data cuboid. Alternatively, if there is one attribute but either the spatial or the temporal dimension contains nested hierarchies (e.g. year and day in the case of time), GTS can also be organized as a data cuboid where rows refer to locations, columns to the years, depths to the days, for instance, and elements to the value of the attribute (Figure 1.1c). For example, the daily precipitation series recorded at stations over years.

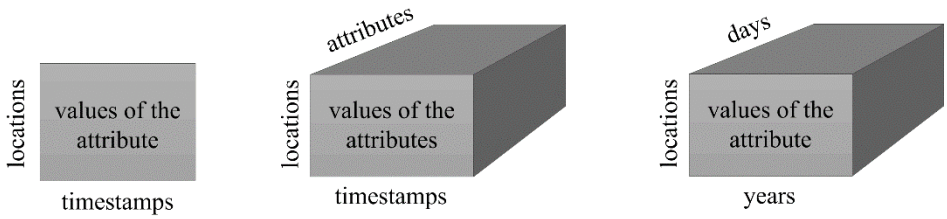


Figure 1.1: Various formats of GTS under different situations. a) GTS with one attribute; b) GTS with two or more attributes; c) GTS with one attribute and nested hierarchies in spatial or temporal dimension.

Since raw data is limited in directly providing useful information (Fayyad 1996), many efforts have been made to extract actionable knowledge from GTS, including the use of data mining and geovisualization. For instance, Crane and Hewitson (2003) applied the clustering analysis for regionalization of individual stations. MacEachren and Brewer (2004) used animation to explore the dynamics of climatic time series. Verbesselt et al. (2010) analyzed the trends of phenological changes by applying regression analysis. However, in terms of data mining analysis, those studies only deal with GTS that fit into a data table (i.e. one single attribute; Figure 1.1a). Methods that are able to analyze other more complex GTS are thus needed. In terms of visual analysis, geovisualization alone is incapable of handing large and complex datasets. In this situation, the combination of data mining and geovisualization can provide possible solutions.

1.3 Clustering methods

Among data mining techniques, classification and clustering methods are widely used in spatio-temporal analytics because they can reduce the amount of data by identifying representative classes/clusters (Han et al. 2009, Shekhar et al. 2009). As such, they greatly facilitate the exploration of large and complex datasets. Compared with classification, clustering, also known as unsupervised classification, is especially useful when limited knowledge of a dataset exists (Guo 2003).

Clustering methods for pattern analysis have drawn attention in many applications and for various types of (spatio-temporal) data (Jain et al. 1999, Berkhin 2006, Han et al. 2009, Kriegel et al. 2009). Depending on the dimensions of data involved in the analysis, clustering methods can be classified into: one-way clustering, co-clustering and tri-clustering methods.

1.3.1 One-way clustering

One-way clustering, also called traditional clustering, organizes the data into groups based on the similarity of elements along a single data dimension (Dhillon et al. 2003). Such a clustering analysis has been widely used in many areas (Han et al. 2009). For instance in biology, to identify groups of genes in terms of similar expression patterns (Ben-Dor et al. 1999). In information retrieval, to identify groups of documents with similar words (Steinbach et al. 2000). In business management, to find out the groups of customers with similar consumption behaviour (Wan et al. 2005).

In terms of GTS, one-way clustering typically analyses along either the spatial or the temporal dimension separately, i.e. the rows or the columns of a data table that contain GTS. In the case of an analysis from a spatial dimension, locations (rows) are clustered by the similarity of the attribute's values along all timestamps (columns). The resulting clusters are groups of locations with similar behaviour and this is why this approach is also called spatial clustering. In the case of an analysis from a temporal dimension, timestamps (columns) are clustered with similar values of the attribute along all locations (rows) and the resulting clusters are groups of timestamps with similar behaviour. Because of this, this approach is also called temporal clustering.

Previous studies that analysed patterns from GTS employed traditional clustering methods. Most of these studies applied these methods for either spatial clustering, for instance, in climatology for regionalization (Crane and Hewitson 2003), or for temporal clustering for grouping years with similar phenology (e.g. years with late summers) (Ahas and Aasa 2003). Only a few studies used traditional clustering methods for both types of clustering to analyze the spatial and the temporal variations in the GTS, for instance, for company relocation (Guo et al. 2006) or crime rates analysis (Andrienko et al. 2010). In these studies, issues related to the modifiable spatial and temporal resolutions of the data should draw attention. Openshaw and Taylor (1979) showed that the spatial resolution at which the analysis is performed affects the explored patterns and defined it as one part of Modifiable Areal Unit Problem (MAUP). Recently, Coltekin et al. (2011) found that the identified patterns also depend on the temporal resolution of the analysis. They defined it as one part of Modifiable temporal unit problem (MTUP; refer to section 1.5 for details), which is a parallel concept to MAUP. Due to the fact that MAUP has been widely studied (Jelinski and Wu 1996, Dungan et al. 2002, Dark and Bram 2007, Arbia and Petrarca 2011), this thesis concentrates on the study of MTUP.

1.3.2 Co-clustering

Unlike one-way clustering, co-clustering considers two dimensions of the data and groups data elements into co-clusters with similar values along these dimensions (Dhillon et al. 2003). Initially proposed by Hartigan (1972), co-clustering, also called bi-clustering, has drawn increasing attention in recent years for pattern analysis in many applications, especially bio-informatics (Madeira and Oliveira 2004, Banerjee et al. 2007). For instance, Cheng and Church (2000), Cho et al. (2004) and Pensa and Boulicaut (2008) applied co-clustering methods to gene expression data to focus on small subsets of genes and conditions of interest. In text content analysis, a co-occurrences table for word-document is co-clustered for both document and word categorization (Takamura and Matsumoto 2002, Dhillon et al. 2003). Also in multimedia content analysis, Qiu (2004) and Cai et al. (2008) respectively applied co-clustering methods to images and auditory scenes to facilitate information retrieval. Besides, in recommender system (e.g. movies), co-clustering helps to analyze the patterns in the ratings of various users to build a prediction model (Hofmann 2004).

In terms of GTS, co-clustering methods can be used to concurrently analyze the spatial and temporal dimensions of data organized as a table. By simultaneously grouping rows and columns in the data table, they identify groups of data elements formed by intersecting each row- and column-cluster. Consequently, co-clustering methods result in groups (co-clusters) that contain similar attribute values along the spatial and temporal dimensions. Thus, patterns extracted by co-clustering methods describe the space-time varying behaviour of an attribute. To the best of our knowledge, co-clustering methods have never been used for exploring the concurrent spatial and temporal patterns from GTS.

1.3.3 Tri-clustering

Tri-clustering identifies groups of data elements by considering their similarity along the three dimensions of GTS that fit into a data cuboid (Zhao and Zaki 2005). Although most tri-clustering methods are relatively new, they have already been used in several studies (Sim et al. 2013). For instance, Zhao and Zaki (2005) introduced the TRICLUSTER algorithm and applied it to a 3D gene expression dataset. Ji et al. (2006) proposed the CubeMiner algorithm to find the frequent co-occurrences of gene-sample-time. A 3D cluster model named S^2D^3 is proposed by (Xu et al. 2009) to also identify gene-sample-time tri-clusters in the microarray dataset. Besides, Sim et al. (2010) developed the MIC algorithm and applied it to a 3D financial dataset. However, CubeMiner can be only applied to

binary 3D datasets while the clusters identified by S^2D^3 are not axis-parallel, which makes it difficult to understand the resulting patterns. Even though TRICLUSTER and MIC can be applied to quantitative datasets to identify axis-parallel tri-clusters, they focus on identifying significant tri-clusters. In this context, significant clusters are those regarded as more meaningful than others for answering an specific task and usually of few amount (Sim et al. 2013). As such, these tri-clustering algorithms are incapable of fully analysing GTS. A new tri-clustering algorithm is thus needed to analyze 3D GTS. This new tri-clustering algorithm should enable the identification of all tri-clusters by simultaneously partitioning the data cuboid across its three dimensions. As an example, let's take a 3D GTS formed when studying various attributes in space and time (Figure 1.1b). By grouping rows, columns and depths at the same time, this tri-clustering algorithm partitions the data cuboid into sub-cuboids formed by intersecting each row-, column- and depth-cluster. These sub-cuboids, also called tri-clusters, contain data elements with similar values along locations, timestamps and attributes. Patterns extracted from tri-clusters are thus able to describe the variations of values along all three dimensions of the GTS.

1.4 Geovisualization

Clustering methods are capable of identifying complex patterns buried in the data, but they lack the ability to represent patterns (Guo 2003). Geovisualization techniques are thus needed to support the visual representation of the identified patterns and thereby facilitate their understanding and further interpretation.

Geovisualization refers to the integration of graphic representation approaches from cartography, scientific visualization, exploratory data analysis and GIScience (Kraak 2000, MacEachren and Kraak 2001). It provides theories, approaches and especially techniques for the visual representation and exploration of the data (MacEachren and Kraak 2001). Various geovisualization techniques have been developed and applied in diverse fields to directly exploit the patterns in data (Dykes et al. 2005).

According to the characteristics of the data to be represented or explored, appropriate geovisualization techniques need to be selected given the task at hand (Koua 2005). For instance, to visualize the spatial information in spatio-temporal data, a traditional geographical map is the most typical and suitable representation. Whereas a timeline provides an overview of data related to the temporal aspect, which could be linear (for linear time, e.g. year) or circular (e.g. season) (Kraak 2005). In order to visualize complex phenomenon or process, these basic representations, seen as the interface to the data, can be developed to other forms

or variants. For instance to display the dynamics of spatial and temporal information, animations or small multiples can be developed from maps (Kraak et al. 1997, Maceachren et al. 2003) and ringmaps can be developed from circular timelines (Zhao et al. 2008). Moreover, when different aspects of data need being visualized, multiple representations are necessary as they combine more than one geovisualization techniques and offer alternative views of the data (Kraak 2003). For instance, the combination of the geographical map and ringmap (Zhao et al. 2008) offers the views of both spatial and dynamic temporal information. However, in respect of large and complex datasets, geovisualization is limited in revealing patterns since clutters and overlappings tend to appear (MacEachren and Kraak 2001, Guo 2003). Clustering methods can offer a solution to the issue by pre-processing the datasets.

Several studies have used geovisualization techniques and clustering methods to explore patterns from the data. For instance, Gahegan and Brodaric (2002) combined the dynamic geographical map and self-organizing maps (SOMs), the latter used as a clustering tool. Swayne et al. (2003) used parallel coordinates plots (PCP) to visualize hierarchical clustering results. Guo (2009) developed an integrated framework consisting of PCP, geographical maps and SOMs to analyze patterns from GTS. Andrienko et al. (2010) used time graphs, geographical map and SOMs matrix map to represent SOMs clustering results of GTS. The challenge to develop such frameworks is to choose or develop suitable clustering methods to extract patterns and then select appropriate geovisualization techniques to visualize them.

1.5 Modifiable temporal unit problem

Due to the increasingly available spatio-temporal data, research problems related to the temporal dimension have started to draw attention. Before the MTUP was formally defined, Li and Chou (2000) found that the modifiable temporal resolution of the data causes different clustering results. Hudson et al. (2011) found the same problem when analyzing the records of species. Coltekin et al. (2011) summarized that differences in patterns can be found when the same phenomenon are analyzed at different temporal resolutions, and formally defined the MTUP. As such, the clustering analysis of GTS in this research should also consider the impacts of different temporal resolutions on the extracted patterns.

1.6 Research objectives

The main objective of this research is to combine clustering methods and geovisualization techniques to enable the full exploration of spatial and temporal patterns from GTS. More precisely, this research proposes and develops approaches based on one-way clustering, co-clustering and tri-clustering methods to allow the full extraction of patterns from GTS, and then uses appropriate geovisualization techniques to visualize these patterns for understanding and interpretation.

To achieve the above objectives, the following research questions must be answered in this research:

Q1. How can a one-way clustering method be combined with geovisualization techniques to separately explore the spatial and temporal patterns from GTS? How does the MTUP affect the explored patterns?

Q2. How can a co-clustering method be combined with geovisualization techniques to concurrently explore the spatial and temporal patterns from GTS?

Q3. How can a co-clustering method and *k*-means be combined to enable the full exploration of spatio-temporal patterns from GTS?

Q4. How to develop a tri-clustering algorithm that enables the full extraction of patterns from GTS that fit into a data cuboid?

1.7 Thesis outline

Together with the introduction and the synthesis, this PhD thesis consists of six chapters. Four of these chapters have been published in, or are submitted to, international peer-reviewed journals. The following paragraphs summarize the contents of all chapters.

Chapter 1 describes the research background and context, states the research objectives and questions, and outlines the structure of this thesis.

Chapter 2 presents an analytical approach that combines a one-way clustering method and geovisualization techniques to analyze spatial and temporal patterns from GTS in an independent fashion. This analysis is done at multiple temporal resolutions to study the MTUP. A dataset of daily average temperatures collected

at 28 Dutch meteorological stations from 1992 to 2011 is used as case study in this chapter.

Chapter 3 introduces the use of a co-clustering method to identify groups of spatio-temporal data that have similar values along the spatial and the temporal dimensions. Heatmaps, small multiples and ringmaps are then used to visualize these groups or co-clusters. This chapter is illustrated with the same dataset used in Chapter 2 and the analysis is done at various temporal resolutions to study the MTUP.

Chapter 4 presents an analytical approach based on the co-clustering method used in Chapter 3 and *k*-means, where the latter is used to refine the co-clusters. Then heatmaps, small multiples and timelines are used to visualize the results. Together with a temperature-driven phenological model, this approach is used to explore European spring phenological patterns.

Chapter 5 develops a new tri-clustering algorithm. This algorithm enables the analysis of 3D GTS and the identification of tri-clusters, which are then refined by *k*-means. By applying to the dataset used in Chapter 2 but regarding it here as a GTS with one attribute and the nested temporal hierarchies (year and day), this tri-clustering analysis explores the spatio-temporal patterns of intra-annual variability in temperature. These patterns are then visualized using both 3D and 2D heatmaps, small multiples and timelines.

Chapter 6 provides an in-depth reflection on the results obtained in Chapters 2 to 5 by discussing their inter-relationships, answers the research questions, lists the main contributions, and provides recommendations for future studies.

Chapter 2 Visual discovery of synchronization in weather data at multiple temporal resolutions*

*This chapter is based on the paper: Wu, X., R. Zurita-Milla & M.-J. Kraak (2013). Visual Discovery of Synchronization in Weather Data at Multiple Temporal Resolutions. *The Cartographic Journal*, 50(3), 247-256.

Abstract:

Analyzing spatio-temporal weather patterns is fundamental to better understand the system Earth. Such patterns depend on the spatial and temporal resolution of the available data. Here this chapter studies a particular spatio-temporal pattern, namely synchronization, and how it is affected by different temporal resolutions and temporal heterogeneity. Twenty years of daily temperature data collected in 28 Dutch meteorological stations is used as case study. This chapter proposes an analytical approach based on self-organizing maps (SOMs) that allows exploring the data from two perspectives: (1) station-based, in which spatially synchronous weather stations are grouped into clusters; (2) year-based, in which temporal synchronization is analyzed using a calendar year as basic unit and similar years are clustered. Clusters are identified using SOMs U-matrix maps and displayed in cluster maps. Next, the spatial distribution of synchronous stations is displayed in the geographic space. Trend plots are used to illustrate trends in every cluster and the temperatures of stations and years are compared with the corresponding cluster's representative values to identify anomalies in the temperature records. The analysis is repeated at daily, weekly and monthly resolutions to study the effects of different temporal resolutions on synchronization. Also daily spatial synchronization results for all years with those for groups of daily synchronous years are analyzed to study the effects of temporal heterogeneity. Results show that synchronization results are different at different temporal resolutions. Monthly results are the most stable ones both in station-based and year-based. It is also observed that spatial synchronization results are simplified when considering temporal heterogeneity.

Key words: geovisualization, multiple temporal resolutions, self-organizing maps, synchronization, temporal heterogeneity

2.1 Introduction

Long and consistent time series of weather data are available in most countries. This data is fundamental to quantify climate change as well as to study a wide range of environmental processes. Thus, discovering and analyzing spatio-temporal weather patterns is essential for understanding the system Earth. One particularly interesting pattern is synchronization, which refers to the degree of temporal similarity between two or more time series (Hudson et al. 2011). The identification of weather synchronous regions or years helps to understand the impact of climate change on living organisms. For instance, years with synchronous temperature patterns present synchronous phenological development.

Analyzing weather data is not trivial because of large amount of data and, more importantly, because it analyzes the data with spatial and temporal dimensions. This sharply increases the complexity of discovered patterns and makes it necessary to bring in geovisualization (Dykes et al. 2005), which provides the visual forms of the patterns to simulate the visual thinking and thereby help exploit spatial and temporal information in them.

Another factor that hinders the identification of spatio-temporal patterns is that they change when analyzed at multiple temporal scales. Coltekin et al. (2011) identified the so-called Modifiable temporal unit problem (MTUP) which illustrates that the use of different temporal resolutions result in the identification of different patterns even when studying one phenomenon. The MTUP is, however, not new. Hudson et al. (2011) identified differences in synchronized species at seasonal and monthly resolutions and Jong et al. Jong and Bruin (2012) illustrated possible impacts of operating at different temporal resolutions when analyzing remotely-sensed vegetation indices.

Besides this, time is heterogeneous (Andrienko et al. 2010) and patterns may change when considering temporal heterogeneity. For instance, daytime's temperature is usually higher than that of nighttime in a day. Then the fluctuations of the daytime's temperature are different from these of a whole day's.

This chapter presents an analytical approach to discover synchronization in temperature data and to explore the MTUP and temporal heterogeneity effects on synchronization. The approach relies on a data mining method (self-organizing maps) and on various geovisualization techniques to study spatial and temporal patterns. As a case study, this chapter analyses 20 years of daily temperature data for 28 weather stations in the Netherlands.

2.2 Discovering synchronization in time series data

Synchronization can be identified separately in space and time. Thus, if there is data for *m* weather stations for *n* years, spatial synchronization means there is a degree of similarity among the evolution of temperature recorded by *x* (<=*m*) stations for *y* (<=*n*) years. Temporal synchronization means that for a chosen temporal unit (i.e. a year), stations recorded similar temperatures. In other words, spatial synchronization results in spatial clusters (i.e. regions with similar temperature) and temporal synchronization clusters years that behaved in a similar way (e.g. "dry" and "wet" years). This means that a robust and visually efficient clustering method is required to detect spatial and temporal synchronies.

Self-organizing maps (SOMs; Kohonen 2001) are used here to discover synchronization. SOMs map n-dimensional data onto usually two-dimensional grids where similar input data are mapped to the same or nearby grids. SOMs results can be efficiently visualized and explored. The so-called Unified distance matrix (U-matrix) map can be used to visualize similarity (distance) among grid cells using colors and thus is an objective tool to recognize clusters. Another way to visualize is to regard every cell as an n-dimensional vector and cells with similar vectors form topologically contiguous regions (clusters) in the SOMs grid. Finally, SOMs results can be projected onto a geographical map to see the spatial distributions of these clusters.

Many SOMs studies can be found in literature (Kohonen 2001, Crane and Hewitson 2003, Koua 2005, Agarwal and Skupin 2008, Andrienko et al. 2010, Hudson et al. 2011). The most related to the work in this chapter are Koua (2005), Andrienko et al. (2010) and Hudson et al. (2011). Koua (2005) employed SOMs to extract clusters and focused on exploring correlations and relationships of different attributes at different places between clusters. However, his/her study only considered spatial clustering at a single temporal granularity. Andrienko et al. (2010) employed SOMs to group data from spatial and temporal perspective separately, resulting in 'space-in-time SOMs' and 'time-in-space SOMs'. However, they did not compare clustering results at multiple temporal resolutions and did not check for trends and anomalies in the original input data. Hudson et al. (2011) applied SOMs to group eight Eucalypt species based on monthly and seasonal (flowering) records to illustrate different clustering results at different temporal granularities. However, they focused on attribute synchronization only, and did not consider spatial perspective.

This chapter extends above approaches in three aspects: (1) Considering effects of different temporal resolutions on synchronization; (2) Considering

effects of temporal heterogeneity on synchronization; (3) Comparing clustering results with original input data to explore trends and anomalies in the original data. In this chapter, a calendar year is used as the basic unit in temporal synchronization. This is a meaningful unit because it corresponds to the standard development cycle of plants and animal (Zhang, Friedl, et al. 2003, Kramer and Hanninen 2009, Peñuelas et al. 2009).

2.3 Data

To illustrate this chapter, Dutch daily temperature data is used. In particular, data collected at 28 meteorological stations for 20 years (1st January 1992 to 31st December 2011) is used. This data was downloaded freely from the website of the Royal National Meteorological Institute (known by its Dutch acronym, the KNMI).

Even though The Netherlands is a relatively small country, covering about 41,500 km^2 and spreading over 3^o-7^{o}E, 49^o-53^{o}W. It is located where a maritime climate meets a more continental one. This results in different temperatures and patterns in the southwest and northeast of the country.

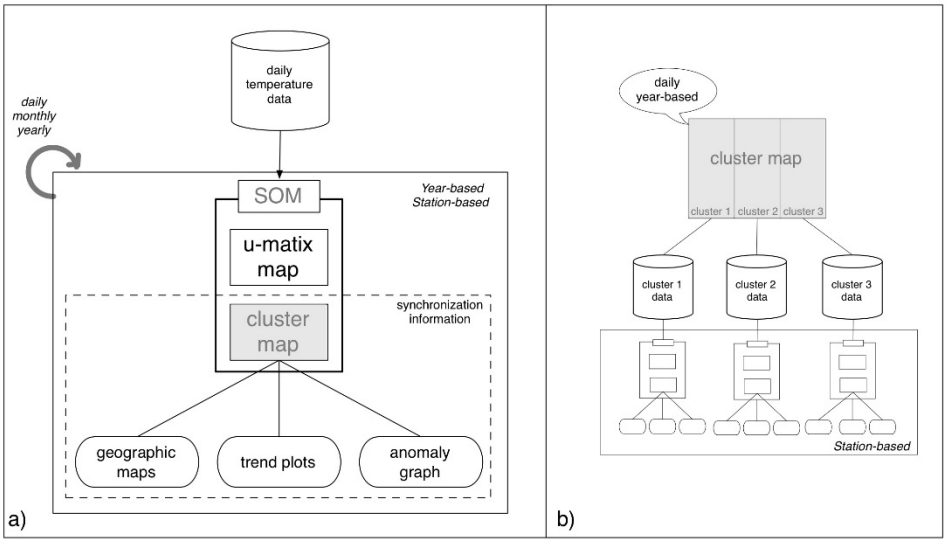


Figure 2.1: Methods to discover MTUP effects (a) and temporal heterogeneity effects (b) on synchronization.

2.4 Methods

2.4.1 SOMs clustering

The temperature data was first organized in a table where each column represents the daily temperature and each row contains values for each station (Figure 2.2a). Then, a Thiessen polygon map was generated around stations to show the spatial distribution of the stations and to define the area influenced by each station.

The highlighted column and row in the table represent SOMs clustering from spatial and temporal perspectives. For clarity this chapter calls these perspectives: station-based (Figure 2.2c) and year-based (Figure 2.2d) from here on. In the first case, daily temperature data was organized for the SOMs input where each row showed temperature of a single station and each column represented temperature of every day; in the second case the basic unit in temporal synchronization was chosen as 'a calendar year'. The sequence of stations was geographically ordered from south-west to north-east to be able to find trends (Figure 2.2b).Then the data was year sorted, columns for each year were organized as each row for the SOMs input where each row showed one year and each column indicated temperature of all stations in that year. The last day of February in leap years was discarded for equal SOMs input. Then each column was normalized to be in the range of [0 1] by subtracting the minimum and dividing by the subtraction of maximum and minimum and each row was seen as an input vector.

The MATLAB SOMs Toolbox 2.0, (http://www.cis.hut.fi/projects/somtoolbox/) was used to do the clustering and appropriate parameters such as the number and topology type of grid cells, neighborhood function, rough training and fine tuning steps were used for training the SOMs. In station-based the number of grid cells was fixed to 28, arranged on a hexagonal grid of 4 x 7. This number was chosen because the case study dataset has 28 stations. If none of stations were synchronous, each station would occupy one cell. For the same reason 20 grid cells, the number of years, was chosen for year-based analysis, arranged on a hexagonal grid of 4 x 5. Neighborhood function and rough training steps in this chapter were adopted from those in (Kohonen et al. 1996) both in station-based and year-based. Fine tuning steps was chosen as 15000 in station-based and 10000 in year-based to achieve more stable SOMs results.

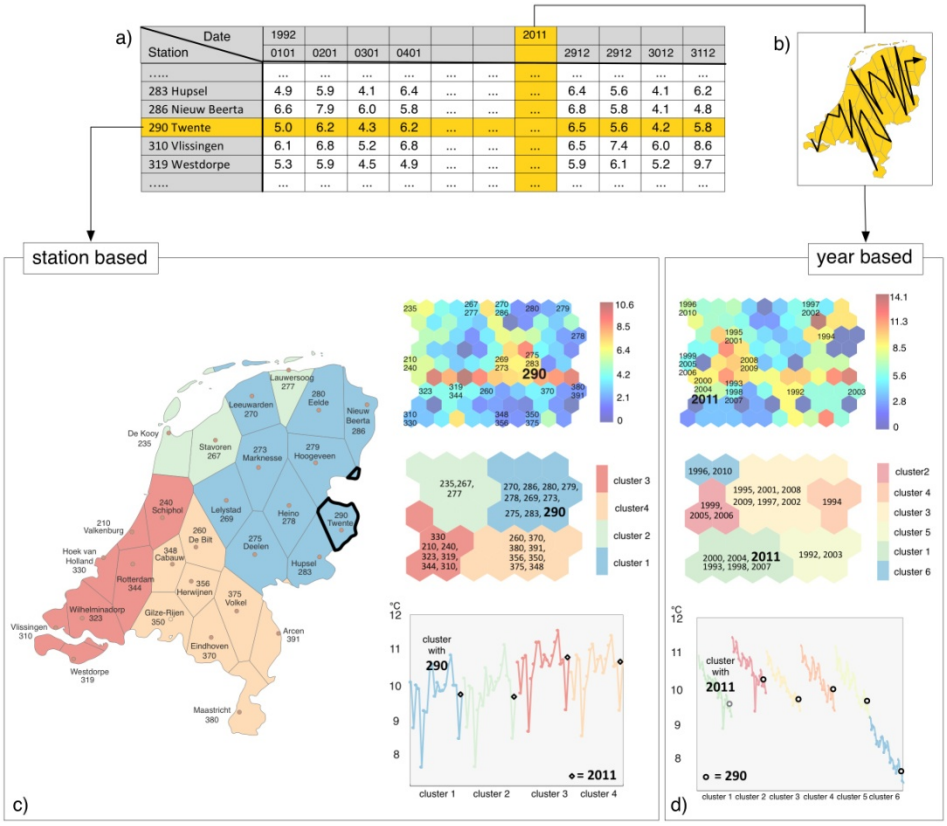


Figure 2.2: Discovering synchronization in station-based and year-based.

The U-matrix map was firstly used to visualize SOMs results for visually identifying clusters (upper right of Figures 2.2c, 2.2d). This map shows more units than SOMs cells because it not only contains distances from each cell to its neighbours but also the mean distances of surrounding values. Figure 2.2c shows the U-matrix map in station-based where stations in the same cell or nearby cells with blue colors between them can be grouped into one cluster and these stations are seen as synchronous.

Next this chapter used the trained SOMs to display clusters with colors (middle right in Figures 2.2b, 2.2d). Blue meant relatively low temperature while red indicated relatively high. The value of each cluster is the average value of cells with labels in that cluster. Clusters were ordered from low to high temperature (blue to red in the legend).

Finally the geography map was colored based on the legend of the cluster map to consistently display the spatial distribution of synchronous stations (left of Figure 2.2c).

2.4.2 Trend plot

A trend plot is a line plot that focuses on the trend of elements in a data set, a cluster in this chapter (Harris 1999). Trend plots offer a supplement to visualizations of the SOMs results. In station-based, they illustrate the temporal patterns for every spatial cluster (down right of Figure 2.2c) and in year-based, they show the spatial trends for synchronous years (down of Figure 2.2d). The color scheme of the trend plot was consistent with that of cluster maps.

2.4.3 Anomalies graph

The training of the SOMs is a procedure of refinement until the values of the cells are representative "generalizations" of the input data. Moreover, the value of each cluster is also different from the value of each cell with labels in that cluster. Therefore, the values of the stations or years belonging to a cluster are different from the value of that cluster. The differences are calculated using equation 2.1.

$$Diff_{ij} = \frac{s_i - c_j}{c_j} \tag{2.1}$$

Where s_i is the value of i^{th} input vector and c_j is the value of j^{th} cluster which the i^{th} input vector belongs to.

Differences are illustrated in Anomalies Graph (Figure 2.3), which helps to identify large differences, called anomalies in this chapter. The user can interactively explore the graph to identify anomalous dates or years (Figure 2.3).

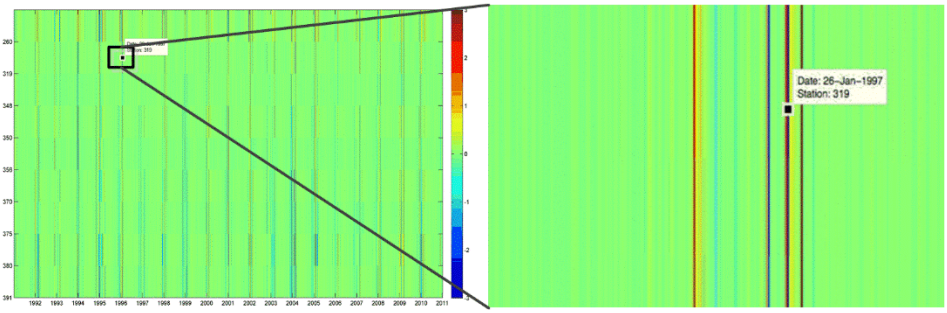


Figure 2.3: Differences displayed in Anomalies Graph.

2.4.4 Aggregation

In this chapter, temperature data was averaged to obtain weekly and monthly temperature from daily data. This is the most common method to downscale weather data (Estrella et al. 2007).

In station-based, the seven days average daily temperature along 20 years starting from the 1st of January 1992 was assigned to weekly data, the indivisible last four days discarded. In year-based, the seven days average daily temperature along 365 days of every station starting from the 1st of January was assigned to weekly data, with the temperature of indivisible last one day also assigned to weekly data. Monthly data was assigned the average temperature of every month both in station-based and year-based.

2.5 Results and Discussions

2.5.1 Spatial synchronization

Figure 2.4 displays synchronization results from spatial perspective at daily (Figure 2.4a), weekly (Figure 2.4b) and monthly (Figure 2.4c) resolutions. There are four clusters of synchronous stations at daily resolution and three at both weekly and monthly. The reduction of the numbers is justified because of the loss in details when daily temperature data is aggregated to weekly and monthly data.

Changes of elements in each cluster at different resolutions are explicitly displayed in the clusters (Figures 2.4a-II, 2.4b-II, 2.4c-II) and in the geography maps (Figures 2.4a-III, 2.4b-III, 2.4c-III). This is possibly due to the reduction in the number of clusters; also, it could because the daily temperature in those stations varies with more low values than high ones and therefore weekly temperature becomes low after aggregation.

Trend plots at different resolutions show similar trends from 1992 to 2011 for each cluster (Figures 2.4a-IV, 2.4b-IV, 2.4c-IV). It shows that in every group the 5th and 19th points, year 1996 and 2010, have the lowest temperatures. The 15th and 16th points, year 2006 and 2007, have the highest temperatures. The slope of every group increases monotonically due to global warming.

Finally, the anomalies graphs display differences between temperature value for every station and corresponding cluster at different resolutions (Figures 2.4a-V, 2.4b-V, 2.4c-V). Differences range from -3 to 3. From the graphs it is seen that at all resolutions most differences are around zero and there are big differences at the beginning of every year for all stations. One explanation for

this could be that temperature varies the most at the beginning of each year. However, at daily resolution there are big differences for a period at the end of

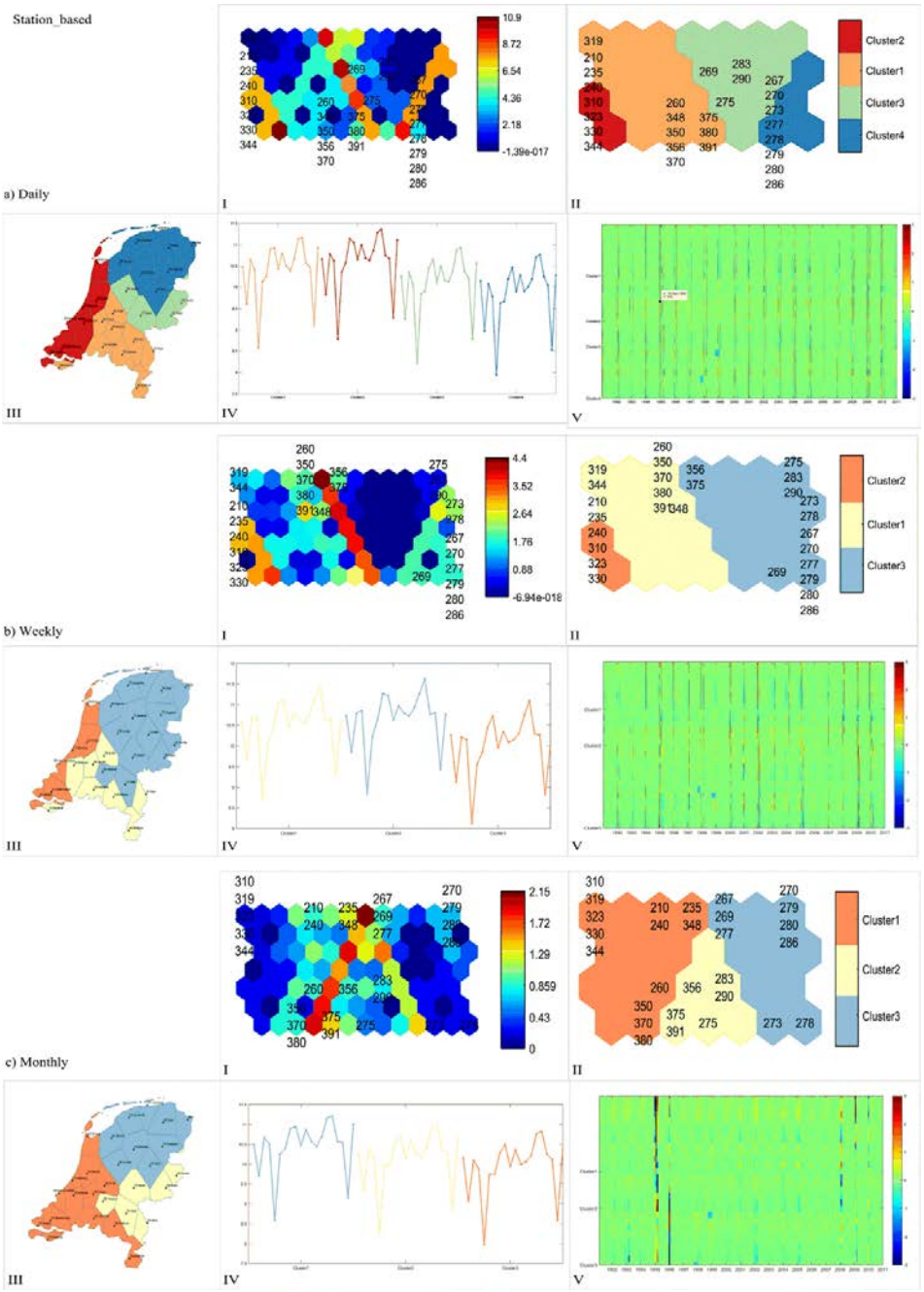


Figure 2.4: Synchronization results at daily (a), weekly (b) and monthly (c) resolutions in station-based.

1998 and 1999 for cluster3, which is because the temperature records in that period are lost and assigned zero. At the weekly resolution, the obvious differences in that period can also be seen but with shorter length due to coarser temporal granularity. At monthly resolution, it is noticeable that at the beginning and end of 1996 and also at the beginning of 2009 there are big differences. This indicates that the monthly temperature in those periods varies a lot from year to year even for synchronized stations.

In summary, different temporal resolutions indeed influence spatial synchronization results in weather data. Synchronization results at monthly resolution are regarded as more stable because differences in the anomalies graph are smallest.

2.5.2 Temporal synchronization

Figure 2.5 displays synchronization results from temporal perspective at daily (Figure 2.5a), weekly (Figure 2.5b) and monthly (Figure 2.5c) resolutions. There are four clusters of synchronized years at all resolutions but there are changes of years in each cluster at different temporal resolutions, especially from weekly to monthly. Changes can be clearly seen in the cluster maps (Figures 2.5a-II, 2.5b-II, 2.5c-II). This may due to variable weekly temperature for those years or because there are no much difference between temperatures among those years and then the division of those years is ambiguous.

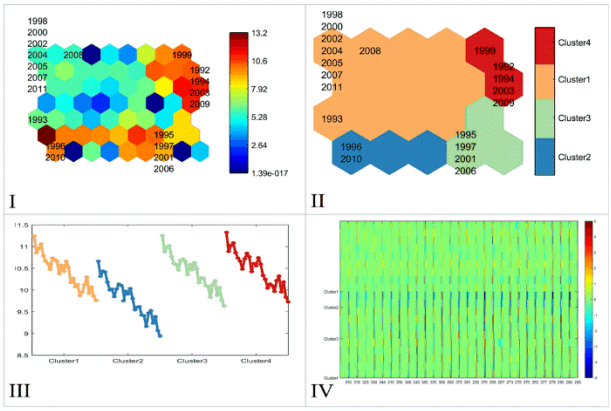
The trend plots show similar trends for temperature of all stations in Netherlands (Figures 2.5a-III, 2.5b-III, 2.5c-III). A decreasing trend of temperature from south-west to north-east can be seen. Additionally, the temperature of cluster2 is obviously lower than that of the other clusters. This cluster contains the years 1996 and 2010, which were also identified as low temperature years in the station-based trend plots.

Finally, the anomalies graphs display differences between temperature value for every year and corresponding cluster at different resolutions (Figures 2.5a-IV, 2.5b-IV, 2.5c-IV). The graphs show that at all resolutions most differences are around zero although the legend ranges between -5 and 5. At daily resolution, there are large differences at the beginning of almost every station for all the synchronized groups, especially for the cluster2. At weekly resolution, the obvious differences at the beginning of synchronized years in cluster4 for all stations can be seen.

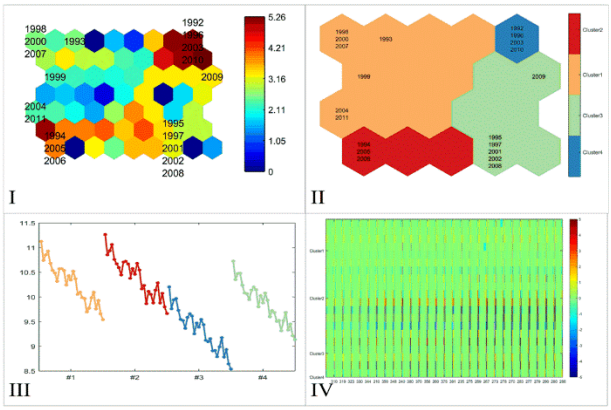
In summary, different temporal resolutions do affect temporal synchronization results. Also synchronization results at monthly resolution are most stable based on smaller value of differences in Anomalies Graph.

Year_based

a) Daily



b) Weekly



c) Monthly

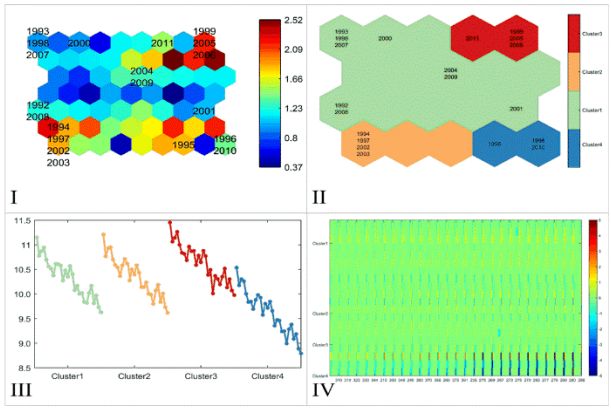


Figure 2.5: Synchronization results at daily (a), weekly (b) and monthly (c) temporal resolutions in year-based.

2.5.3 Temporal heterogeneity effects on synchronization

There are four clusters of daily synchronous years. Figure 2.6 displays spatial synchronization results using 20 years and spatial synchronization results using each group of synchronized years.

There are four, two, two, two clusters of synchronized stations using each group of synchronized years and four clusters of synchronized stations using all 20 years. The reduction of the number is obviously resulting from temporal homogeneity of synchronized years in temperature.

Stations in each cluster using each group of synchronized years and using 20 years are clearly different, displayed in cluster and geography maps. This is possibly due to the reduction in the number of clusters.

In summary, through comparison this chapter shows that spatial synchronization results are considerably different when considering temporal heterogeneity or not. Generally speaking, synchronization results are simplified when considering temporal heterogeneity.

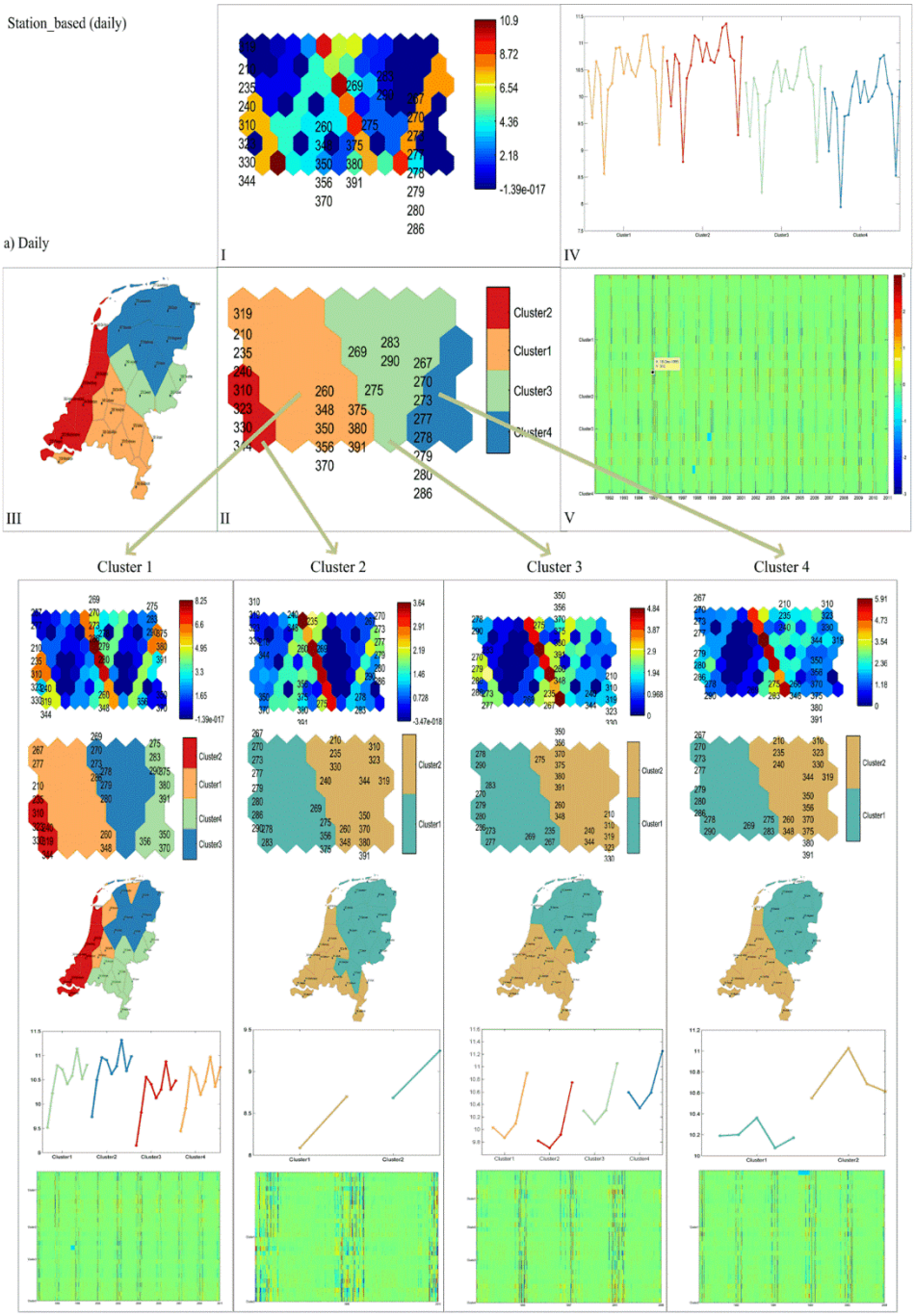


Figure 2.6: Spatial synchronization results using 20 years (Station-based daily) and spatial synchronization results using each group of daily synchronized years in four clusters.

2.6 Conclusions

This chapter has worked on an approach to visually discover synchronization of temperature records and to analyze the effects of changing the temporal resolution (MTUP) as well as the effects due to temporal heterogeneity. The analysis, which relied on self-organizing maps, was made from a spatial and from a temporal perspective: station-based where synchronous stations were discovered; year-based where temporal synchronization was explored using 'year' as basic unit. U-matrix maps were used to visually identify clusters of synchronous stations and years. Cluster maps were used to display clusters using colors. Geographic maps were used to consistently display the spatial distribution of the synchronous stations. Trend plots were used to analyze the temporal trends of synchronous stations and the spatial trends of synchronous years. Anomalies graphs were used to analyze differences between input data and corresponding clustering results. The analysis was repeated at multiple temporal resolutions from spatial and temporal perspective to study the effects of different temporal resolutions on the synchronization. Then spatial synchronization results for all years were compared with the results for groups of daily synchronous years to study the effects of temporal heterogeneity.

This chapter has shown that synchronization results are different at daily, weekly and monthly resolutions both in station-based and year-based. Temporal aggregation results in a reduction in the number of clusters and changes of stations in each cluster are observed in station-based analysis while only changes of years in each cluster were observed in the year-based analysis. Synchronization results at monthly resolution are seen as most stable both in station-based and year-based.

This chapter has also shown that spatial synchronization results are different when considering temporal heterogeneity. A reduction in the number of clusters and changes of stations in each cluster were observed when considering temporal heterogeneity.

These results show that the analytical approach is effective to visually discover spatio-temporal patterns in temperature data and to explore MTUP effects and temporal heterogeneity effects. Future work would add other weather parameters and explore the link between weather patterns and intra- and inter-annual phenological patterns at national to continental scales.

Chapter 3 Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data*

*This chapter is based on the paper: Wu, X., R. Zurita-Milla & M.-J. Kraak (2015). Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science*, 29**,** 624-642.

Abstract:

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 this chapter uses 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 one or more attributes observed at fixed geographical locations). In addition, this chapter presents 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 the study in this chapter, Dutch daily average temperature data collected at 28 weather stations from 1992 to 2011 was used. 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 three 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.

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

3.1 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 and Kraak 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 comes 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 analysis allows considering groups of 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 analysis 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 chapter only spatio-temporal clustering analysis of GTS is considered. GTS contain time evolving values for one or more observed attributes that are 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 methods can be found in literature (Crane and Hewitson 2003, Zhang, Huang, et al. 2003, Guo et al. 2006, Wu et al. 2008, Andrienko et al. 2010, Hagenauer and Helbich 2013, Wu et al. 2013). Crane and Hewitson (2003) clustered precipitation records of individual stations into regional datasets using self-organizing maps (SOMs) to analyze regional patterns. Zhang, Huang, 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. (2010) and Wu et al. (2013) all analyzed 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) analyzed spatio-temporal patterns in socio-economic data using SOMs in a hierarchical structure where spatial and temporal patterns are analyzed 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 and Andrienko 2010). 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 chapter 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, this chapter presents three geovisualization techniques: heatmaps, small multiples and ringmaps that support the exploration of the results of this co-clustering algorithm.

3.2 Study area and data

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

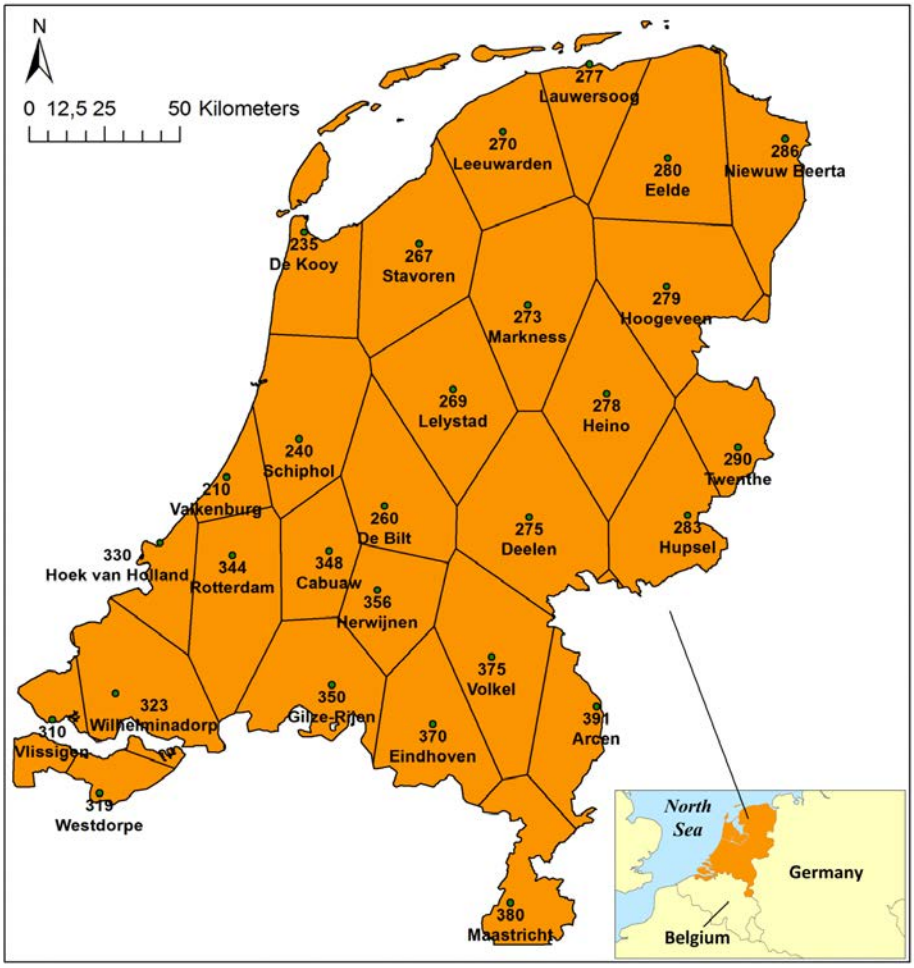


Figure 3.1: Location of the Netherlands in Europe (bottom right) and the Thiessen polygon map of the Netherlands to define influenced area of each station.

The Netherlands is located in the northwest of Europe (bottom right of Figure 3.1). 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 km2, Dutch weather is influenced by both maritime (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 dataset, 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 3.1 where each polygon is labelled by the station ID (e.g. 290) and its name (e.g. Twente).

3.3 Clustering methods

3.3.1 Clustering and co-clustering methods

GTS, such as the temperature series available for this chapter, can be organized in a 2D data matrix where rows indicate space (in this 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 methods. That is, methods 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 methods, however, treat objects and attributes equally (Han et al. 2011) 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 methods 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 methods (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 preserves a set of linear summary statistics (e.g. row/column and/or 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. This chapter choses the I-divergence because Banerjee et al. (2007) empirically proved its superiority. Besides, this chapter choses 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, this chapter call it BBAC_I as the I-divergence is used. The next section explains this co-clustering algorithm in more detail.

3.3.2 Bregman block average co-clustering algorithm with I-divergence (BBAC_I)

The BBAC_I algorithm enables the co-clustering of any 2D data matrix with positive and real valued elements that typically represent a joint probability distribution or co-occurrences between two random variables. The 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 $\{s_1,...,s_m\}$ and *T* in the timestamp sets $\{t_1,...,t_n\}$. Accordingly, the co-clustered data matrix is $O(\hat{S},\hat{T})$, where $\hat{S}$ take values in the station-cluster sets $\{\hat{s}_1,...,\hat{s}_k\}$ ($k \le m$) and $\hat{T}$ in the timestamp-cluster sets $\{\hat{t}_1,...,\hat{t}_l\}$ ($l \le n$).

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

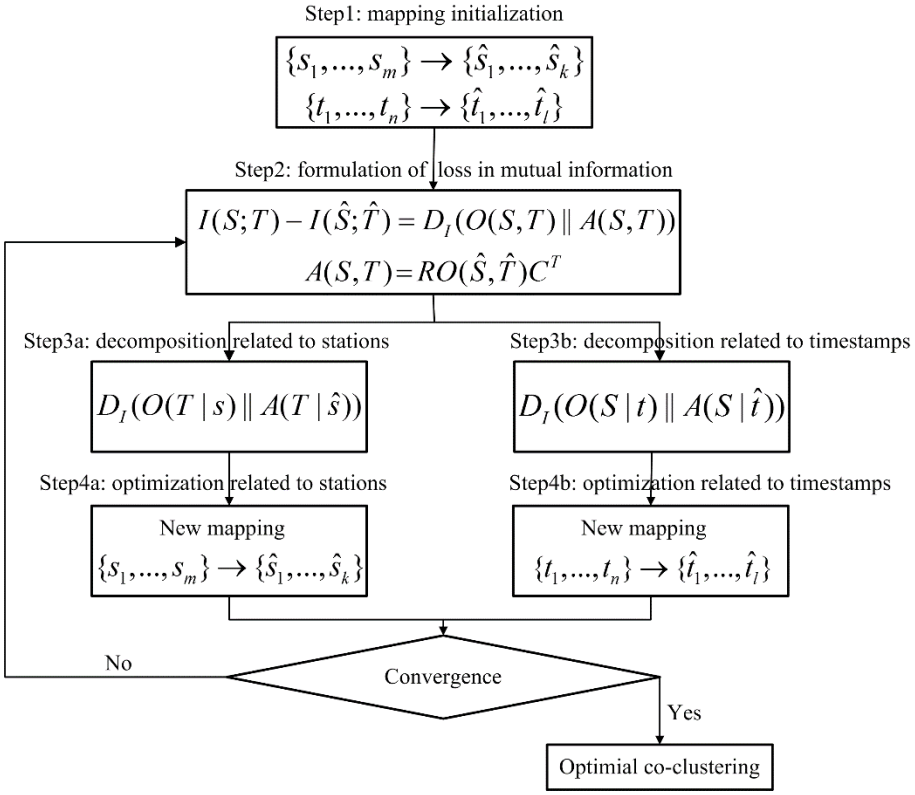


Figure 3.2: Summarized five steps of the Bregman block average co-clustering algorithm with I-divergence (BBAC_I).

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, the summarized five steps of this co-clustering algorithm is presented (Figure 3.2):

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:

$$I(S;T) - I(\hat{S};\hat{T}) = D_I(O(S,T) \| A(S,T)) \qquad (3.1)$$

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:

$$A(S,T) = RO(\hat{S},\hat{T})C^T \tag{3.2}$$

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 3.1 of loss in mutual information can be further represented as:

$$D_I(O(S,T) \| A(S,T)) = \sum_{\hat{s}} \sum_{\hat{t}} \sum_{s \in \hat{s}} \sum_{t \in \hat{t}} o(s,t) \log \frac{o(s,t)}{a(s,t)} \tag{3.3}$$

Step 3: to decompose equation 3.3 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 3.3 can be decomposed into the I-divergence of mapping from stations to station-clusters (step 3a):

$$D_I(O(T \mid s) \| A(T \mid \hat{s})) = \sum_{\hat{s}} \sum_{s \in \hat{s}} o(T \mid s) \log \frac{o(T \mid s)}{a(T \mid \hat{s})} \tag{3.4}$$

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

$$D_I(O(S \mid t) \| A(S \mid \hat{t})) = \sum_{\hat{t}} \sum_{t \in \hat{t}} o(S \mid t) \log \frac{o(S \mid t)}{a(S \mid \hat{t})} \tag{3.5}$$

Step 4: As aforementioned, the optimal co-clustering is to minimize the loss in mutual information. Now since equation 3.3 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 3.4 (step 4a):

$$i = \arg \min_{i \in \{1,\ldots,k\}} D_I(O(T \mid s) \| A(T \mid \hat{s}_i)) \tag{3.6}$$

And similarly find the new mapping of each timestamp to timestamp-clusters that minimizes equation 3.5 (step 4b):

$$j = \arg \min_{j \in \{1,\ldots,l\}} D_I(O(S \mid t) \| A(S \mid \hat{t}_j)) \tag{3.7}$$

Step 5: to re-calculate the loss of mutual information using the new mapping according to equation 3.3. 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 3.3 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.

In the following this chapter illustrates how BBAC_I yields the optimal co-clustering with an example. Suppose that this chapter wants to co-cluster the co-occurrence matrix between seven stations and five timestamps:

$$O(S,T)=\begin{bmatrix} 2.5 & 3.0 & 3.2 & 5.1 & 5.3 \\ 2.5 & 3.2 & 3.0 & 5.0 & 5.5 \\ 5.2 & 5.2 & 5.0 & 3.1 & 3.2 \\ 5.3 & 5.0 & 5.4 & 3.2 & 3.1 \\ 5.0 & 5.2 & 5.1 & 3.3 & 3.1 \\ 7.8 & 8.0 & 8.0 & 5.2 & 5.3 \\ 7.5 & 7.8 & 7.6 & 5.1 & 5.2 \end{bmatrix} \quad (3.8)$$

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

$$O(\hat{S},\hat{T}) = \begin{bmatrix} 2.9000 & 5.2250 \\ 5.1556 & 3.1667 \\ 7.7833 & 5.2000 \end{bmatrix} \quad (3.9)$$

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 3.1 shows how BBAC_I iteratively yields the local optimal co-clustering for the example matrix *O(S,T)* illustrated in equation 3.8. Each iteration in Table 3.1 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 3.9 and minimizes the loss in mutual information.

Table 3.1: BBAC_I in Figure 3.2 iteratively yields the optimal co-clustering for the example *O(S,T)* in equation 3.8.

$O(S,T)$

Step 1 & 2 of Figure 3.2

$A(S,T)$	$\hat{t}_2$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_2$	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	$\hat{s}_3$	4.4500	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			

Step 3 & 4 of Figure 3.2

$A(S,T)$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_1$	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			

Step 2 & 3 & 4 of Figure 3.2

$A(S,T)$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_2$	$O(\hat{S},\hat{T})$	$\hat{t}_1$	$\hat{t}_2$
$\hat{s}_1$	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			

3.4 Geovisualization techniques

In order to fully exploit co-clustering results, appropriate geovisualization techniques must be used. Here this chapter relies 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.

3.4.1 Heatmap

A heatmap is a graphical representation of a data matrix where the individual values in the matrix are visualized using a range of colors. 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-color scale provides a rapid view of all the data.

3.4.2 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 one color 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 color of each map interpreted from average temperature values in that timestamp-cluster indicates which timestamp-cluster the timestamp belongs to.

3.4.3 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 colors. The colors are interpreted from the average temperature value of timestamps in each timestamp-cluster for all stations.

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

The BBAC_I algorithm described in Section 3.3.2 was used to hierarchically analyze the spatio-temporal patterns in the Dutch temperature series at yearly, monthly and daily resolutions.

This was done using the workflow illustrated in Figure 3.3. First, the daily temperature data was averaged to create a yearly temperature dataset, which was organized as a 28 (stations) by 20 (years) data matrix. 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 dataset was created by calculating the, per station, average monthly temperature using the temperature records of all the years belonging to

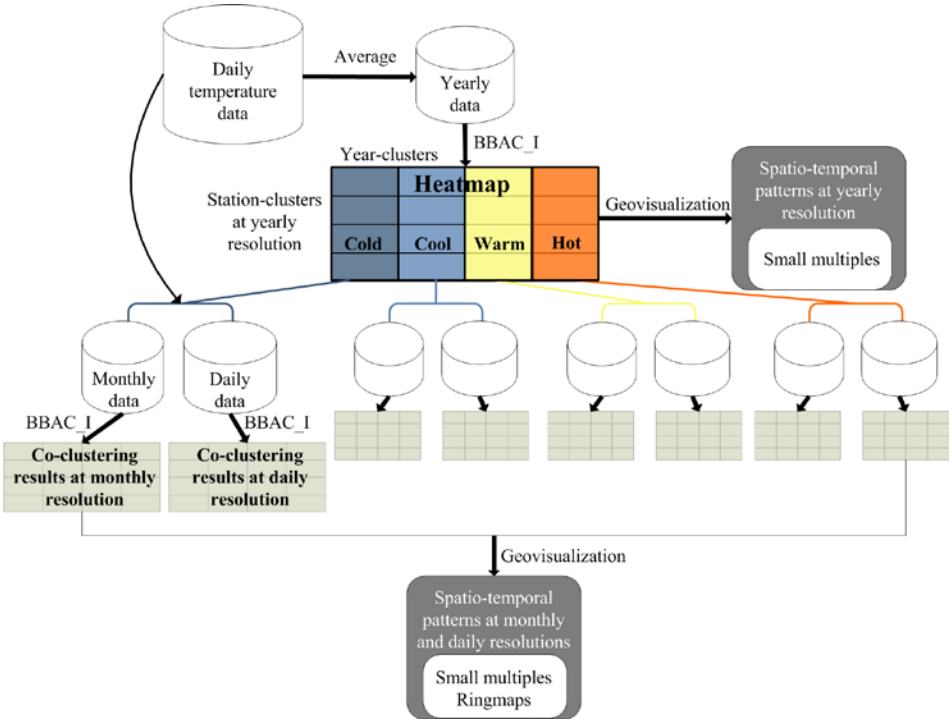


Figure 3.3: Workflow of the experiment at different temporal resolutions.

that year-cluster. This resulted in four monthly temperature matrices of 28 (stations) by 12 (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 (stations) by 365 (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 methods, the selection, optimization and evaluation of cluster numbers in co-clustering ones remains challenging and depends on specific applications. In this chapter, 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 four 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 chapter; 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 chapter was downloaded from http://www.ideal.ece.utexas.edu/software/bregcc_code.tar.

3.6 Results and discussions

3.6.1 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.

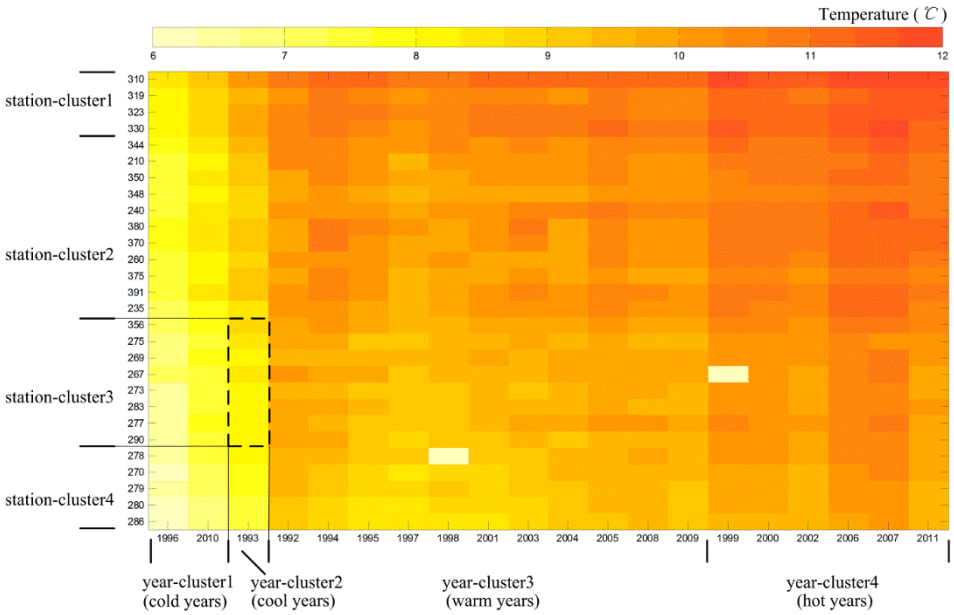


Figure 3.4: Heatmap to visualize the co-clustering results at yearly resolution, where light yellow indicates low temperature and red means high temperature. x-axis: years arranged according to year-clusters with increasing temperature values; y-axis: stations ordered according to station-clusters with 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 heatmap in Figure 3.4 straightforwardly displays the station-clusters, year-clusters as well as station-year co-clusters in the results. Small multiples in Figure 3.5 show the spatial distribution of station-clusters (Figure 3.5a) and temporal distribution of year-clusters (Figure 3.5b). Besides, the small multiples in Figure 3.6 illustrate the spatial patterns of station-year co-clusters.

The re-ordered yearly temperature matrix is showed in Figure 3.4 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 there are 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 3.1) 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 3.5a shows the spatial distribution from station-cluster1 to station-cluster4: the colored area indicates the region of the station-cluster and the color symbolizes the average temperature values in that

station-cluster. The higher the temperature, the darker the color. Figure 3.5a clearly shows that the Netherlands is divided into four regions from left to right: southwest, center-southwest, center-northeast and northeast. These four regions reveal a decreasing temperature pattern across the country, which confirms the expected temperature patterns: maritime (milder) in the southwest and more continental in the northeast. Figure 3.5a also shows almost all stations in each station-cluster are geographically adjacent. Each map of the small multiples in Figure 3.5b shows the year-cluster to which that particular year belongs to. This time, the color represents the average temperature values in that year-cluster. Figure 3.5b shows that yearly temperature in the Netherlands from 1992 to 2011 are mostly "warm" or "hot", with only three 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.

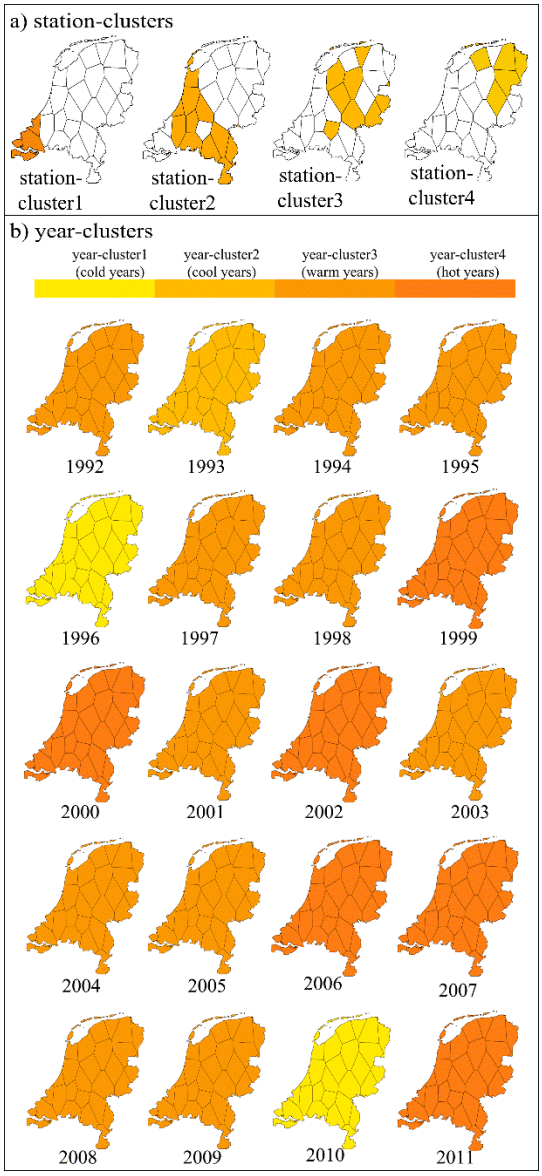


Figure 3.5: Small multiples to display spatial distribution of station-clusters and temporal distribution of year-clusters. a) Small multiples to show the spatial distribution of four station-clusters: within each map the colored area indicates the region of the station-cluster. The higher is the average temperature values of the station-cluster, the darker the color 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 color is.

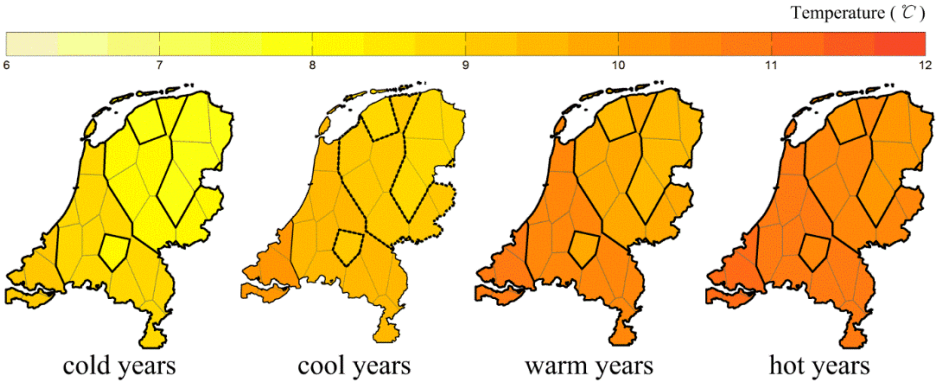


Figure 3.6: Small multiples to display spatial patterns in station-year co-clusters for each year-cluster, where light yellow indicates low temperature and red means high temperature; Region within dashed boundary lines in the second map corresponds with the co-cluster within the dashed rectangle in Figure 3.4.

Since the elements mapped to each station- and year-cluster are clearly illustrated in the heatmap, Figure 3.4 also displays the station-year co-clusters straightforwardly. For instance, the dashed rectangle in the bottom left of Figure 3.4 shows one of the sixteen 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" 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 3.6 displays the spatial patterns of station-year co-clusters shown in Figure 3.4: 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, center-southwest, center-northeast and northeast are indicated by thick lines. Each of regions corresponds to each of the station-year co-clusters shown in Figure 3.4 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 center-northeast region bordered with a dashed line in the "cool" year corresponds with the dashed rectangle displayed in the heatmap in Figure 3.4.

Moreover, the heatmap in Figure 3.4 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.

3.6.2 Spatial-temporal patterns at monthly and daily resolutions

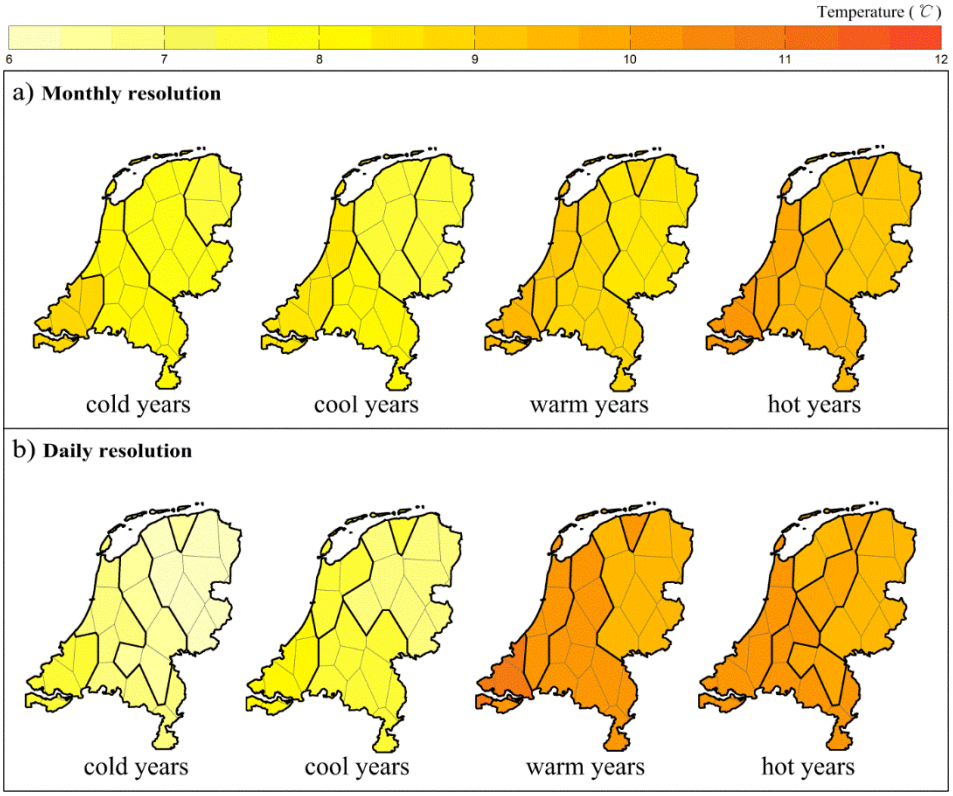


Figure 3.7: Small multiples to display spatial patterns for station-month and station-day co-clusters. a) Small multiples display spatial patterns in station-month co-clusters for year-clusters; b) Small multiples from display spatial patterns in station-day co-clusters for year-clusters. Light yellow indicates low temperature and red means high temperature.

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 by 12) temperature matrices were mapped to one set of 4 by 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 by 365) were mapped to one set of 4 by 4 station-day co-clusters and in all there were also four sets associated to “cold”, “cool”, “warm” and “hot” years.

Figure 3.7 displays the spatial patterns via small multiples for the four sets of station-month co-clusters (Figure 3.7a) and four sets of station-day co-clusters(Figure 3.7b) and with the same color scheme used for the yearly resolution (Figure 3.6).

In Figure 3.7a 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 3.7a shows the decreasing temperature patterns from southwest to northeast for each year-cluster.

In Figure 3.7b 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 3.7b shows the same decreasing temperature patterns from southwest to northeast. It can be noticed that the temperature in “warm” and “hot” years are similar. I suppose this is because the representative daily data for “hot” years has more variations than those for “warm” years.

Figure 3.8 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 color scheme of this figure extends the one used previously with blue 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 twelve months clockwise from January to December and the color 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.

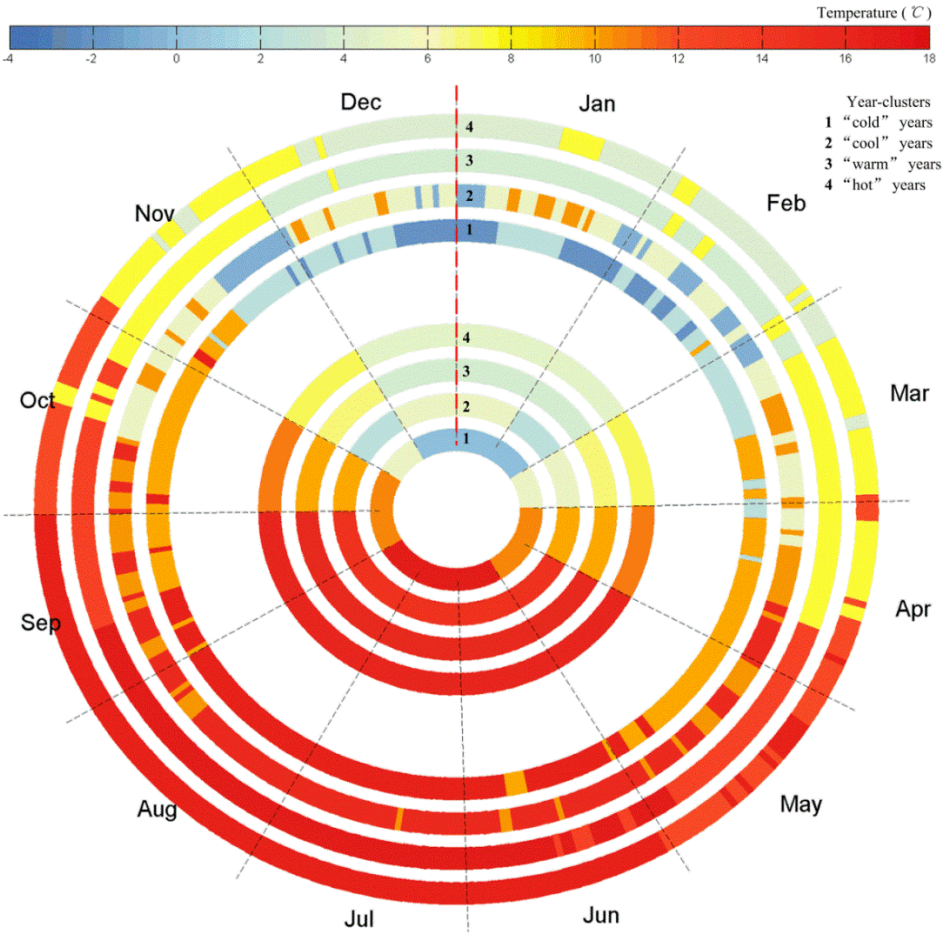


Figure 3.8: Ringmap to show temporal patterns for station-month co-clusters and station-day co-clusters. The inner four circles display temporal patterns for the four sets of station-month co-clusters for year-clusters; The outer four circles displays temporal patterns for the four sets of station-day co-clusters for year-clusters. Blue means very low temperature; light yellow indicates low temperature and red means high temperature.

Combining the ringmap with the small multiples in Figure 3.7a, it is seen 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 3.8 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 1st January to 31st December and the color 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 several anomalous days are further noticed, 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 3.7b, it can be observed 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 Figures 3.7a and 3.7b 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. It is suggested 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 3.8. With respect to the temporal patterns, the inner four and outer four circles of Figure 3.8 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.

3.7 Conclusions

This chapter introduced the use of the BBAC_I to analyze 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, this chapter 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 Modifiable temporal unit problem (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, further work in the following areas will be anticipated: (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 3.4, 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 attribute – yearly, monthly and daily averaged temperatures in this chapter. Future work is needed to adapt the BBAC_I to allow the co-clustering of more attributes; (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.

Chapter 4 A novel analysis of spring phenological patterns over Europe based on co-clustering*

*This chapter is based on the manuscript: Wu, X., R. Zurita-Milla & M.-J. Kraak (2016). A novel analysis of spring phenological patterns over Europe based on co-clustering. *Journal of Geophysical Research: Biogeosciences.* In press.

Abstract:

The study of phenological patterns and their dynamics provides insights into the impacts of climate change on terrestrial ecosystems. Here this chapter presents a novel analytical approach, based on a co-clustering method, which enables the concurrent study of spatio-temporal patterns in spring phenology. The approach is illustrated with a long-term time series of first leaf dates (*FLD*) over Europe, northern Africa and Turkey calculated using the extended spring index models and the European E-OBS daily maximum and minimum temperature datasets (1950 to 2011 with a spatial resolution of 0.25 degrees).This *FLD* dataset was co-clustered using the Bregman block average co-clustering with I-divergence (BBAC_I) and the results were refined using *k*-means. These refined co-clusters were mapped to provide a first spatially-continuous delineation of phenoregions in Europe. The results show that the study area exhibits four main spatial phenological patterns of spring onset. The temporal dynamics of these phenological patterns indicate that the first years of the study period tend to have late spring onsets and the recent years have early spring onsets. The results also show that the study period exhibits twelve main temporal phenological patterns of spring onset. The spatial distributions of these temporal phenological patterns show that western Turkey tends to have the most variable spring onsets. Changes in the boundaries of other phenoregions can also be observed. These results indicate that this co-clustering based analytical approach effectively enables the simultaneous study of both spatial patterns and their temporal dynamics and of temporal patterns and their spatial dynamics in spring phenology.

Key words: co-clustering, Europe, extended spring index models, spatio-temporal patterns, spring phenology

4.1 Introduction

The spatio-temporal inhomogeneity of climate change is well documented (IPCC 2013). For instance in Europe, the increase of surface air temperature in high latitudes is different than that in other places (Haylock et al. 2008, EEA 2012) and the decadal increase of average temperature over the period 2002-2011 is higher than over the period 1850-1899 (Brohan et al. 2006, Smith et al. 2008, Hansen et al. 2010). As a result, the impacts of climate change on terrestrial ecosystems are also inhomogeneous across space and time (Menzel et al. 2006). In this respect, it is important to study dynamics in phenological patterns from both spatial and temporal dimensions as they provide insights into the inhomogeneous impacts of climate change on terrestrial ecosystems (Walther et al. 2002, Parmesan and Yohe 2003).

Phenology is the science that deals with the study of life cycle phases in plants and animals driven by environmental factors (Lieth 1974, Schwartz and Chen 2002). Several studies have demonstrated that plant phenology is one of the most reliable bioindicators of climate change (Menzel et al. 2006, Schwartz et al. 2006, Gordo and Sanz 2010). Phenological spring events (e.g. first leaf appearance) are reported to be more sensitive to climate change than events occurring in other seasons (Matsumoto et al. 2003, Doi and Katano 2008, Gordo and Sanz 2010). Consequently, indices that measure the onset of spring are ideal biological indicators of climatic variability in space and time (Schwartz 1998, Schwartz et al. 2006).

Several methods are available to monitor spring phenology. One of such methods relies on time series of remotely sensed satellite images to derive the so-called start-of-spring (SOS). However, there is no universally accepted method to characterize SOS from satellite data and results depend on the method and/or sensor used (Schwartz et al. 2002, White et al. 2009). Besides, satellite data is only available since the 1980s and thus it is insufficient for studying long-term phenological responses to climate change. Another method uses ground phenological observations of selected species and biological events (e.g. leafing or blooming). Long-term phenological records exist but they exhibit a poor spatial coverage (White et al. 2005, Menzel et al. 2006, Studer et al. 2007). A third method relies on the use of phenological models. Often these models are calibrated using ground phenological observations and allow predicting spring onset in regions where no observations exist. This chapter uses the extended spring index models (Schwartz et al. 2013; Section 4.2.2) to consistently characterize spring onset over large areas and for long time periods.

One of the most popular methods to explore patterns in spatio-temporal data is clustering analysis, which identifies groups of similar data elements and enables analysts to consider them at a higher level of abstraction (Andrienko et al. 2009). In environmental studies, clustering is particularly useful because regional analysis for a given time period (e.g. a season) is generally more informative than analyzing a collection of observations made at a few locations and timestamps (Zirlewagen and Von Wilpert 2010). Thus, several studies can be found on the analysis of phenological patterns using clustering methods (Ahas and Aasa 2003, White et al. 2005, Ahas et al. 2007, Gu et al. 2010, Kumar et al. 2011, Mills et al. 2011, Zhang et al. 2012, Zurita-Milla et al. 2013). Ahas and Aasa (2003) used a clustering method to group phenological events by years to identify early and late years in terms of seasonal rhythm. Ahas et al. (2007) used clustering analysis to group years in terms of human activities' rhythm to characterize urban and rural dynamics. White et al. (2005) used an iterative *k*-means clustering algorithm to identify phenologically and climatically similar clusters in space, which they termed phenoregions. Based on this latter work, Kumar et al. (2011) and Mills et al. (2011) developed and implemented a parallel *k*-means clustering algorithm to identify phenoregions in conterminous United States (CONUS). Gu et al. (2010) applied a widely used clustering method, ISODATA, to generate another map of phenol regions for CONUS. Zhang et al. (2012) combined principal component analysis (PCA) and *k*-means++ to generate a phenol regions map for a smaller geographical region (the Upper Colorado River Basin) and Zurita-Milla et al. (2013), used self-organizing maps to cluster time series of satellite data to identify the main phenological patterns in the Kruger national park (South Africa).

All these phenological clustering studies can be categorized into two types (refer to Figure 4.1): (1) spatial clustering for spatial pattern analysis (Figure 4.1b); and (2) temporal clustering for temporal pattern analysis (Figure 4.1c). In spatial clustering, the locations are regarded as objects and each timestamp as an attribute. Clustering results are locations with similar values of phenological variable(s) along all timestamps (the thick rectangle in Figure 4.1b is an example of a spatial cluster). In the temporal clustering, the time stamps are regarded as objects and each location as an attribute. Clustering results are time stamps with similar values of phenological variable(s) along all locations (the thick rectangle in Figure 4.1c is an example of a temporal cluster). However, patterns identified by only using spatial clustering lack the ability to describe the time-varying behaviour present in the data and vice versa (Deng et al. 2011). For instance in Figure 4.1b values in location-cluster1 are more similar in timestamp-1 and

timestamp-2 rather than along all timestamps while in Figure 4.1c values in timestamp-cluster1 are more similar in location-1 and location-2 rather than along all locations. This deficiency demands a clustering method that allows the simultaneous discovery of spatial and temporal patterns. Co-clustering enables this type of analysis.

Co-clustering methods, unlike one-way clustering ones such as *k*-means, treats locations and timestamps equally (Han et al. 2011). This is done by mapping locations to location-clusters and timestamps to timestamp-clusters at the same time (Figure 4.1d). Co-clustering identifies homogeneous spatio-temporal co-clusters (co-clusters for short), formed by intersecting each location- and timestamp-cluster. The values of phenological variables in timestamps at locations that belong to the same co-cluster are similar. The thick rectangle in Figure 4.1d shows an example of a co-cluster.

Each co-cluster contains elements that are similar to each other. However, as pointed out by Wu et al. (2015), the values in different co-clusters might still be similar due to the need to arbitrarily predefine the number of location- and timestamp-clusters before applying the algorithm to the dataset. Also, co-clustering methods assign complete rows/columns to location- and year- clusters. Consequently, complex spatio-temporal patterns might not be fully captured. To deal with this, co-clustering could be run with a relatively large number of co-clusters that could be subsequently grouped to create axis-parallel irregular (i.e. non-rectangular) co-clusters.

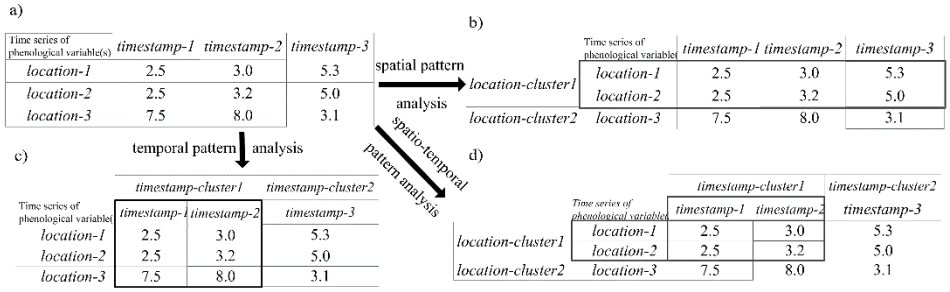


Figure 4.1: Spatial or/and temporal clustering analysis in phenological dataset. a) time series of phenological variable(s); b) location-clusters resulting from spatial clustering analysis; the thick rectangle is an example of a spatial cluster; c) timestamp-clusters resulting from temporal clustering analysis; the thick rectangle is an example of a temporal cluster; d) spatio-temporal co-clusters, co-clusters for short, resulting from spatio-temporal clustering analysis; the thick rectangle is an example of a co-cluster intersected by a location-cluster and a timestamp-cluster.

Fully considering the above listed issues, this chapter proposes a novel analytical approach based on a co-clustering method that enables the exhaustive analysis of complex spatial and temporal phenological patterns. Briefly, this chapter first generates time series of date of first leaf using the extended spring index models. Then, this chapter co-clusters the time series and refines the results using *k*-means and, finally, maps the results to reveal the main spring phenological patterns as well as their dynamics.

4.2 Materials and methods

4.2.1 Materials

The European E-OBS dataset (v10.0; http://www.ecad.eu/download/ensembles/download.php; Haylock et al. 2008), which is a product of the European Climate Assessment and Dataset (ECA&D) project, is used in this chapter. This gridded dataset runs from 1950 to 2013 and covers the whole Europe as well as Northern Africa and Turkey. This chapter downloaded geo-referenced daily maximum (TX) and minimum (TN) temperatures with a spatial resolution of 0.25°×0.25°.

The completeness of TX and TN was examined as the extended spring index models (section 4.2.2) require daily temperature records and will not calculate the date of first leaf if there are more than 20 missing TX or TN values per month

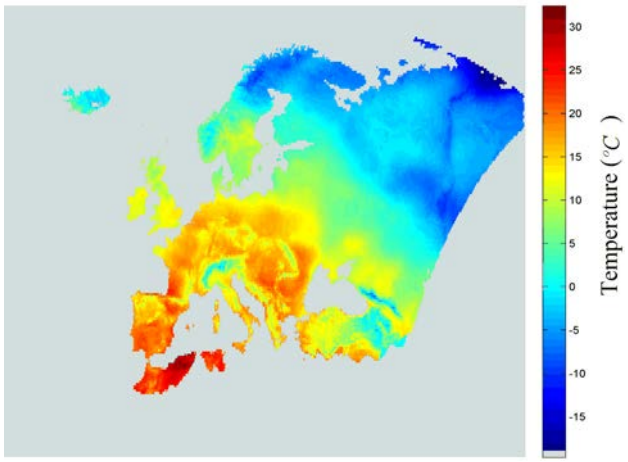


Figure 4.2: The spatial extent of the filtered dataset used in this chapter, depicted by the spatial distribution of the maximum temperature for an arbitrary day (21st March 2010) as an example.

for the first six months of the year. Thus these records were excluded. Also, grid cells for which the spring phenological index could not be calculated for more than 20 years were excluded from further analysis. The application of the criteria resulted into a new dataset that contains 28225 grid cells (98.5% of all the cells covered by the original E-OBS dataset) covering the period 1950 to 2011 (62 out of 64 years). The spatial extent of the filtered dataset is depicted in Figure 4.2 that, as an example, shows the spatial distribution of TX on an arbitrary day (21st March 2010).

4.2.2 Extended spring index models

The so-called spring indices (SI) are a suite of regression-based models that can be used to characterize spring onset (Schwartz 1997, Schwartz and Reiter 2000, Schwartz et al. 2006). These models predict the first leaf dates (*FLD*) and first bloom dates (*FBD*) for three key species: lilac (*Syringa chinesis "Red Rothomagensis"*), honeysuckles (*Lonicera tatarica "Arnold Red" and Lonicera. Korolkowii "Zabeli"*). The model inputs are daily maximum and minimum temperatures for a site as well as its latitudinal information – used as a proxy for day length. The SI have been extensively validated in the Northern Hemisphere to generate reliable phenological timings for the above mentioned species (Schwartz et al. 2006). However, the original SI can only produce outputs for locations where the chilling and warmth requirements are satisfied (Schwartz et al. 2013). This is a typical characteristic of sequential phenological models (Chuine 2000) where the model should first accumulate chilling degrees and then start accumulating temperature to predict leafing, blooming or other phenological phases. To be able to cover large areas, here this chapter uses the recently developed extended spring indices (SI-x) where chilling is not required anymore (Schwartz et al. 2013). The SI-x models have been validated in continental United States with extensive ground phenological observations and they were found as effective as the original SI models (Schwartz et al. 2013). Thus, the SI-x must perform well in Europe too, given the fact that the SI are already validated in this continent (Schwartz et al. 2006). Besides, the three SI species are also widely present in Europe and monitored by Europe-wide programs, such as the European Phenology Network (van Vliet et al. 2003) and the International Phenological Gardens in Europe (Menzel et al. 2006).

This chapter focuses on the *FLD* model output as it is the earliest spring bioindicator. This index from the SI-x models refers to the average of the first leaf dates of aforementioned three key species. Nevertheless, *FLD* has been proved to have a larger scope and found relevant also for the general phenological

onset of grasses and shrubs growth (Schwartz and Chen 2002, Schwartz et al. 2006), and even fruit trees (Schwartz et al. 2013). The *FLD* is calculated as follows:

$$FLD_{ij} = SI - x(TX_{ij}, TN_{ij}, lat_i), [i]_1^{28225} and [j]_1^{62} \quad (4.1)$$

Where TX is the daily maximum temperature, TN the daily minimum temperature and *lat* represents the latitude of the location. The index *i* symbolizes the locations (i.e. the 28 225 grid cells in the study area) and *j* the years (62 years). The TX and TN are used to accumulate degree-days and synoptic events (i.e. warm peaks in temperature) from the first of January of each year. Notice that TX and TN need to be transformed from Celsius to Fahrenheit before passing them to the model because the SI-x was calibrated using this unit. Latitudinal information is used in this model as a proxy for day length to account for photoperiodic changes (Basler and Körner 2012). For a detailed explanation of the SI-x models and for the code to calculate the *FLD* please see (Ault et al. 2015).

4.2.3 Co-clustering analysis

Co-clustering methods have been used for pattern analysis in many fields (Dhillon et al. 2003, Cho et al. 2004, Banerjee et al. 2007, Wu et al. 2015). Dhillon et al. (2003) proposed the information theoretic co-clustering (ITCC) algorithm, which uses the I-divergence metric to find optimum co-clusters in the input 2D matrix while preserving the row, column and co-cluster averages during the process. They applied the ITCC for word-document analysis to find groups of inter-related documents and words. Cho et al. (2004) proposed the minimum sum-squared residual co-clustering (MSRC) algorithm, which employs the squared Euclidean distance for optimization, to obtain subsets of genes and conditions with similar expression values. Banerjee et al. (2007) developed the so-called Bregman co-clustering algorithm, which is a meta co-clustering algorithm with the above two algorithms as special cases. In their application of the Bregman co-clustering for word-document analysis, Banerjee and colleagues empirically proved the superiority of the I-divergence metric. Recently, Wu et al. (2015) applied the Bregman block average co-clustering algorithm with I-divergence (BBAC_I), which is a special case of the Bregman co-clustering that preserves co-cluster averages, to time series of temperature values and successfully identified observations with similar temperatures along both the spatial and the temporal dimensions. This latter algorithm is also employed in this chapter to identify co-clusters with similar *FLD* values along locations and years. The BBAC_I algorithm allows to co-cluster positive data matrices with real-

Algorithm Bregman block average co-clustering algorithm with I-divergence (BBAC_I)

Input: O_{FLD}, *lc (number of location-clusters), yc (number of year-clusters)*
Output: optimized lc× yc co-clusters
Begin

1. *Initialization: random mapping from l locations to lc location- clusters and y years to yc year-clusters;*
2. *Calculation of the loss function*

$$f_{loss} = D_I(O_{FLD} \| \hat{O}_{FLD})$$

3. *Iterations:*

Begin

3.1 update mapping from locations to location-clusters

$$i = \arg\min_{i\in\{1,\ldots,lc\}} D_I(O_{FLD} \| \hat{O}_{FLD})$$

3.2 update mapping from years to year-clusters

$$j = \arg\min_{j\in\{1,\ldots,yc\}} D_I(O_{FLD} \| \hat{O}_{FLD})$$

End

Until convergence

End

Figure 4.3: Pseudocode for Bregman block average co-clustering algorithm with I-divergence (BBAC_I).

valued elements which represent co-occurrences or joint probability between two random variables (Wu et al. 2015). It regards co-clustering as an optimization issue in information theory where mutual information measures the amount of shared information between two variables, and yields the optimal co-clusters by minimizing the loss in mutual information between the original data matrix and the co-clustered one.

The *FLD* data can be regarded as a co-occurrence matrix (O_{FLD}) between a spatial variable taking values in 28225 locations and a temporal variable taking values in 62 years. The pseudocode of the BBAC_I algorithm (Figure 4.3) illustrates the iterative process to optimize the partitions of the *FLD* data matrix to the co-clusters. The first step of the algorithm is to perform an initial random mapping of the locations to location-clusters and of the years to year-clusters. This produces the co-clustered matrix ($\hat{O}_{FLD}$). Then, in the second step, the loss in mutual information between the original and the co-clustered matrices is measured by applying the equation listed in this step, where $D_I(\cdot \| \cdot)$ denotes the I-divergence between two matrices. In the third step, the algorithm starts an iterative process to update the mapping from locations to location-clusters and years to year-clusters to minimize the loss function. This function has been

proven to be monotonically decreasing after each iteration (Banerjee et al. 2007). The iterations cease when the loss function converges to a local minimum, i.e. the change in the loss is below a predefined threshold. For the detailed description of the BBAC_I, please see (Wu et al. 2015). Since a global minimum cannot be guaranteed by the local minimum (Dhillon et al. 2003), thus, in practice, the co-clustering process is repeated with several random mappings to find the optimal local minimum. The optimal co-clustering results are consequently yielded. Finally, the rows and columns of the *FLD* matrix are re-ordered so that locations/years belonging the same location-/year-cluster are arranged together. In particular, the co-clusters are arranged according to their average *FLD* values: low *FLD* values are positioned at the bottom left and high *FLD* values at the top right corner of the re-ordered matrix.

Re-ordered matrices typically exhibit a check board pattern because of the assignment of full rows and columns to clusters in co-clustering methods. To better capture the spatio-temporal patterns in the FLD dataset, the well-known k-means algorithm was used to regroup the results into k axis-parallel non-rectangular co-clusters, also named irregular co-clusters. The mean and variance of the FLD values in each regular co-cluster were used as input features for k-means and the number of clusters (i.e. k) was optimized using the Silhouette method (Rousseeuw 1987). Since k-means clustering is NP-hard and only guarantees the local optimal result, this process is repeated several times to obtain the optimal local solution.

Since the two-step clustering (first BBAC_I and then k-means) used in the co-clustering analysis contains both local optimization algorithms, the stability of the clustering results must be considered. Especially considering that the BBAC_I results feed into k-means as input features, multiple trials of each clustering process are essential to provide the stable final results.

4.2.4 Spatio-temporal phenological patterns

The main spatio-temporal phenological patterns in the *FLD* dataset were explored by analyzing the irregular co-clusters from a spatial and a temporal perspective. For this, the chapter visualized the irregular co-clusters using three sets of small multiples (i.e. a series of geographic maps) and two sets of linear timelines (i.e. a linear line that visualizes events in a chronological order). The first set of small multiples was used to show the spatial distribution of each irregular co-cluster or phenoregions, which indicate self-similar clusters in terms of *FLD* in this chapter. The second set of small multiples was used to display the unique spatial patterns found in the study area (one map per pattern). These

spatial patterns were extracted from the irregular co-clusters by combining grid cells with the same variation over the study area. In other words, these spatial patterns were created by combining phenoregions falling into the same years. A linear timeline was used to show the temporal dynamics of these spatial patterns over the whole study period. The percentage of each phenol region per spatial pattern as well as the number of years that exhibit each spatial pattern were also calculated to characterize the main spatial patterns and supports the study of their temporal dynamics. The third set of small multiples was used to display the spatial extent of unique temporal patterns found during the whole study period. These temporal patterns were extracted from the irregular co-clusters by combining years with the same variation along the study period and arranging them chronologically. Each timeline is used to show each temporal pattern and thus the number of timelines is equal to the number of unique temporal patterns.

4.3 Results

4.3.1 First leaf dates (*FLD*)

The *FLD* values were calculated for all valid grid cells and years in the E-OBS temperature datasets. As an example, Figure 4.4 shows their spatial distribution over the study area in 2010. The greener the color, the earlier the *FLD*. As expected, the *FLD* increases from the south to north of Europe. The range of variation is fairly wide with very early spring dates (end of January and beginning of February) for Southern Europe, Turkey and northern Africa. Late to very late spring dates (end of June and the start of July) occur in Scandinavia and northern Russia as well as in a few “cold islands” corresponding to the Alps and the Caucasus Mountains. It is important to note that although the spatial patterns of *FLD* (Figure 4.4) resemble those of TX (Figure 4.2), these patterns are fundamentally different. Not only because the former displays the spring phenological pattern of one year while the latter shows the patterns of a particular day in that year but also because the SI-x model captures non-linearity in the accumulation of temperature. Therefore the *FLD* is more than an accumulation of degree days.

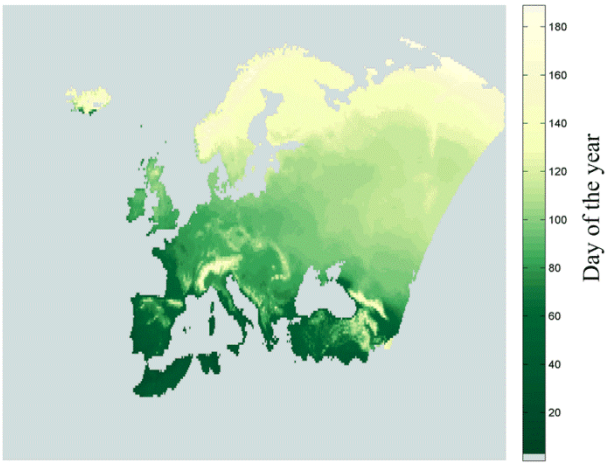


Figure 4.4: The spatial distribution of the first leaf dates over Europe in 2010. The greener the color, the earlier the first leaf date.

4.3.2 Regular and irregular co-clusters

The BBAC_I algorithm was applied to the *FLD* matrix, which was created by reorganizing the input *FLD* maps so that all valid locations appear as rows and the yearly *FLD* values as columns. The number of year-clusters was set to four after testing values from four to ten in the step of one because the BBAC_I algorithm always returned this number of year-clusters. This means that the loss function of BBAC_I achieves its minimum with this value for this particular dataset. The number of location-clusters was set to 45 after testing values in the range 20 to 60 with an interval of five. Again the reason for choosing this value is that it minimizes the value of the loss function. This chapter empirically observed that setting the threshold for convergence of the loss function to 10^{-6} and the number of iterations to 2000 was sufficient to guarantee stable results. Finally, the number of random mapping initializations was set to 200 to help find a global minimum of the loss function.

This BBAC_I parametrization produced the 180 (45 × 4) regular co-clusters displayed as a heatmap in Figure 4.5. The values of x-axis are the 62 years covered by the dataset arranged according to their membership to the year-clusters, which are sorted from early to late *FLD* values. Most of the years belonging to year-cluster1 are recent years whereas the ones in year-cluster4 correspond to the first years of the study period. The values of y-axis are the 28 225 grid cells arranged from location-cluster1 to location-cluster45 from top to bottom. The higher the number of location-cluster the earlier the *FLD*. Thus, based on the arrangement of location- and year-clusters along the axes, the regular

co-cluster location-cluster45/year-cluster1 has the earliest *FLD* values while the regular co-cluster location-cluster1/year-cluster4 has the latest *FLD* values. Figure 4.5 also shows several co-clusters with almost the same *FLD* values. For example, location-cluster 45/year-cluster1 and location-cluster45/year-cluster2. This problem, which tends to occur with any co-clustering methods where the number of clusters is specified a priori, requires the refinement of the co-clusters. Here this is done by re-grouping the co-clusters using *k*-means. To do this, the mean and the variance of each co-cluster were calculated and used as input features for *k*-means (Figure 4.6). As expected, the mean of *FLD* values increases from the first to the last co-cluster. However, the *FLD* variance presents many fluctuations, with a peak at the 115th co-cluster. The most probable reason for these fluctuations are anomalous TX and/or TN values in the original dataset. To cope with this, this chapter named one of *k* irregular co-clusters as "abnormal".

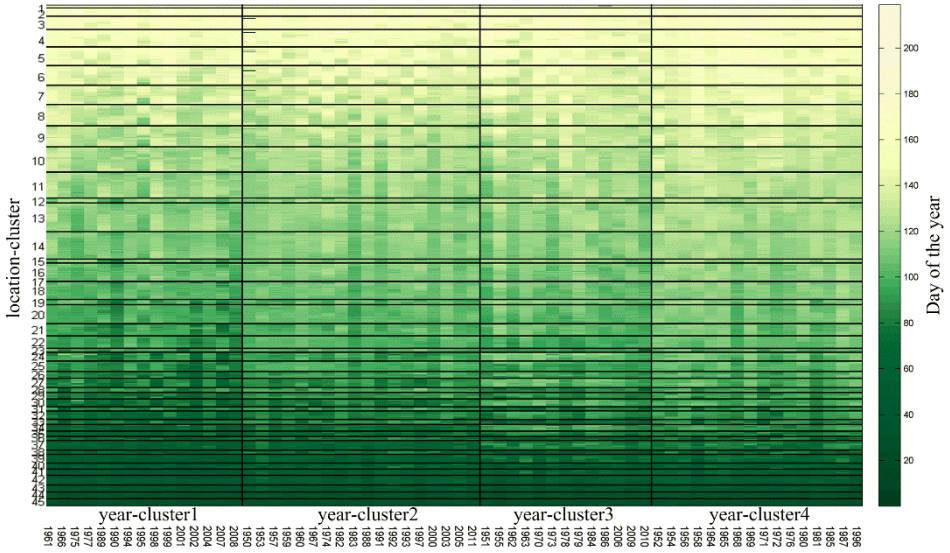


Figure 4.5: 180 (45 × 4) regular *FLD* co-clusters, intersected by each location-cluster and year-cluster and indicated with thick lines. The greener the color, the earlier the date. X-axis from left to right: 62 years arranged according to the order from year-cluster1 to year-cluster4 with increasingly late *FLD*; Y-axis from bottom to top: 28 225 grids arranged according to the descending sequence from location-cluster45 to location-cluster1 with increasing late *FLD*.

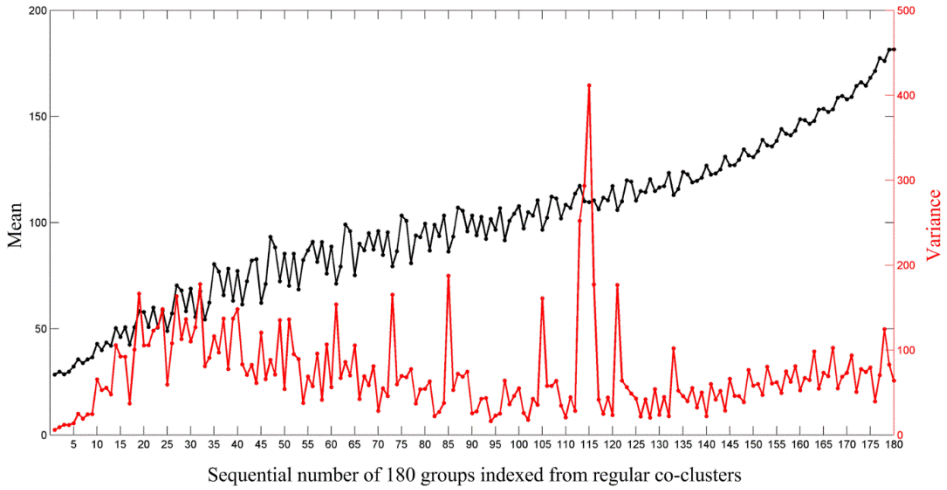


Figure 4.6: The mean and variance of the *FLD* values within each of the 180 (45 × 5) regular co-clusters. The number is indexed from location-cluster45 to location1 and from year-cluster1 to year-cluster4.The left y-axis shows the mean of *FLD* values and the right y-axis show the variance of *FLD* values.

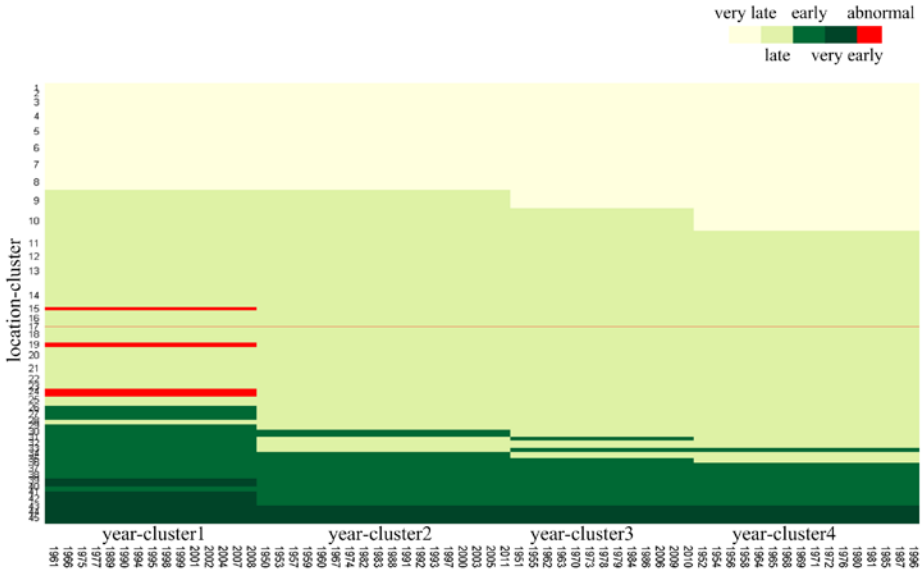


Figure 4.7: The five irregular *FLD* co-clusters named “very late”, “late”, “early” “very early” and “abnormal”. The composition of irregular co-clusters from regular ones is also displayed in this heatmap by preserving the co-cluster arrangement in Figure 4.5.

The value of *k* was optimized by testing values from three to ten in the step of one and using the Silhouette method. Results showed that the optimum number of irregular co-clusters is five and therefore, besides the “abnormal” co-cluster,

this chapter defined "very early", "early", "late", and "very late" co-clusters according to their average *FLD* values, which (rounded) correspond to 37th, 67th, 105th and 152nd days of the year. These irregular co-clusters discretize the whole *FLD* dataset and, hence, they are referred as the main *FLD* categories. Figure 4.7 shows the composition of the irregular co-clusters using the same format as that for the regular co-clusters (Figure 4.5). Figure 4.7 also shows that some regular co-clusters (e.g. location-cluster39/year-cluster1) belonging to one *FLD* category (e.g. "very early" *FLD*) are embedded in another category (e.g. "early" *FLD*).This is because BBAC_I re-orders locations and years into co-clusters using the average value of a complete location and year (i.e. full row or column of the original *FLD* matrix). For instance, the average value of one location can be higher than another when considering the whole time series but lower when considering a subset of years.

4.3.3 Spatial-temporal patterns in *FLD*

The spatial distribution of each *FLD* category is displayed in the small multiples shown in Figure 4.8. It shows that the phenoregions for "late" *FLD* located mostly in Russia and Eastern Europe have the largest extents while those for "very early" *FLD* located mostly in Southern Europe and northern Africa have the smallest extents. It also shows that there are at most four phenoregions (as this is the number of year-clusters) for each *FLD* category due to different spatial extents. For instance for "very late" *FLD* located mostly in Northern Europe, the first two phenoregions have the same spatial extent while for "very early" *FLD* the last three phenoregions have the same extent. For both "early" *FLD* located mostly in Western Europe and Turkey and "late" *FLD*, there are four different phenoregions and thus there are more spatial variations for this two categories. It is also observed that the "abnormal" *FLD* category only happens in a very small area of southern Iceland (for all 62 years), and the borders between Western and Eastern Europe, Scandinavian countries and northern United Kingdom (UK) in recent years.

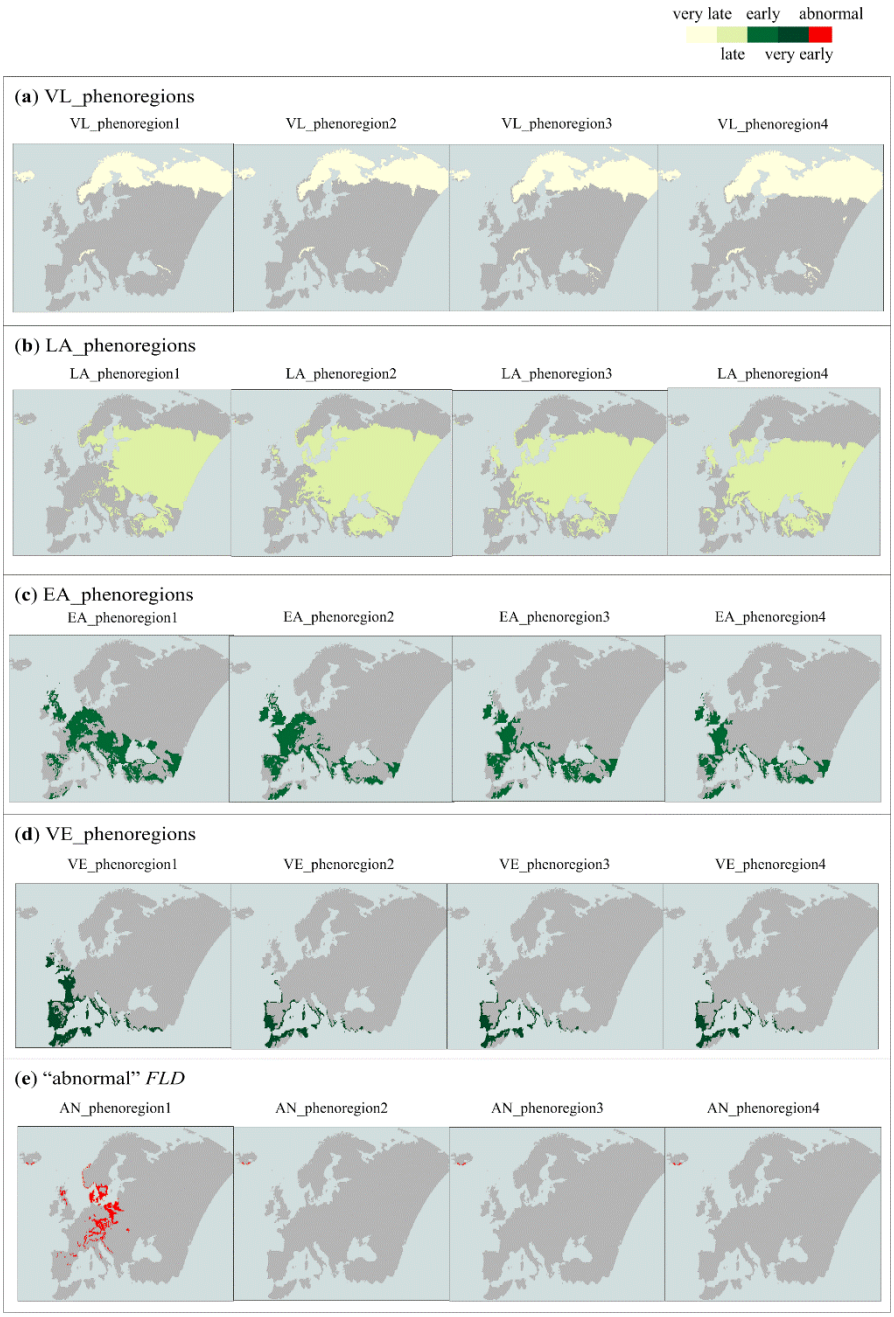


Figure 4.8: Phenoregions (self-similar clusters in terms of *FLD*) for each of the five *FLD* categories. For instance, the map for VL_phenoregion1 displays the phenoregion for "very late" *FLD* in accordance with year-cluster1. For categories with less than four, four maps were still used for consistency.

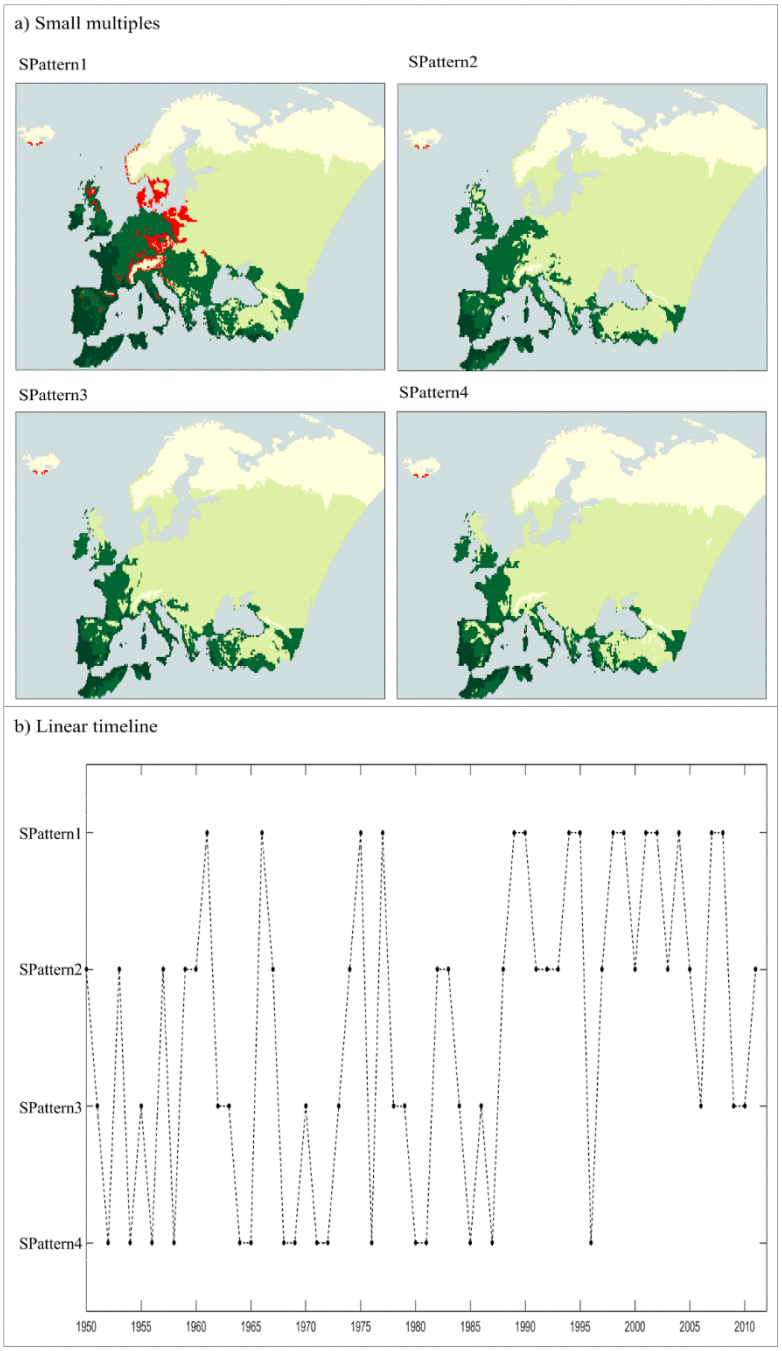


Figure 4.9: a) The small multiples to show four unique spatial patterns (SPattern1~SPattern4) over Europe; and b) the linear timeline to show the temporal dynamics of these spatial patterns from 1950 to 2011.

Table 4.1: Percentages of phenoregions per spatial pattern and the number of years that exhibit each spatial pattern

Spatial patterns	VL_phenoregions	LA_	EA_	VE_	Number of years
SPattern1	24.4%	46.4%	16.7%	9.1%	15
SPattern2	24.2%	57.9%	13.7%	4.2%	18
SPattern3	28.4%	54.9%	12.5%	4.2%	13
SPattern4	33.4%	51.7%	10.7%	4.2%	16

Grid cells with the same variation over the whole study area in Figure 4.7, that is, phenoregions falling into the same year-cluster in Figure 4.8, compose four unique spatial patterns over Europe. These spatial patterns, which were named SPattern1 to SPattern4, were showed in the small multiples in Figure 4.9a and their temporal dynamics from 1950 to 2011 was showed in the linear timeline in Figure 4.9b.The percentages of phenoregions for each *FLD* category and the number of years taken by these patterns were showed in Table 4.1. Among the four spatial patterns, SPattern1 has the highest percentages of phenoregions for "very early" and "early" *FLD* and lowest for "late" *FLD* while SPattern4 has the highest of phenoregions for "very late" *FLD* and lowest for "very early" and "early" *FLD*. Thus, variations from SPattern1 to SPattern2 or SPattern3 result mainly in 4.9% decrease in "very early" phenoregions located in the Iberian Peninsula, western France and southern UK, and also 4% decrease in "early" phenoregions in the Balkan Peninsula and northern UK. Thus such variations indicate increasingly late *FLD* and spring onsets and so do variations from SPattern2 or SPattern3 to SPattern4. Besides aforementioned decrease in "very early" phenoregions, variations from SPattern1 to SPattern4 also result in 6% decrease in "early" phenoregions located in Germany and Austria, and also 9% increase in "very late" phenoregions in Scandinavia. Therefore such variations indicate the trend for the very late *FLD* and spring onsets. The timeline in Figure 4.9b shows that there is a general trend towards earlier *FLD* values from early to recent years and from north to south of Europe. It shows that more than two thirds (eleven out of fifteen) of SPattern1 occurred after 1987 while others of SPattern1 did not occur until 1960. It also shows that few (one out of sixteen) SPattern4 occurred after 1987 in 1996. Thus the period of 1950 -1960 had very late spring onsets and the period of 1988-2011 had very early spring onsets except 1996. The first big variation of spring onsets during the study period occurred from 1965 (SPattern4) to 1966 (SPattern1) where large areas in Western and Southern

Europe, UK and Scandinavia experienced increasingly very early *FLD*. The biggest variation of spring onsets occurred in the period of 1975-1977 where these areas experienced increasingly very late and then very early *FLD*.

Years with the same variation over the whole study period in Figure 4.7 compose fourteen unique temporal patterns. In Figure 4.10 each timeline shows one temporal pattern with variations among *FLD* categories over the time period 1950-2011 and the map below shows the geographical extent of that temporal pattern. Among the temporal patterns except those related to "abnormal" *FLD*, there are four temporal patterns with one *FLD* category for all years, indicating very late (first pair), late (sixth pair), early (twelfth pair) and very early (fourteenth pair) spring onsets separately. The maps show that very late and late spring onsets are prevalent in Northern and Eastern Europe respectively and the geographical extent of the latter is the largest among all. Other eight temporal patterns are with variations between different *FLD* categories, indicating changes between very late and late spring onsets (third and fourth pairs), late and early spring onsets (seventh to eleventh pairs), early and very early spring onsets (thirteenth pair). The maps show that Western Europe and the Balkan Peninsula, which predominately have variations between late and early *FLD*, have the most complex spring onsets. In general the more south the regions and the more recent the years are, the more are the variations biased towards early spring onsets. It also shows that western Turkey has the most intensive variations of spring onsets because *FLD* varied 32 times between late and early categories over the study period.

The validity of these identified spatio-temporal phenological patterns will be discussed in the following session by comparing them with those in previous studies. Besides, the potential reasons that might be responsible for these patterns will also be discussed.

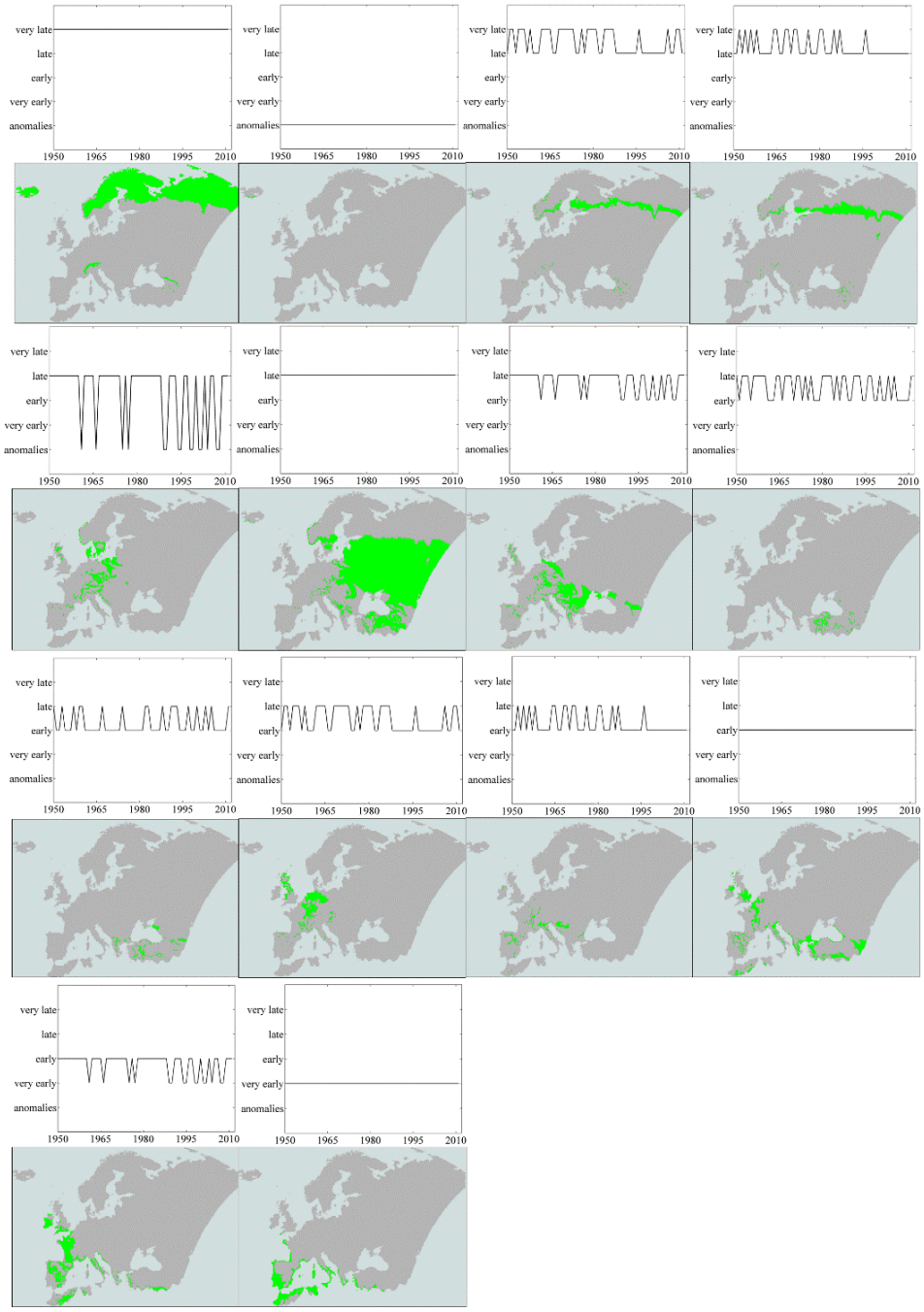


Figure 4.10: Each linear timeline to display each of fourteen unique temporal patterns with variations among *FLD* categories over the time period 1950-2011, and the map below to show the geographical extent of the temporal pattern.

4.4 Discussion

Previous studies on satellite-derived land surface phenology and ground phenological observations have provided information on spring dynamics in Europe using a wide variety of datasets. Through the analysis of AVHRR (Advanced Very High Resolution Radiometer) NDVI (Normalized Difference Vegetation Index) time series for the period 1982-2001, Stöckli and Vidale (2004) found that the period 1985-1987 had late spring onsets and that the years 1989, 1990, 1994 and 1995 had early ones. The results of this chapter confirm those findings. As showed in Figure 4.9, the period 1985 to1987 indeed had late sprint onsets (SPattern3 and Spattern4) and the four years had very early spring onsets (SPattern1). This could be attributed to climate changes with increasing temperature in this period, which are reflected by the fluctuation of atmospheric CO2 assimilation by vegetation (Keeling et al. 1996). Jeong et al. (2011) and Fu et al. (2014) extended the study period till 2008 and 2011 separately and found that spring onset (start of the growing season in their work) advanced significantly in the period 1982-1999 and that 'earlier springs' weakened in the period 2000-2008 and 2000-2011 respectively. Although with a different later changing point at 1988, this chapter also shows that the pattern in the recent period 1988-2011 (stable very early spring onsets) except 1996 is different with that in previous years. This weakened advance of spring onsets might be caused by asymmetric warming patterns of springs between the two periods. As stated by (Fu et al. 2014), the reduced spring warming in the latter period would result in the retarded advance of spring onsets. The different changing point of the two periods might be caused by the extension of the study period to include the period 1950-1981. Most years of this period have late spring onsets, which would lead to the weakened early spring onsets at an earlier year over the whole period. Using ground phenological records of leaf unfolding from 1951 to 1996, Menzel (2000) concluded that the increasing temperature mostly took place in the late 1980s and 1990s. This chapter also shows consistent spring onset dynamics that spatial patterns varied from SPattern4 to SPattern1 from 1987 to 1989 (Figure 4.9), indicating increasingly very early spring onsets. However, this chapter observed stable early spring onsets in the late 1990s. This discrepancy might be caused by two reasons. The first one is the extended study period in this chapter and spatial coverage. Second, as mentioned earlier, the FLD index calculated using the SI-x models indicates the spring onset of the three key species and other relevant plants (grasses, shrubs and fruit trees). Whereas the phenological records used in (Menzel 2000) include different species. It could thus result in the discrepancy in

the trend of spring onset. Chmielewski and Rötzer (2001) analyzed the ground observational values of the timing of leaf unfolding for the period 1969-1998 and reported that spring onsets were generally early in 1990s except 1996, due to the long and strong winter of 1995/1996. These results agree well with the findings in this chapter. Chmielewski and Rötzer (2001) also analysed regional patterns of the timing of leaf unfolding and found weak trends of early spring onsets in northern Scandinavia and of late ones in south-eastern Europe. This chapter also found these regional trends (first and seventh pairs in Figure 4.10). The observed mildly decreasing temperature in the last few years of their study period may explain the trends in these regions. Ahas et al. (2002) extended the study period of this latter study to 1951-1998 and reported that first third of the study period mostly has late spring onsets and the last third has early spring. Such trends are can also be observed in Figure 4.9 in this chapter. All of these results confirm the effectiveness of co-clustering and SI-x based analysis to capture the spatio-temporal phenological patterns.

In contrast to satellite-derived phenological metrics and ground observations as cited above, phenological models have the advantage of predicting phenological metrics of spring onsets over large regions and for long time periods. For instance, Schwartz et al. (2006) analyzed European *FLD* patterns from 1955 to 2002 using the original spring index models (SI) and temperature data from meteorological stations. The stable spring onsets in Eastern Europe reported in their study are consistent with this chapter (sixth pair in Figure 4.10). It might be caused by the weak spring warming in this region (Jones and Moberg 2003). However, this chapter found variations between late and early spring onsets with the bias towards the latter in recent years in parts of Germany. This contrasts with their conclusions, which indicate that central Europe (mainly Germany) exhibits the strongest trend of earlier spring onset. This inconsistency might be caused by two reasons. The first one is different study periods. The extended study period in this chapter might lead to different trends of spring onsets. The second reason is different sources of datasets. Schwartz et al. (2006) used temperature data from individual meteorological stations and the resulting trend exists in these isolated and un-evenly distributed stations. Whereas the temperature data from a 0.25 degree interpolated grid cells is used in this chapter and the resulting trend spreads over the spatial continuous region. Thus, it could result in different trends. Nevertheless, the temporal dynamics of four spatial phenological patterns explored in current chapter is confirmed by the similar temporal variation of departure from long-term average showed in their work by a simple curve.

However, more detailed spring phenological patterns over Europe are described in this chapter by the four spatial patterns showed in maps (Figure 4.9).

Even though SI-x do not consider phenological strategies such as chilling requirements (Polgar and Primack 2011), they retain the utility and accuracy of SI which take above strategies into account (Schwartz et al. 2013). This fact enables the extension of the analysis into areas such as northern Africa in this work or those analyzed by (Zhang et al. 2007) where chilling requirements are not met (Schwartz et al. 2013). Thus, together with the use of the E-OBS temperature datasets in this chapter, SI-x allowed to delineate spatially contiguous phenoregions in Europe, northern Africa and Turkey. These phenoregions are more informative than the analysis of temporal dynamics based on un-evenly distributed locations (meteorological stations or ground observations). Moreover, the approach in this chapter allows visualizing the boundaries of the main phenoregions and studying their changes in time. To the best of my knowledge, this chapter reports the first spatially-continuous phenoregions in Europe (northern Africa, Turkey) while several phenological regionalizations already exist for CONUS (Gu et al. 2010, Kumar et al. 2011, Mills et al. 2011, Zhang et al. 2012). The results in this chapter also confirm that SI-x work well for large-scale analysis. However, one potential caveat of the SI-x models is that they do not consider other environmental variables than temperature and, indirectly, day length. Therefore, the SI-x models might not fully capture the phenological dynamics of areas or species that are driven by, for instance, precipitation.

Finally, in contrast to previous phenological studies, that analyzed spring phenology from a separate view, this is the first study that applies co-clustering to the exploration of spring phenological patterns. Co-clustering allows the simultaneous analysis of the spatial and temporal dimensions of the data. This resulted in the simultaneous identification of not only the main spatial patterns in the study area and their temporal dynamics, but also the identification of the main temporal patterns over the study period and their spatial distributions.

4.5 Conclusions

This chapter has presented a novel analytical approach that allows to simultaneously study spatio-temporal patterns over large areas and long time periods. In more details, the analysis first used the extended spring index models to calculate time series of *FLD* from the E-OBS daily maximum and minimum temperature datasets over Europe, northern Africa and Turkey from 1950 to 2011. Then the *FLD* dataset was co-clustered by applying the Bregman block average

co-clustering algorithm with I-divergence (BBAC_I) and *k*-means. This co-clustering based approach identified irregular co-clusters that contain similar *FLD* values along both locations and years. Finally these irregular co-clusters were visually explored to reveal the main spatio-temporal patterns in spring phenology.

The developed analytical approach allowed the delineation of spatially-continuous phenoregions in Europe for the first time. Besides, the analysis in this chapter identified four unique spatial patterns and showed that the early years of the study period (1950-1960) have very late spring onsets, especially in northern Europe and Russia. The period 1961-1987 has many variations in the timing of spring onset. The largest variation took place in the years 1975 to 1977 where the spring onset occurred from very early to very late and back. This chapter also observed that recent years (from 1988 onwards) tend to have stable early springs, especially in the Iberian Peninsula and northern Africa. One exception is the year 1996 that had a very late spring onset. The results also identified twelve unique temporal phenological patterns as well as their spatial distributions. Four of these temporal patterns indicate stable spring onsets and eight patterns indicate variable spring onsets. The analysis of their spatial distribution showed that very late and late spring onsets prevail in Northern and Eastern Europe while variations between late and early spring onsets prevail in Western Europe and the Balkan Peninsula, with the most intensive variations located in western Turkey.

For the future work, the identification of the driving forces behind the patterns found in this chapter is anticipated. This could be done by linking the patterns (e.g. phenoregions and temporal dynamics) with environmental variables. More generally, further work could investigate the role of decadal and inter-annual climatic variability, e.g. from the influence of North Atlantic Oscillation (NAO) in the variations of spatial and/or temporal spring onset patterns.

Chapter 5 Tri-clustering geo-referenced time series for analyzing patterns of intra-annual variability in temperature*

*__This chapter is based on the manuscript:__ Wu, X., R. Zurita-Milla, E. Izquierdo-Verdiguier & M.-J. Kraak (2016). Tri-clustering geo-referenced time series for analyzing patterns of intra-annual variability in temperature. *Annals of the American Association of Geographers.* In revision.

Abstract:

Clustering analysis is often used to explore patterns in geo-referenced time series (GTS). However, most clustering studies only analyze GTS from one or two dimension(s) and are not capable of analyzing the 3D GTS defined by one spatial, one temporal and any third (e.g. attribute) dimensions. Here this chapter develops a novel clustering algorithm called Bregman cuboid average tri-clustering algorithm with I-divergence (BCAT_I), which enables the analysis of 3D GTS. BCAT_I simultaneously groups the data along its dimensions to form regular tri-clusters. These tri-clusters are subsequently refined using *k*-means to fully capture spatio-temporal patterns in the data. By applying BCAT_I to time series of daily average temperature in the Netherlands (28 weather stations from 1992 to 2011), this chapter identified the refined tri-clusters with similar temperature values along the spatial (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. The results show that in the last two thirds of the study period, there is an intense variability of spring and winter temperatures at the northeast & center of the Netherlands. For the same period, an intense variability of spring temperatures is also visible at the southeast of the country. The 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 spatio-temporal data.

Key words: data mining, geovisualization, geo-referenced time series, intra-annual variability, tri-clustering

5.1 Introduction

Geo-referenced time series (GTS) describe the time-evolving behavior of one or more attributes that are typically recorded at fixed locations and uniform temporal 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 a type of spatio-temporal data and, as such, they "live" in the n-dimensional space formed by their spatial, temporal and (multi)attribute dimensions (Guo et al. 2006). This chapter focuses on the analysis of 3D GTS with one spatial, one temporal and any third (e.g. attribute) dimensions. Such data are naturally modelled and viewed as a data cuboid (Harinarayan et al. 1996, Han et al. 2011), in which each cell stores the value of each attribute observed at one location in one timestamp. In particular, this chapter presents a novel tri-clustering 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 datasets by assigning similar data elements to the same group (Andrienko et al. 2009). By this means, clustering analysis 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 traditional clustering methods to analyze patterns in the datasets (e.g. (Hagenauer and Helbich 2013, Helbich et al. 2013, Grubesic et al. 2014)). In these studies, the authors group data elements along spatial dimension (i.e. locations) with similar values of the attribute on the full data space. Recently, Wu et al. (2015) and Wu et al. (2016) used co-clustering methods to perform 2D clustering for pattern analysis in GTS, that is, to simultaneously group data elements along the spatial and temporal dimensions of the data. However, neither traditional clustering nor co-clustering methods are capable of analyzing 3D GTS. Hence this chapter focuses on the use of a tri-clustering method for such analysis.

Tri-clustering methods have already been applied in other fields. For instance, Zhao and Zaki (2005) developed a tri-clustering algorithm called TRICLUSTER that identifies gene-sample-time clusters in 3D microarray datasets; Ji et al. (2006) proposed the CubeMiner algorithm to mine frequent co-occurrences of gene-sample-time in 3D microarrays too and Sim et al. (2010) presented a tri-clustering algorithm called MIC to mine correlated 3D subspace clusters from financial datasets. However, CubeMiner is only applicable to binary datasets and TRICLUSTER and MIC only aim at searching for significant clusters instead of exhaustively identifying all clusters in the 3D dataset. A cluster is seen as

significant if it is intrinsically outstanding or more interesting than other clusters for a specific tasks and clustering methods (Sim et al. 2013). As such, none of these tri-clustering algorithms are able to fully analyze 3D GTS by providing a complete partition of the dataset. The issue necessitates of a new tri-clustering algorithm that is capable of identifying all clusters in 3D GTS. To this end, this chapter expands the previous works on co-clustering analysis (Wu et al. 2015, Wu et al. 2016) and presents a tri-clustering algorithm specifically designed to analyze GTS that fit into data cuboids.

The main objective of this chapter is therefore to develop a tri-clustering algorithm that allows the simultaneous analysis of GTS along its three dimensions. The possibilities of this algorithm are demonstrated by analyzing a GTS of daily temperatures. Such data naturally fits into a cuboid where each cell contains a temperature value indexed by its location and timestamps (year and day) of measurement. This application of tri-clustering paves the way towards the analysis of spatio-temporal patterns of intra-annual variability in temperature records, thereby supporting the study of the ecological impacts of climate change (Walther et al. 2002).

5.2 Methods

In this section, the tri-clustering algorithm is firstly presented and then the need for refining the results by regrouping the tri-clusters using *k*-means is explained.

5.2.1 Bregman cuboid average tri-clustering algorithm with I-divergence (BCAT_I)

This section describes the development of the Bregman cuboid average tri-clustering algorithm with I-divergence as similarity metric (BCAT_I). 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 co-clustering algorithm with I-divergence (BBAC_I used by (Wu et al. 2015, Wu et al. 2016)). BCAT_I enables the simultaneous clustering of the elements along all dimensions of a data cuboid filled with positive real-values data. Such a data cuboid can be regarded as co-occurrences among three random variables: the stations (*S*), the years (*Y*) and the days of the year (*D*). In this set-up the rows of the data cuboid refer to the stations that represent the fixed locations, the columns

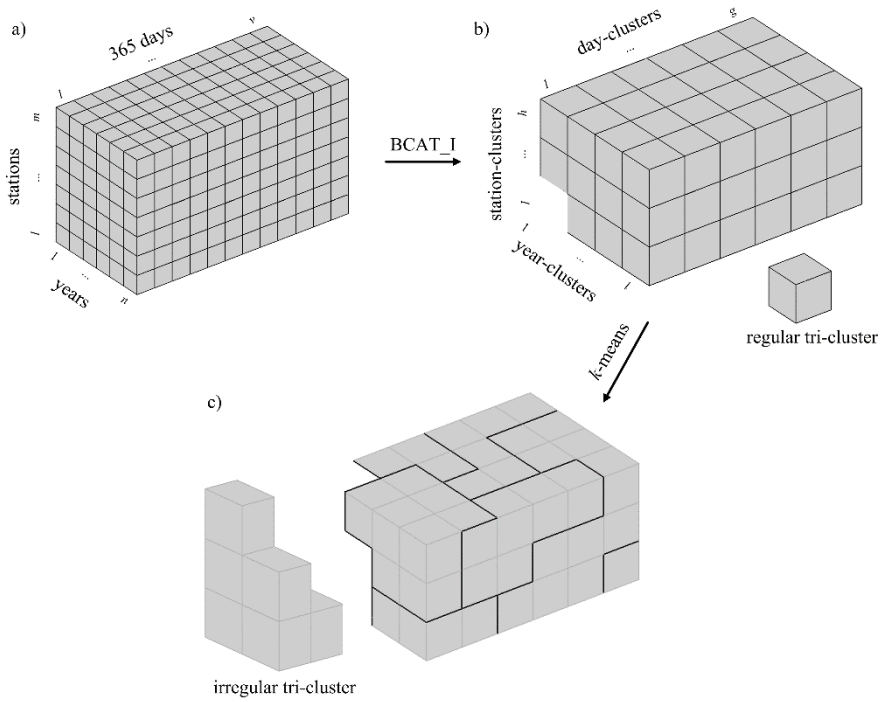


Figure 5.1: a) The data cuboid with size $m \times n \times v$. The rows refer to stations, the columns to years and depths to 365 days; b) $h \times l \times g$ regular tri-clusters (sub-cuboid) after applying BCAT_I to the data cuboid. The rows refer to station-clusters, the columns to year-clusters and depths to day-clusters; c) k irregular tri-clusters refined from regular ones with k-means. The axes arrangement is the same as that for regular tri-clusters.

to the years and the depths to the 365 days that belong to each year (for convenience, the 29th 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 5.1a).

By concurrently grouping stations to station-cluster, years to year-clusters and days to day-clusters, the tri-clustering algorithm seeks for tri-clusters that contain similar temperature values along the three dimensions of the input data cuboid. These tri-clusters are defined by the intersection of station-, year-, and day-clusters (Figure 5.1b). Like the BBAC_I, the composition of the tri-clusters is optimized by using the I-divergence metric, whose superiority over other metrics (e.g. Euclidean distance) has been empirically proved (Banerjee et al. 2007). As such, the tri-clustering problem can be regarded as an optimization one where the optimal results minimize the loss of mutual information between the original data cuboid and the tri-clustered one.

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 5.2: The pseudocode of Bregman cuboid average tri-clustering algorithm with I-divergence (BCAT_I).

Figure 5.2 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 $\mathbf{O}' \in \mathbb{R}^{m \times vn}$, of which the rows are the m stations and the columns are $v \times n$ days/years. This allows the definition of $\widehat{\mathbf{O}'}$, which is calculated as the averages of elements of $\mathbf{O}'$ that belong to the same cluster according to the current mapping. Then an approximate matrix of $\mathbf{O}'$, named $\mathbf{A} \in \mathbb{R}^{m \times vn}$, is created by

expanding $\widehat{\mathbf{O}'}$ based on the mapping and the values of the same cluster to be preserved. After that, the loss of the mutual information between $\mathbf{O}'$ and **A** is calculated. By minimizing the information loss, the optimal mapping from stations to station-clusters is produced and **R** is updated. The optimization proceeds with the year and day clustering to update **C** and **T**. Once the information loss is minimal, the update of the **R**, **C** and **T** matrices stops and this yields the optimal tri-clustering results (Appendix contains a detailed explanation of the tri-clustering algorithm). Finally, the original data cuboid **O** is re-ordered following the optimized binary matrices **R**, **C** and **T**, to group together the elements that belong to the same tri-cluster. For the particular application in this chapter, 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 tri-clusters have increasing temperature values from the bottom-left-front corner to top-right-back corner of the re-ordered cuboid. To simplify the analysis and visualization of the tri-clusters their values are set to the average of the elements that belong to each tri-cluster.

5.2.2 Refinement of BCAT_I result

The BCAT_I partitions the data cuboid into $h \times l \times g$ regular tri-clusters. The need to predefine the numbers of clusters in the tri-clustering algorithm leads to a potential issue that different (tri-)clusters might still have similar values (Wu et al. 2015, Wu et al. 2016). To mitigate this problem and better capture the patterns in GTS, this chapter suggests using *k*-means to re-group the regular tri-clusters into *k* irregular tri-clusters (Figure 5.1c). It also suggests using the mean and variance of data elements belonging to each of the regular tri-clusters as inputs for *k*-means. The number of irregular tri-clusters (*k* in *k*-means) is optimized by using the Silhouette method (Rousseeuw 1987) because it produces results that correlate well with human evaluations of clustering results (Lewis et al. 2012). Finally, like with the regular tri-clusters, the values of the irregular tri-clusters are set to the value of the *k*-means centroids.

5.3 Using BCAT_I to explore spatio-temporal 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 et al. 2008, Williams and Middleton 2008). As stated by Doi et al. (2008), patterns of annual averages of weather variables do not properly capture those of intra-annual variability whereas the latter have stronger impact on ecosystems than the former. This motivates the experiment in this chapter, 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 tri-clustering analysis allows the exploration of the spatio-temporal patterns of intra-annual temperature variability in the Netherlands.

5.3.1 Data

This chapter uses Dutch daily average temperature data collected at 28 meteorological stations over 20 years (from 1st of January 1992 to 31st of December 2011). This temperature data and the geographical coordinates of each station were obtained from the Royal Netherlands Meteorological Institute (KNMI, http://www.knmi.nl/). Using the stations coordinates and the boundary of the Netherlands, a Thiessen polygon map that defines the area influenced by each station is generated (Figure 5.3). In this map, each polygon is labelled with

Figure 5.3: Study area: Thiessen polygon map of the Dutch meteorological stations.

both the station ID given by the KNMI (e.g. 290) and the name of the station (e.g. Twente).

The location of the Netherlands in Europe (bottom right of Figure 5.3) 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 with 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 is smaller than that in the (south)east (Lenderink et al. 2011).

5.3.2 Experiment design

As discussed in section 5.2, the Dutch temperature data was first organized in a data cuboid of 28 (stations) by 20 (years) by 365 (days). Then BCAT_I was used to tri-cluster this cuboid. The number of station- and year-clusters were empirically set to 4 following the results of (Wu et al. 2015) who used the annual averages of the same temperature dataset for the co-clustering analysis. The number of day-clusters was fixed to 8 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 4 (station-clusters) by 4 (year-clusters) by 8 (day-clusters) regular tri-clusters, which were subsequently refined into *k* irregular tri-clusters. After that, results were displayed using several (geo)visualization techniques. Both 3D and 2D heatmaps were used to visualize both the regular and irregular tri-clusters. One set of small multiples and two sets of timelines were used to show the composition of the regular tri-clusters (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 tri-clusters, for each year-cluster the values of day-clusters along station-clusters were examined. 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 Netherlands exhibit 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 regions, each 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 the day-clusters with the same spatial pattern were combined and chronologically visualized in one timeline. The values of the irregular tri-clusters to which these day-clusters belong were used to define the colors used in the timelines. These timelines were arranged in aligned with geographic maps in the second set of small multiples that indicate corresponding spatial patterns.

5.4 Results and discussion

5.4.1 Regular and irregular tri-clusters

The application of BCAT_I to the Dutch daily temperature dataset yielded 128 (4 × 4 × 8) regular tri-clusters. The 3D heatmap (center) and four 2D heatmaps (side subplots) in Figure 5.4 display these tri-clusters in the re-ordered data cuboid. In the 3D heatmap, rows indicate station-clusters, columns indicate year-clusters, depths indicate day-clusters and their intersections (sub-cuboids) indicate regular tri-clusters. As an example, the sub-cuboid marked by the thick line in Figure 5.4 shows the regular tri-cluster (4, 3, 1). This tri-cluster is formed

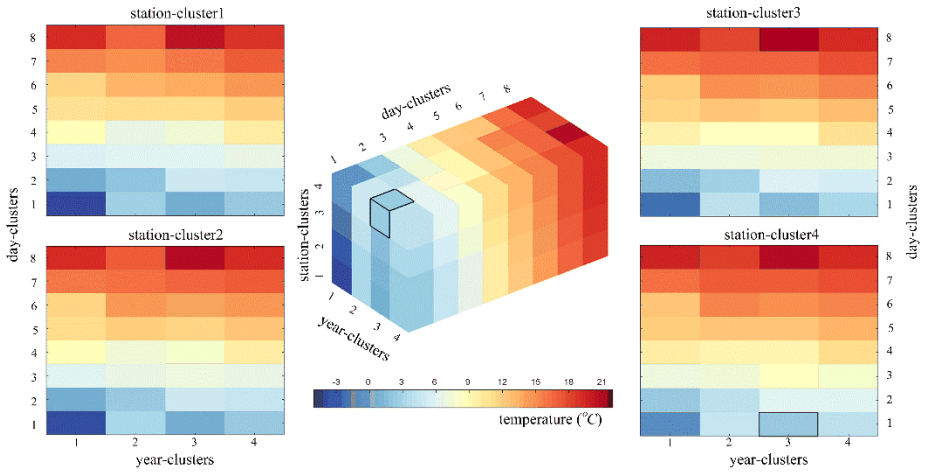


Figure 5.4: The resulting 128 (4 × 4 × 8) regular tri-clusters from BCAT_I in 3D heatmap (middle) and 2D heatmaps (side subplots) illustrating each of station-clusters. In the 3D heatmap, the rows indicate station-clusters, columns indicate year-clusters and depths indicate day-clusters. In each 2D heatmap, x-axis indicates year-clusters and y-axis indicates day-clusters. The regular tri-cluster (4, 3, 1), intersected by station-cluster4, year-cluster3 and day-cluster1, is highlighted by the think lines in both 3D and the 2D heatmaps.

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

Both 3D and 2D heatmaps clearly show that the newly developed tri-clustering algorithm exhaustively identifies all tri-clusters in the dataset, unlike only significant ones in other tri-clustering algorithms (Zhao and Zaki 2005, Sim et al. 2010). This feature is necessary to allow the full analysis of 3D GTS. Also, the example of regular tri-cluster shows that it contains one more dimension than the co-cluster (Wu et al. 2015) in terms of days. Compared with BBAC_I that analyzes GTS from two dimensions, BCAT_I considers three dimension of GTS and allows the analysis of the original data also along day dimension without the loss in details. Consequently, the resulting tri-clusters contain more information than co-clusters, which is essential for the exploration of intra-annual variability. Both the 3D and 2D heatmaps in Figure 5.4 show that many regular tri-clusters have similar temperature values. Besides, the heatmaps show that the tri-cluster with the highest temperature, which is supposed to be (4, 4, 8) in this re-ordered

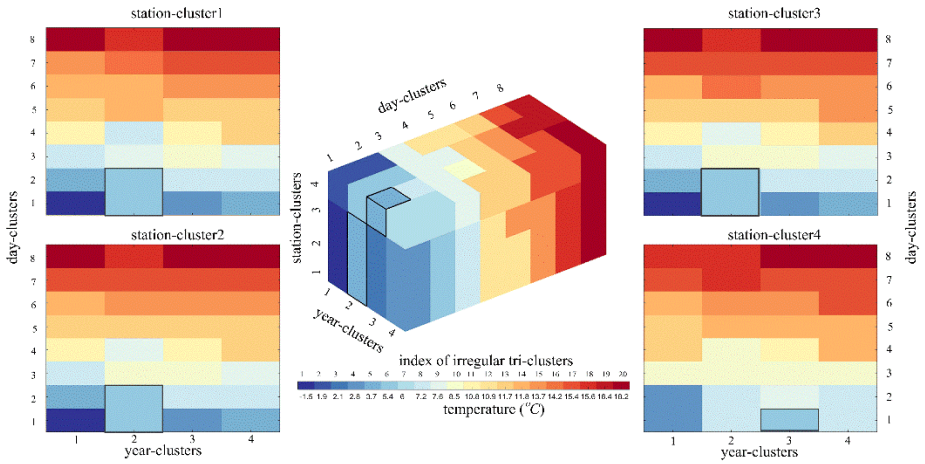


Figure 5: The resulting 20 irregular tri-clusters from *k*-means in 3D heatmap (middle) and 2D heatmaps (side subplots). The axes arrangement in all heatmaps is the same as that in Figure 5.4. The same example of irregular tri-cluster (number 5) is highlighted by the thick lines in both the 3D and the 2D heatmaps.

data cuboid, is (4, 3, 8) instead. It is supposed that is because the re-order is based on average values of whole station-, year- and day-clusters. This demands the refinement of these tri-clusters using *k*-means.

After testing *k* values from four to 30 in steps of one, the Silhouette method identified 20 as the optimal number of irregular tri-clusters. These irregular tri-clusters are indexed and shown in Figure 5.5 using a 3D heatmap (center) and four 2D heatmaps (side subplots). The legend contains discrete values for each of 20 irregular tri-clusters, unlike the one in Figure 5.4 with continuous values to indicate representative temperatures assigned to regular tri-clusters. These discrete values of irregular tri-clusters suggest the usefulness of the refinement by *k*-means.

The thick lines in the heatmaps of Figure 5.5 show one same example of irregular tri-cluster (in this case number 5). By using the same axes arrangement in Figures 5.4 and 5.5, the composition of each irregular tri-cluster from regular ones can be observed. For example, the irregular tri-cluster number 5 is composed from the following regular tri-clusters: (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 sub-cuboids with similar temperatures. This grouping completely identifies similar temperature values of the original dataset along spatial, temporal and day dimensions, which thus enables the full exploration of the complex patterns in the data.

Figure 5.6 shows the spatial distribution of station-clusters using small multiples (panel a) and the temporal distributions of year- and day-clusters using two sets of timelines (panels b and c, respectively). The small multiples show that the temperature-wise 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 with that in the east of the country. This division is the same as that in the co-clustering analysis by BBAC_I (Wu et al. 2015) that used annual averages of the same dataset while different from that in the clustering analysis by SOMs (Wu et al. 2013) using the same dataset. It is supposed that might because the divided regions are only affected by the involvement of the temporal information in the data (i.e. year and day in tri-clustering analysis and year in co-clustering analysis). The linear timeline of year-clusters shows that 80 percent of the study period (16 out of 20 years), especially the period of 1999-2011, 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) has very low temperature values, and the remaining 10 percent (years 1993 and 1998) has low temperature values. This

distribution of year-clusters over the study period is also supported by that in (Wu et al. 2013, Wu et al. 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. It indicates that temperature is much changeable in the spring (Jaagus and Ahas 2000).

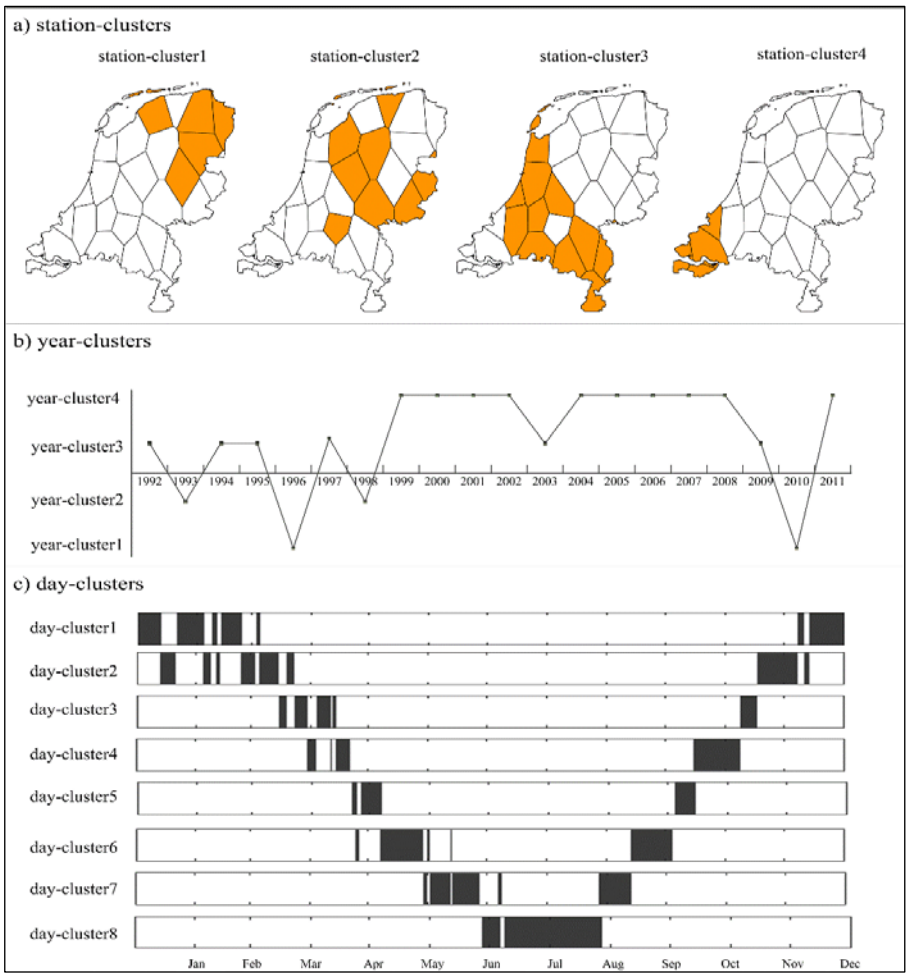


Figure 5.6: The composition of regular tri-clusters. a) the small multiples to display the spatial distribution of station-clusters 1 to 4; b) a linear timeline to show temporal distribution of year-clusters 1 to 4 and c) the timelines to show the temporal distribution of day-clusters 1 to 8.

The combination of Figures 5.5 and 5.6 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 the southwest region. The warmest temperatures occurred in the summer period of all the 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 co-clustering analysis (Wu et al. 2015), except that more complete information about similar temperature values especially daily-wise is discovered in this chapter.

In respect of visualization, it is worth pointing out that it would be more enlightening to have the interactive 3D heatmaps and multiple linked views for previous figures. In that case, the 3D heatmaps in Figures 5.4 and 5.5 can be rotated and edited as transparent to allow the highlighting of any selected ir/regular tri-cluster. At the same time, the corresponding regular/irregular tri-cluster that compose/comprise it, in 3D and 2D heatmaps and also corresponding distributions in Figure 5.6 can be also highlighted. An interactive interface that integrates aforementioned features is planned as future work.

5.4.2 Spatio-temporal patterns of intra-annual variability

Spatio-temporal patterns of intra-annual variability in Dutch daily average temperature from 1992 to 2011 were analyzed from the 20 irregular tri-clusters and displayed in the small multiples and timelines in Figure 5.7.

The small multiples in Figure 5.7a 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 showed in the heatmaps for irregular tri-clusters (Figure 5.5), the value of day-clusters 1 to 5 and 8 is the same for station-clusters 1 to 3 and different for station-cluster4. It 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 is that the Netherlands is divided into two regions: northeast & 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 to 4); (2) the northeast &

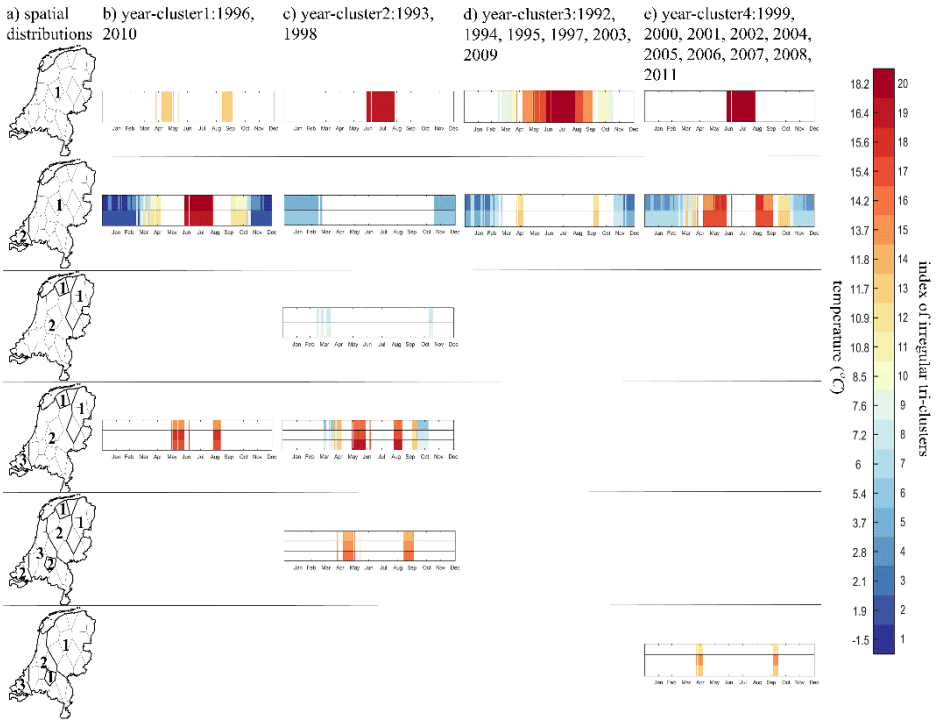


Figure 5.7: Spatio-temporal patterns of intra-annual variability in Dutch daily temperature from 1992 to 2011. a) the small multiples to show the unique spatial patterns of intra-annual variability. b-e) the timelines to show temporal patterns of temperature variability within each of four year-clusters. Each timeline is aligned with corresponding spatial pattern.

center of the country as region1 (station-clusters 1 to 3) and the southwest as region2 (station-cluster4); (3) the northeast as regoin1 (station-cluster1) and the center & southwest as region2 (station-clusters 2 to 4); (4) the northeast as region1 (station-cluster1), the center as region2 (station-clusters 2 and 3) and the southwest as region3 (station-cluster4); (5) the northeast as region1 (station-cluster1), the center-northeast & southwest as region2 (station-clusters 2 and 4) and the center-southwest as region3 (station-cluster3); (6) the northeast & 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 5.7a) 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. It might because this two areas are more directly

influenced by continental and maritime climates respectively. Besides, it is observed that in the last two spatial patterns, the station 356 (Herwijnen, see Figure 5.3) is isolated from the region it belongs to by another region with different variabilities. The possible reasons for this requires further analysis with local environmental variables.

The timelines in Figures 5.7b to 5.7e 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 tri-clusters by combining day-clusters with the same spatial patterns and visualizing them chronologically. The timelines are aligned with the corresponding spatial patterns (Figure 5.7a). As an example, year-cluster1 is discussed 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: northeast & 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 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 5.7a and 5.7b shows the spatio-temporal patterns of variability in temperature within year-cluster1 (1996, 2010). In this two cold years, temperatures in most days exhibit different variabilities at the northeast & center of the Netherlands from that at 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 & center of the country experienced an intense variability in both winter and spring temperatures while the southwest area only underwent such a variability in spring temperatures. Such changeable temperature in winter is worthy of special notice because such phenomenon in spring and its effect have been well known (Chmielewski and Rötzer 2001).

Figure 5.7c for year-cluster2 (1993, 1998), together with Figure 5.7a, shows that this two years experienced the most complex temperature variability. Except winter and first half summer, days in this two years are scattered with different spatial patterns, which indicate that temperatures in these days experienced intense variability across the whole Netherlands. Such result is supported by the findings of (Wu et al. 2015) who found that the days in 1993 had the most changeable temperatures. Besides, 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). It 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 5.7a and 5.7d, temperatures in most days from spring to autumn in year-cluster3 underwent the same variability throughout the Netherlands. Temperatures in other days the northeast & center and the southwest of the country experienced different but both mild variabilities. It could be because in years belonging to year-cluster3, only one of maritime and continental climates completely influences the whole country in most days from spring to autumn whereas the latter becomes dominant in other days.

The combination of Figures 5.7a and 5.7e shows the spatio-temporal patterns of temperature variability in temperature within year-cluster4, in which most elements are recent and hot years. In these years, temperatures at the northeast & center of the Netherlands experienced more variabilities than that at the southwest in most days except summer.

Thus, the Netherlands has complex spatio-temporal patterns of intra-annual variability in temperature, despite of its relatively small area. For most days in the whole study period, the variability in temperature defines two regions in the country: the northeast & center and the southwest. Both in cold years (i.e. 1996, 2010) and hot years (i.e. years belong to year-cluster4), the northeast & center of the country experienced an intense variability in the spring and winter temperatures while the southwest only experienced such a variability in spring temperatures. Whereas 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 while in other seasons, especially the spring, both climates become influential.

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

5.5 Conclusions

This chapter presents a newly developed tri-clustering 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) dimensions. Unlike one-way clustering or co-clustering, BCAT_I

is capable of simultaneously clustering the data along all three dimensions of the data cuboid. The resulting tri-clusters can be subsequently refined using k-means to identify an optimal number of irregular tri-clusters that enables the full exploration of spatio-temporal patterns hidden in the 3D GTS.

In this chapter, BCAT_I was used to analyze time series of Dutch daily average temperature 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 tri-clustering analysis identified groups of stations and years that have similar within-year variability in the average daily temperatures. Displayed using various geovisualization techniques, the results show that the Netherlands has six unique spatial patterns of intra-annual temperature variability associated to 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 & center of the Netherlands. Such a variability is also found in the spring temperatures at the southeast of the country. Besides this, it is found that temperatures in most days of 1993 and 1998 experienced an intense variability across the whole country. However, it is also found that summer temperatures are homogeneous throughout the country for most of the study period.

These results indicate the possibilities of this newly developed tri-clustering algorithm to effectively analyze the complex patterns in daily temperature series. Nevertheless, it is important to notice that the proposed BCAT_I algorithm and subsequent refinement of the tri-clusters are generic. As such, they can be applied to any 3D GTS. For instance, they could be used to generate environmental zones by applying it 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 tri-clustering 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 spatio-temporal data.

Chapter 6 Synthesis

6.1 Introduction

Exploring patterns from spatio-temporal data helps to provide actionable information and knowledge to various stakeholders. This improves decision-making in spatial and temporal applications. Clustering methods can be used to extract patterns from various types of (spatio-temporal) data. However, they are insufficient to represent these patterns. In this context, geovisualization techniques can be used to visualize the extracted patterns. Despite previous studies using both clustering methods and geovisualization techniques, several challenges still exist in the exploration of geo-referenced time series (GTS), one important type of spatio-temporal data. These can be summarized as follows: (1) the exploration of differences in spatial and/or temporal patterns caused by the so-called Modifiable temporal unit problem (MTUP); (2) the exploration of concurrent spatial and temporal patterns; (3) the full exploration of spatio-temporal patterns from complex datasets; and (4) the exploration of patterns from 3D GTS. Aiming to answer these challenges and the research questions stated in section 1.6, the objective of this research is to combine one-way clustering, co-clustering, tri-clustering methods, and various geovisualization techniques, for the full exploration of patterns from GTS.

The four research questions were answered in the core chapters (2, 3, 4 and 5) of this thesis. This chapter starts with a reflection about the connections among the core chapters by summarizing and discussing their inter-relationships. Then it answers each of the research question and, after that, provides general conclusions. According to the above mentioned four challenges, the main contributions of this research are discussed and, finally, recommendations for future work are outlined.

6.2 Reflections: connecting the dots

(1) Applying a one-way clustering method to independently explore spatial and temporal patterns from GTS at different temporal resolutions

Given the complexity of space and time, a comprehensive analysis is needed to consider the space- and time-varying behaviour present in GTS. To this end, Andrienko et al. (2010) applied separate but complementary clustering analysis to explore patterns in the data along the spatial and temporal dimensions. Inspired by their work and concerned with MTUP, Chapter 2 developed an analytical approach that explores one specific pattern, synchronization, from a spatial and a temporal perspective and at different temporal resolutions. The approach is based

on self-organizing maps (SOMs) and applied various geovisualization techniques for representing the results.

Using Dutch daily average temperatures, Chapter 2 explored spatial and temporal synchronization in temperature at daily, weekly and monthly resolutions. Spatial synchronization means clusters of stations with similar temperature variations along all years. Temporal synchronization means clusters of years (a calendar year is used as basic unit in the analysis) with similar temperature variations along all stations. It is found that even within the same cluster in the results, for instance, station-cluster at daily resolution (Figure 2.4a-II), the variations of temperature are not the same at different temporal periods (Figure 2.4a-V). This motivates the use of a co-clustering method that enables the identification of similar values along both spatial and temporal dimensions of GTS (Chapter 3).

The same temperature dataset is also used in Chapter 3 for a co-clustering analysis and in Chapter 5 for a tri-clustering analysis. Moreover, the number of station-clusters identified in Chapter 2 is also adopted in the analysis shown in Chapter 3. As such, patterns explored by different clustering analyses are made comparable for cross-verification of the results among different chapters.

(2) Applying a co-clustering method to explore concurrent spatial and temporal patterns from GTS

The co-clustering method (BBAC_I) used in Chapter 3 enables the concurrent exploration of spatial and temporal patterns from GTS. BBAC_I grouped locations and timestamps at the same time, which resulted in co-clusters that contain similar values along both dimensions. Geovisualization techniques were then used to display these co-clusters and explore the corresponding patterns.

Using the Dutch daily temperature dataset, Chapter 3 identified regions as well as subsets of years/months/days that contain similar values. The results showed the concurrent spatial and temporal patterns in the dataset. These patterns were visualized by heatmaps, small multiples and timelines.

The results of this analysis showed that there are two aspects that can be further improved. The first one is caused by the existence of similar values in different co-clusters, especially neighboring ones. To solve the issue, the co-clustering resulting should be refined (Chapter 4). The second aspect concerns the analysis of 3D GTS. BBAC_I only allows the analysis of GTS with one attribute (i.e. of GTS that fit into a data table). If GTS with more than one attributes or with nested spatial or temporal hierarchies are to be explored, an extension of BBAC_I is needed (Chapter 5).

(3) Applying a co-clustering method and *k*-means to fully explore spatio-temporal patterns from GTS

The analytical approach presented in Chapter 4 combines BBAC_I and *k*-means to enable the full exploration of concurrent spatio-temporal patterns from GTS. BBAC_I was first applied to identify regular co-clusters, which were subsequently grouped by *k*-means into irregular co-clusters. By analyzing these irregular co-clusters, the spatio-temporal patterns were fully explored.

Together with a temperature-driven phenological model and a European gridded temperature dataset, the approach was applied to analyze time series of a phenological index, first leaf dates (*FLD*). The analysis found four main spatial phenological patterns of *FLD* over the study area and extracted their temporal dynamics over the study period. The analysis also identified twelve main temporal *FLD* patterns and their spatial distributions.

Chapter 4 improves the co-clustering analysis presented in Chapter 3 by refining co-clusters with similar values. As such, the patterns in GTS can be fully explored: not only spatial patterns and their dynamics along the temporal dimension, but also temporal patterns and their dynamics along the spatial dimension. Besides, Chapter 3 reports results of an initial experiment of the co-clustering analysis and, thus, is based on a dataset covering a relatively small territory and a short temporal period. After the applicability of the co-clustering method to analyze GTS is proved, Chapter 4 presents a more detailed study using a dataset that covers a large territory for a long time period and with some outliers. As such, Chapter 4 proves that the proposed co-clustering analysis is applicable to large and complex datasets.

(4) Developing and applying a tri-clustering method to fully explore patterns from 3D GTS

The tri-clustering method (BCAT_I) developed in Chapter 5 allows the analysis of GTS that fit into a data cuboid. By simultaneously grouping GTS along its three dimensions, BCAT_I identified tri-clusters formed at the intersection of each location-, timestamp- and attribute-cluster. These tri-clusters were then refined by *k*-means. By analyzing these irregular tri-clusters, patterns of 3D GTS were fully explored.

Using the same dataset in Chapters 2 and 3, but this time fitting it into a data cuboid with one spatial (station) and two nested temporal (year and day) dimensions, the tri-clustering analysis allowed the full exploration of patterns of intra-annual variability in temperature: six unique spatial patterns of intra-annual variability associated to four groups of years were found. It is important to point

out that BCAT_I is generic and can be applied to any 3D GTS. Thus it provides the possibility of exploring GTS with more than one attribute. Besides, the numbers of station- and year-clusters used in Chapter 3 are also adopted in Chapter 5 to facilitate the comparison of patterns. Moreover, similar to Chapter 4, Chapter 5 also refines the tri-clustering results. As such, similar values are exhaustively grouped into refined tri-clusters that allow the full exploration of the spatio-temporal patterns of intra-annual variability in Dutch temperatures.

6.3 Answers to research questions and general conclusion

RQ1: *How can a one-way clustering method be combined with geovisualization techniques to separately explore the spatial and temporal patterns from GTS? How does the MTUP affect the explored patterns?*

In Chapter 2, SOMs are selected as a one-way clustering method to explore synchronization, a specific spatio-temporal pattern. GTS with one attribute, in this case, the daily average temperatures at 28 Dutch weather stations from 1992 to 2011 are used as a case study after organizing it as a data table. On the one hand, regarding stations as rows and daily temperature for all years as columns, SOMs can be applied for spatial clustering. This analysis identifies synchronous stations (i.e. stations with similar temperature along all years). Then, U-matrix maps, SOMs cluster maps, trend plots, anomalies graphs and geographic maps can be used to visualize the spatially synchronous stations. On the other hand, regarding years (a calendar year is chosen as basic unit) as columns and daily temperature for all stations as rows, SOMs can also be applied for temporal clustering, to identify synchronous years. Then, the same visualization techniques except geographic maps can be used to display temporal patterns of synchronicity. As such, it is demonstrated that the combination of SOMs and geovisualization techniques can be used to explore both spatial and temporal patterns in a separate fashion.

The spatial and temporal clustering analyses are repeated at daily, weekly and monthly resolutions to examine MTUP effects. Results show that in terms of spatial synchronization, MTUP affects both elements and numbers of station-clusters. In terms of temporal synchronization, MTUP only affects the elements assigned to year-clusters. These results indicate that different temporal resolutions affect both spatial and temporal patterns explored from GTS.

RQ2: *How can a co-clustering method be combined with geovisualization techniques to concurrently explore the spatial and temporal patterns from GTS?*

Chapter 3 presents the use of a co-clustering method – the Bregman block average co-clustering algorithm with I-divergence (BBAC_I) to concurrently explore the spatial and temporal patterns from GTS. BBAC_I is a special case of a meta co-clustering algorithm, named Bregman co-clustering algorithm, which includes various loss function, e.g. squared Euclidean distance, to measure the distortion between the original and co-clustered data and also different sets of linear summary statistics of the original data, e.g. row/column averages, to be preserved in the co-clustered data. Chapter 3 chose I-divergence as the loss function because of its empirically proved superiority and co-cluster averages as the preserved statistics because the variations of values within each co-cluster along locations and timestamps need be considered. After an initial mapping of locations and timestamps, an objective function to measure the loss of information is minimized to find the optimal co-clusters. Various geovisualization techniques can then be used to display the patterns exhibited by these co-clusters: heatmaps can be used to show co-clusters straightforwardly; small multiples can be used to display the spatial and temporal varying behaviour present in GTS by visualizing the co-clusters in geographic maps; ringmaps can be used to display temporal varying behaviour associated to cyclic timestamps.

RQ3: *How can a co-clustering method and k-means be combined to enable the full exploration of spatio-temporal patterns from GTS?*

Chapter 4 presents an analytical approach that combines BBAC_I (the co-clustering method) and *k*-means to fully explore concurrent spatio-temporal patterns in GTS. BBAC_I can be first applied to GTS to identify a relative large number of homogenous location-timestamp co-clusters. Then *k*-means can be used to create irregular co-clusters by re-grouping the original co-clusters. By combining locations with the same variation of the attribute in the irregular co-clusters, spatial patterns can be extracted and displayed in small multiples. The temporal dynamics of these spatial patterns along all timestamps can be displayed in a linear timeline. Also by combining timestamps with the same variation of the attribute in the irregular co-clusters and by arranging them in a chronological way, temporal patterns can be extracted and displayed in a linear timeline. The spatial distributions of these temporal patterns can be displayed in another set of small multiples.

Three important parameters are needed by this approach: the numbers of location- and timestamp-clusters for BBAC_I, and the number of refined clusters

for *k*-means. The numbers of location- and timestamp-clusters can be empirically set by optimizing the objective function of the co-clustering algorithm for the specific dataset and the number of *k* can be set by methods that evaluate clustering results, e.g. the Silhouette method.

RQ4: *How to develop a tri-clustering algorithm that enables the extraction of patterns from GTS that fit into a data cuboid?*

Chapter 5 develops a new tri-clustering algorithm that allows the exploration of patterns in 3D GTS. This algorithm, named Bregman cuboid average tri-clustering algorithm (BCAT_I), is an extension of BBAC_I. Similar to BBAC_I, this algorithm measures the distortion between the original GTS and the tri-clustered ones with an objective function based on the I-divergence metric. Through an iterative process, BCAT_I minimizes the objective function and yields the optimal tri-clusters. These tri-clusters can then be refined by *k*-means to enable the full exploration of patterns from 3D GTS.

Using the same dataset as in Chapter 2 but fitting it into a data cuboid with one spatial and two nested temporal (year and day) dimensions, the tri-clustering analysis identified refined tri-clusters that were used to study intra-annual variability in Dutch daily temperatures. By analyzing these refined tri-clusters, six spatial patterns of intra-annual variability can be extracted and displayed in small multiples. Temporal patterns of intra-annual variability associated to four groups of years can also be extracted and displayed using timelines. As such, the complex patterns in 3D GTS can be fully explored.

The general conclusion of this thesis can be summarized as follows: Complex spatial and temporal patterns can be extracted from GTS by using one-way clustering, co-clustering and tri-clustering methods. Then these patterns can be displayed by using geovisualization techniques to facilitate their understanding and interpretation. As such, the clustering-based approaches contribute to a better understanding of complex patterns in GTS and to improved decision-makings.

6.4 Main contributions

Based on the challenges linked to the exploration of GTS (section 6.1), the main contributions of this PhD thesis are summarized below.

The first contribution is the development of an analytical approach based on SOMs that can be used to explore variations in spatial or temporal patterns caused by different temporal resolutions of the GTS. Previous studies only focused on exploring either spatial or temporal patterns without considering the effects of

MTUP. The analytical approach in Chapter 2 explored patterns at daily, weekly and monthly resolutions. It was shown that the MTUP affects the explored spatial or temporal patterns to different extents. This suggests that MTUP should be further studied in spatio-temporal analytics and that an appropriate temporal resolution should be selected depending on patterns to be explored for a specific task.

The second contribution is the introduction of the use of a co-clustering method for the exploration of concurrent spatial and temporal patterns from GTS. Traditional one-way clustering methods identify clusters of one dimension (e.g. location-clusters) that contain similar values along the other dimension (e.g. timestamps) and describe space- or time-varying behaviour in GTS. Unlike them, the introduced co-clustering method identifies location-timestamp co-clusters with similar values along these two dimensions. These co-clusters allow a better description of the space- and time-varying behaviour in the data.

The third contribution is the presentation of an analytical approach that combines a co-clustering method and *k*-means to fully explore concurrent spatio-temporal patterns in a large and complex dataset. Although the introduced co-clustering method enables the identification of homogenous location-timestamp co-clusters in GTS, the issue that different co-clusters might still have similar values needs to be solved. Chapter 4 used *k*-means to refine the co-clusters. By analyzing the refined co-clusters and using appropriate geovisualization techniques, the approach allows the simultaneous exploration of spatial patterns and their temporal dynamics and of temporal patterns and their spatial dynamics.

The fourth and final contribution of this thesis is the development of a tri-clustering algorithm that allows the exploration of 3D GTS. Both traditional clustering and co-clustering methods are incapable of fully analyzing this type of GTS. Chapter 5 presents a new tri-clustering algorithm, BCAT_I, to identify tri-clusters, which were then refined by *k*-means. The use of BCAT_I was tested by applying it to a particular GTS with one spatial and two nested hierarchical temporal dimensions. It was shown that this developed tri-clustering algorithm can be used to explore complex patterns, i.e. spatio-temporal patterns of intra-annual variability. This chapter proves the possibility of exploring patterns from GTS with more than one attributes and thus allows more information to be involved in the analysis. It also suggests that BCAT_I has the potential to analyze any 3D GTS, no matter the ones with more than one attributes or others with nested hierarchies in spatial or temporal dimension.

6.5 Future work

Several lines for future work can be recommended based on the performed work. For instance, to involve the human analytical skills in the exploration of GTS, a geovisual analytics framework can be developed from the clustering-based approaches in this thesis. Other lines are: the further study of MTUP and modifiable spatial and temporal unit problems (MSTUPs), the optimization of clusters numbers and the applicability of the clustering-based approaches to other types of spatio-temporal data.

6.5.1 The geovisual analytics framework

Focusing on the use of various clustering-based approaches to explore patterns from GTS, this PhD research only used static and stand-alone geovisualizations to represent the results. Although these graphic representations have proved to be effective, the use of interactive visual interfaces is suggested to facilitate the exploration by involving human analytical skills. A clustering-based geovisual analytics framework can thus be developed from this work for the interactive visual exploration of complex patterns in GTS.

To develop the geovisual analytics framework for the exploration of GTS, multiple linked visualizations and the interactive interfaces should be included in addition to the clustering-based approaches proposed in this thesis. The multiple linked visualizations instead of stand-alone ones should be used to help the understanding of the analytical process. Take the co-clustering analysis in Chapter 3 for example. To understand the spatial and temporal distributions of each co-cluster, the observations need be made back and forth between the heatmap (Figure 3.4) to visualize elements of co-clusters and the small multiples (Figure 3.5) to display spatial and temporal distributions. Multiple linked visualizations would make such an understanding easier: if a co-cluster is selected in the heatmap, the corresponding co-cluster would be highlighted in the small multiples to provide a direct view of its spatial and temporal distributions. Moreover, an interactive interface needs to be integrated in the geovisual analytics framework to allow the analyst to interact with other parts and thereby control the analytical process. Take the parameters of the clustering algorithms for example. The numbers of station- and year-clusters for the BCAT_I in Chapter 5 of this thesis are chosen based on the results analyzed using BBAC_I in Chapter 3. However, both the temporal resolutions of the dataset to be analyzed and the clustering algorithms have changed. An interactive interface would let the analyst to interactively choose the parameters depending on the clustering

algorithms and datasets. As such, the analytical process can be optimized and the mined patterns can be better explored by using a geovisual analytics framework.

6.5.2 MTUP and MSTUP

As illustrated in Chapter 2 and 3, MTUP affects the extracted pattern from GTS when they are explored at different temporal resolutions. However, as mentioned in (Coltekin et al. 2011), there are three aspects of MTUP: the temporal extent, the starting point and temporal resolution. Jong and Bruin (2012) and Cheng and Adepeju (2014) have analyzed and confirmed the influences of the three aspects of MTUP on trends extracted from vegetation time series and on clusters identified from spatio-temporal point data. Therefore, the temporal extent and starting timestamp of GTS should also be considered when exploring patterns using clustering-based approaches. To study the impacts of the temporal extent, the study period of the case studies could be shortened or lengthened. For instance, the study period from 1992 to 2011 used in Chapters 2, 3 and 5 can be changed to 1992 to 2010. Then the explored patterns of these two study periods can be compared. To study the impacts of the starting timestamp, different calendars can be used. The Gregorian calendar is used in this thesis, of which the beginning of each year is 1st of January. A different calendar, for instance, lunar calendar with the beginning of year depending on the movement of moon, can be used, and the resulting patterns can be compared.

MTUP concerns the temporal dimension while MAUP focuses on issues related to the spatial dimension, including spatial boundary, spatial zoning and spatial resolution. Consequently, the analysis of GTS needs to pay attention to the intersection of both MTUP and MAUP. This issue, called "Modifiable spatio-temporal unit problems" (MSTUPs; Cheng and Adepeju 2014), should be studied to examine how it affects the extracted patterns from GTS using clustering-based approaches and to help to select suitable spatio-temporal units for specific tasks. To this end, the combination of each aspect from MTUP and MAUP is suggested. For instance, to study the impacts of resolutions, both the original spatial and temporal resolutions of GTS can be aggregated to a coarser level (e.g. from day to month for temporal resolutions). Then the question of how resolutions affect the explored patterns can be answered. Similarly, different combinations of one or more aspects from MTUP and MAUP can be made and their impacts can be examined.

The study of MTUP and MSTUP can benefit from the geovisual analytics framework previously suggested. Through an interactive visual interface, the

analyst can examine and explore the changes of extracted patterns in real-time when interactively modifying different aspects of MTUP and MSTUP.

6.5.3 Optimization of clusters numbers

To assure the quality of the clustering results, the number of clusters should be optimized (Pelleg and Moore 2000). However, the number of clusters has been empirically set. For instance, the numbers of station- and year-clusters for co-clustering analysis in Chapter 3. Although these values have yielded reasonable results, optimizing the numbers is suggested by considering both the clustering algorithm, the specific dataset and the task at hand.

Several studies have worked on finding the optimal number of clusters for one-way clustering methods (Pelleg and Moore 2000, Berkhin 2006). Thus future work should be directed to the selection of the optimal clusters numbers for the co- and tri-clustering methods. The use of Bayesian information criterion (BIC) can assist the optimization of the numbers of clusters (Kass and Wasserman 1995, Cai et al. 2008). Besides, the optimization can be integrated into the recommended geovisual analytics framework, so that the analyst can find the optimum for various datasets, clustering algorithms and interactive tasks.

6.5.4 Application to other types of spatio-temporal data

Focusing on the exploration of GTS, the clustering-based approaches developed in this thesis can also be applied to other types of spatio-temporal data. There are five types of spatio-temporal data according to the classification of (Kisilevich et al. 2010) and trajectories are the one with the most complex form. Trajectories contain sequential values of one or more observed attributes of a moving object. Compared with GTS, they are usually collected at changing locations and at non-uniform temporal intervals.

Previous studies on clustering analysis of trajectories all used one-way clustering (Li et al. 2004, Giannotti et al. 2007, Adrienko and Adrienko 2011). That is, they only considered the space-varying behaviour of the moving objects (e.g. human). Other clustering-based approaches, co-clustering, for instance, should be applied to trajectories to explore the space- and time-varying behaviour present in human movements. Thus future work is directed to adapt BBAC_I so that the concurrent mapping will be from various locations (instead of fixed locations) to location-clusters and from timestamps with non-uniform intervals (instead of uniform intervals) to timestamp-clusters. Similarly, patterns in other types of data can be explored in the future work by adapting clustering-based

approaches developed in this thesis to help a better understanding of spatio-temporal data.

Appendix

Bregman cuboid average tri-clustering algorithm with I-divergence (BCAT_I)

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 $\mathbf{R} \in \mathbb{R}^{m \times h}$, $\mathbf{C} \in \mathbb{R}^{n \times l}$ and $\mathbf{T} \in \mathbb{R}^{v \times g}$, 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 tri-clustering:

$$f_{loss} = I(S; Y; D) - I(\hat{S}; \hat{Y}; \hat{D}) \quad \text{(A1)}$$

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, firstly the original temperature cuboid is reshaped as a data matrix $\mathbf{O}' \in \mathbb{R}^{m \times vn}$ 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 $o'_{s,dy}$ ($dy \in \{1, \dots, v \times n\}$) as elements focuses on each station while retaining information in all days and years. Correspondingly, the loss function in equation A1 can be reformulated as the loss of mutual information before and after mapping the reshaped data matrix:

$$f_{loss} = I(S; DY) - I(\hat{S}; \widehat{DY}) \quad \text{(A2)}$$

BCAT_I measures the loss of mutual information using I-divergence. Therefore, the loss function in equation A2 can be represented as I-divergence between this reshaped data matrix and a matrix that approximates it:

$$I(S; DY) - I\left(\hat{S}; \widehat{DY}\right) = D_I(\boldsymbol{O}' \,||\boldsymbol{A}) \quad (3)$$

Where $D_I(\cdot||\cdot)$ is the I-divergence between two elements (i.e. data matrices). $\mathbf{A} \in \mathbb{R}^{m \times vn}$ is the approximation matrix of $\mathbf{O}'$ with $a_{s,dy}$ as elements. The calculation of **A** is determined by $\mathbf{O}'$, the current mapping and the statistics of $\mathbf{O}'$ to be preserved in the approximation. Due to the reshaping, the current mapping for columns from $\mathbf{O}'$ to $\widehat{\mathbf{O}'}$ is the repetition of T for *n* times because of the data arrangement of columns in $\mathbf{O}'$, and named $\mathbf{T1} \in \mathbb{R}^{vn \times g}$. Since the statistics of **O** to be preserved during tri-clustering are tri-cluster averages to consider variations along stations, years and days, the averages of elements

within the same station- and day/year-clusters that are to be preserved in **A**. **A** is calculated as:

$$\boldsymbol{A} = \boldsymbol{R}\,\widehat{\boldsymbol{O}'}\,\boldsymbol{T}\boldsymbol{1}^{\boldsymbol{T}} \tag{A4}$$

Where $\widehat{\mathbf{O}'}$ is the clustered matrix of $\mathbf{O}'$ and calculated as averages of elements that are intersected by each station- and day/year-clusters; $\mathbf{T1}^{\mathbf{T}}$ denotes the transpose of $\mathbf{T1}$.

Then the I-divergence between the reshaped data matrix and its approximation matrix in equation A3 can be further represented as:

$$D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) = \sum_{\hat{s}} \sum_{\widehat{dy}} \sum_{s\in\hat{s}} \sum_{dy\in\widehat{dy}} o'_{s,dy} log \frac{o'_{s,dy}}{a_{s,dy}} \tag{A5}$$

According to the illustration in (Banerjee et al. 2007), equation A5 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:

$$D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) = \sum_{\hat{s}} \sum_{s\in\hat{s}} o'_{s,\cdot} log \frac{o'_{s,\cdot}}{a_{s,\cdot}} \tag{A6}$$

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 A6, which yields the optimal station clustering result encoded in $\mathbf{R}^*$ in a binary way.

$$j^*(\cdot) = argmin_{j\in[1,h]} D_{I\,i,j}\ \ [i]_1^m \tag{A7}$$

After that, the mapping from years to year-clusters is optimized. Firstly **O** is reshaped as a data matrix $\mathbf{O}' \in \mathbb{R}^{n\times mv}$ with $o'_{y,sd}$ $(sd \in \{1,\dots,m\times v\})$ 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 A1 is re-expressed as the loss of mutual information before and after clustering the reshaped data matrix:

$$f_{loss} = I(Y;SD) - I(\hat{Y};\widehat{SD}) \tag{A8}$$

Measured with I-divergence by BCAT_I, the loss function in equation A8 can be re-expressed as I-divergence between this $\mathbf{O}'$ and its approximation matrix:

$$I(Y;SD) - I\big(\hat{Y};\widehat{SD}\big) = D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) \tag{A9}$$

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

$$\boldsymbol{A} = \boldsymbol{C}\,\widehat{\boldsymbol{O}'}\,\boldsymbol{R1^T} \tag{A10}$$

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

Then equation A9 can be further expressed as:

$$D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) = \sum_{\hat{y}} \sum_{\widehat{sd}} \sum_{y \in \hat{y}} \sum_{sd \in \widehat{sd}} o'_{y,sd} log \frac{o'_{y,sd}}{a_{y,sd}} \tag{A11}$$

Equation A11 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:

$$D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) = \sum_{\hat{y}} \sum_{y \in \hat{y}} o'_{y,\cdot} log \frac{o'_{y,\cdot}}{a_{y,\cdot}} \tag{A12}$$

Since the mapping from each year to different year-clusters results in different values of the loss function in equation A12, the optimization is achieved by minimizing the loss function for the mapping of each year. Such optimization is done according to equation A13. Thus, the mapping from years to year-clusters is updated and the optimal year clustering result is produced and saved in $\mathbf{C}^*$ with binary encoding:

$$q^*(\cdot) = argmin_{q \in [1,l]} D_{I_{p,q}}\ [p]_1^n \tag{A13}$$

The last step is to optimize the mapping from v days to g day-clusters. Firstly **O** is reshaped as a data matrix $\mathbf{O}' \in \mathbb{R}^{v \times nm}$ 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 A1 is reformulated in this step as the loss of mutual information before and after clustering:

$$f_{loss} = I(D; YS) - I(\widehat{D}; \widehat{YS}) \tag{A14}$$

Due to the I-divergence measure used by BCAT_I, equation A14 can be reformulated as:

$$I(D; YS) - I(\widehat{D}; \widehat{YS}) = D_I(\boldsymbol{O}'\ ||\boldsymbol{A}) \tag{A15}$$

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

$$\boldsymbol{A} = \boldsymbol{T}\widehat{\boldsymbol{O}'}\boldsymbol{C1^T} \tag{A16}$$

Where $\widehat{\mathbf{O}'}$ is the clustered matrix of $\mathbf{O}'$ and calculated as average values of elements that are intersected by each day- and year/station-clusters. The current column mapping from $\mathbf{O}'$ 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 $\mathbf{O}'$. $\mathbf{C1^T}$ is the transpose 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:

$$D_I(\boldsymbol{O}' \,||\boldsymbol{A}) = \sum_{\hat{d}} \sum_{\widehat{ys}} \sum_{d \in \hat{d}} \sum_{ys \in \widehat{ys}} o'_{d,ys} log \frac{o'_{d,ys}}{a_{d,ys}} \tag{A17}$$

Then equation A17 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:

$$D_I(\boldsymbol{O}' \,||\boldsymbol{A}) = \sum_{\hat{d}} \sum_{d \in \hat{d}} o'_{d,\cdot} log \frac{o'_{d,:}}{a_{d,\cdot}} \tag{A18}$$

The last step of optimization is minimizing the loss function in equation A18 for each day cluster assignment. Therefore, the mapping from days to day-clusters is updated according to equation A19, which yields the optimal day clustering result and encoded in $\mathbf{T}^*$ in the binary way.

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

Finally the loss in mutual information is re-calculated with the updated mapping $\mathbf{R}^*$, $\mathbf{C}^*$, $\mathbf{T}^*$ according to the loss function in equation A1 until convergence (i.e. the loss achieves a local minimum).

Bibliography

Adrienko, N. & G. Adrienko (2011) Spatial generalization and aggregation of massive movement data. *Visualization and Computer Graphics, IEEE Transactions on,* 17**,** 205-219.

Agarwal, P. & A. Skupin. 2008. *Self-organising maps: Applications in geographic information science*. John Wiley & Sons Ltd.

Ahas, R. & A. Aasa. 2003. Developing comparative phenological calendars. In *Phenology: An Integrative Environmental Science,* ed. M. D. Schwartz, 301-318. Springer Netherlands.

Ahas, R., A. Aasa, A. Menzel, V. Fedotova & H. Scheifinger (2002) Changes in European spring phenology. *International Journal of Climatology,* 22**,** 1727-1738.

Ahas, R., A. Aasa, S. Silm & J. Roosaare (2007) Seasonal indicators and seasons of Estonian landscapes. *Landscape Research,* 30**,** 173-191.

Anagnostopoulos, A., A. Dasgupta & R. Kumar. 2008. Approximation algorithms for co-clustering. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,* 201-210. ACM.

Andrienko, G. & N. Andrienko (2010) Space, time and viusal analytics. *International Journal of Geographical Information Science,* 24**,** 1577-1600.

Andrienko, G., N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak & D. Keim (2010) Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum,* 29**,** 913-922.

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

Arbia, G. & F. Petrarca (2011) Effects of MAUP on spatial econometric models. *Letters in Spatial and Resource Sciences,* 4**,** 173-185.

Ault, T. R., R. Zurita-Milla & M. D. Schwartz (2015) A Matlab© toolbox for calculating spring indices from daily meteorological data. *Computers & Geosciences,* 83**,** 46-53.

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

Basler, D. & C. Körner (2012) Photoperiod sensitivity of bud burst in 14 temperate forest tree species. *Agricultural and Forest Meteorology,* 165**,** 73-81.

Ben-Dor, A., R. Shamir & Z. Yakhini (1999) Clustering gene expression patterns. *Journal of computational biology,* 6**,** 281-297.

Berkhin, P. 2006. A survey of clustering data mining techniques. In *Grouping Multidimensional Data: Recent Advances in Clustering*, 25--71.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett & P. D. Jones (2006) Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research: Atmospheres,* 111**,** D12106.

Cai, R., L. Lu & A. Hanjalic (2008) Co-clustering for auditory scene categorization. *IEEE TRANSACTIONS ON MULTIMEDIA,* 10**,** 596-606.

Carrel, M., M. Emch, P. K. Streatfield & M. Yunus (2009) Spatio-temporal clustering of cholera: the impact of flood control in Matlab, Bangladesh, 1983-2003. *Health Place,* 15**,** 741-52.

Cheng, T. & M. Adepeju (2014) Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection.

Cheng, Y. & G. M. Church. 2000. Biclustering of expression data. In *Proceedings ISMB*, 93-103. AAAI Press.

Chmielewski, F.-M. & T. Rötzer (2001) Response of tree phenology to climate change across Europe. *Agricultural and Forest Meteorology,* 108**,** 101-112.

Cho, H., I. S. Dhillon, Y. Guan & S. Sra. 2004. Minimum sum-squared residue co-clustering of gene expression data. In *Fourth SIAM Int'l Conf. Data Mining*.

Chuine, I. (2000) A unified model for budburst of trees. *Journal of Theoretical Biology,* 207**,** 337-347.

Coltekin, A., S. D. Sabbata, C. Willi, I. Vontobel, S. Pfister, M. Kuhn & M. Lacayo. 2011. Modifiable temporal unit problem. In *ISPRS/ICA workshop "Persistent problems in geographic visualization" (ICC2011)*. Paris, France.

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

Dark, S. J. & D. Bram (2007) The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography,* 31**,** 471-479.

Deng, M., Q. Liu, J. Wang & Y. Shi (2011) A general method of spatio-temporal clustering analysis. *Science China Information Sciences,* 56**,** 1-14.

Dhillon, I. S., S. Mallela & D. S. Modha. 2003. Information-theoretic co-clustering. In *The 9th International Conference on Knowledge Discovery and Data Mining (KDD)*, 89-98.

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

Doi, H. & I. Katano (2008) Phenological timings of leaf budburst with climate change in Japan. *Agricultural and Forest Meteorology,* 148**,** 512-516.

Dungan, J. L., J. Perry, M. Dale, P. Legendre, S. Citron‐Pousty, M. J. Fortin, A. Jakomulska, M. Miriti & M. Rosenberg (2002) A balanced view of scale in spatial statistical analysis. *Ecography,* 25**,** 626-640.

Dykes, J., A. M. MacEachren & M.-J. Kraak. 2005. *Exploring geovisualization*. Elsevier.

EEA. 2012. Climate change, impacts and vulnerability in Europe 2012, an indicator-based report. 1-304. Copenhagen, Denmark: European Environment Agency.

Estrella, N., T. Sparks & A. Menzel (2007) Trends and temperature response in the phenology of crops in Germany. *Global Change Biology,* 13**,** 1737-1747.

Fayyad, U. M. (1996) Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems***,** 20-25.

Fu, Y. H., S. Piao, M. Op de Beeck, N. Cong, H. Zhao, Y. Zhang, A. Menzel & I. A. Janssens (2014) Recent spring phenology shifts in western Central Europe based on multiscale observations. *Global Ecology and Biogeography,* 23**,** 1255-1263.

Gahegan, M. & B. Brodaric. 2002. Computational and visual support for geographical knowledge construction: Filling in the gaps between exploration and explanation. In *Advances in Spatial Data Handling*, 11-25. Springer.

Giannotti, F., M. Nanni, F. Pinelli & D. Pedreschi. 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 330-339. ACM.

Goodchild, M. F. (2009) Geographic information systems and science: today and tomorrow. *Annals of GIS,* 15**,** 3-9.

Gordo, O. & J. J. Sanz (2010) Impact of climate change on plant phenology in Mediterranean ecosystems. *Global Change Biology,* 16**,** 1082-1106.

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

Gu, Y., J. F. Brown, T. Miura, W. J. D. v. Leeuwen & B. C. Reed (2010) Phenological classification of the United States: A geographic framework for extending multi-sensor time-series data. *Remote Sensing* 2**,** 526-544.

Guo, D. 2003. Human-machine collabration for geographic knowledge discovery with high-dimensional clustering. In *The Pennsylvania State University, Department of Geography*.

---. 2009. Multivariate spatial clustering and geovisualization. In *Geographic Data Mining and Knowledge Discovery - 2nd Edition*, eds. H. J. Miller & J. Han, 325-345. London: Taylor & Francis Group.

Guo, D., J. Chen, A. M. Maceachren & K. Liao (2006) A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics,* 12**,** 1461-1474.

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

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

Han, J., M. Kamber & J. Pei. 2011. *Data mining: concepts and techniques: concepts and techniques*. Elsevier.

Han, J., J.-G. Lee & M. Kamber. 2009. An overview of clustering methods in geographic data analysis. In *Geographic Data Mining and Knowledge Discovery,* eds. H. J. Miller & J. Han, 150-187. New York: Taylor & Francis Group.

Hansen, J., R. Ruedy, M. Sato & K. Lo (2010) Global surface temperature change. *Reviews of Geophysics,* 48**,** RG4004.

Harinarayan, V., A. Rajaraman & J. D. Ullman (1996) Implementing data cubes efficiently. *ACM SIGMOD Record,* 25**,** 205-216.

Harris, R. L. 1999. *Information graphics: A comprehensive illustrated reference*. New York, NY, USA: Oxford University Press.

Hartigan, J. A. (1972) Direct clustering of a data matrix. *Journal of American Statistical Association,* 67**,** 123-129.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones & M. New (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research,* 113.

Helbich, M., W. Brunauer, J. Hagenauer & M. Leitner (2013) Data-driven regionalization of housing markets. *Annals of the Association of American Geographers,* 103**,** 871-889.

Hofmann, T. (2004) Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS),* 22**,** 89-115.

Hsu, K.-C. & S.-T. Li (2010) Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources,* 33**,** 190-200.

Hudson, I. L., M. R. Keatley & S. Y. Lee (2011) Using self-organising maps (SOMs) to assess synchronies: an application to historical eucalypt flowering records. *International Journal of Biometeorology,* 55**,** 879-904.

IPCC. 2013. IPCC Synthesis Report.

Jaagus, J. & R. Ahas (2000) Space-time variations of climatic seasons and their correlation with the phenological development of nature in Estonia. *Climate Research,* 15**,** 207-219.

Jain, A. K., M. N. Murty & P. J. Flynn (1999) Data clustering: a review. *ACM Comput. Surv.,* 31**,** 264-323.

Jelinski, D. E. & J. Wu (1996) The modifiable areal unit problem and implications for landscape ecology. *Landscape ecology,* 11**,** 129-140.

Jeong, S. J., H. Chang‐Hoi, H. J. Gim & M. E. Brown (2011) Phenology shifts at start vs. end of growing season in temperate vegetation over the Northern Hemisphere for the period 1982‐2008. *Global Change Biology,* 17**,** 2385-2399.

Ji, L., K. L. Tan & A. K. H. Tung. 2006. Mining frequent closed cubes in 3D datasets. In *The 32nd international conference on Very large data bases*, 811--822.

Jones, P. D. & A. Moberg (2003) Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *Journal of Climate,* 16**,** 206-223.

Jong, R. d. & S. d. Bruin (2012) Linear trends in seasonal vegatation time series and the modifiable temporal unit problem. *Biogeosciences,* 9**,** 71-77.

--- (2012) Linear trends in seasonal vegetation time series and the modifiable temporal unit problem. *Biogeosciences,* 9**,** 71-77.

Jong, R. d., J. Verbesselt, A. Zeileis & M. E. Schaepman (2013) Shifts in global vegetation activity trends. *Remote Sensing,* 5**,** 1117-1133.

Kass, R. E. & L. Wasserman (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the american statistical association,* 90**,** 928-934.

Keeling, C. D., J. F. S. Chin & T. P. Whorf (1996) Increased activity of northern vegetation inferred from atmospheric CO2 measurements. *Nature,* 382**,** 146-149.

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

Kisilevich, S., F. Mansmann, M. Nanni & S. Rinzivillo. 2010. Spatio-temporal clustering. In *Data Mining and Knowledge Discovery Handbook,* eds. O. Maimon & L. Rokach, 855-874. Springer US.

Kohonen, T. 2001. *Self-organizing maps*. Springer-Verlag Berlin Heidelberg.

Kohonen, T., J. Hynninen, J. Kangas & J. Laaksonen. 1996. SOM PAK: The Self-Organizing Map Program Package.

Koua, E. L. 2005. Computation and visual support for exploratory geovisualization and knowledge construction. In *International Institute for Geo-information Science (ITC)&Faculty of Geographical Sciences, Utrecht University*. Utrecht University.

Kraak, M.-J. (2000) About maps, cartography, geovisualization and other graphics. *Geoinformatics Journal,* 3.

--- (2003) Geovisualization illustrated. *ISPRS Journal of Photogrammetry and Remote Sensing,* 57**,** 390-399.

Kraak, M.-J., R. Edsall & A. M. MacEachren. 1997. Cartographic animation and legends for temporal maps: Exploration and or interaction. In *The 18th International Cartographic Conference*, 253-261.

Kraak, M. J. 2005. Timelines, temporal resolution, temporal zoom and time geography. In *the 22nd International Cartographic Conference*. A Coruna Spain.

Kramer, K. & H. Hanninen. 2009. The annual cycle of development of trees and process-based modeling of growth to scale up from the tree to the stand. In *Phenology of Ecosystem Processes*, ed. A. Noormets, 201-227. New York: Springer.

Kriegel, H.-P., P. Kröger & A. Zimek (2009) Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 3**,** 1.

Kumar, J., R. T. Mills, F. M. Hoffman & W. W. Hargrove (2011) Parallel *k*-means clustering for quantitative ecoregion delineation using large data sets. *Procedia Computer Science,* 4**,** 1602-1611.

Lenderink, G., H. Mok, T. Lee & 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**,** 3033-3041.

Lewis, J. M., M. Ackerman & V. R. d. Sa. 2012. Human cluster evaluation and formal quality measures: A comparative study. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)*, 1870-1875. Citeseer.

Li, S.-T. & S.-W. Chou. 2000. Multi-resolution spatio-temoral data mining for the study of air pollutant regionalization. In *the 33rd Hawaii International Conference on System Sciences*. Hawaii, USA.

Li, X. & M.-J. Kraak. 2012. Exploring multivariable spatio-temporal data with the time wave: case study on meteorological data. In *Advances in Spatial Data Handling and GIS. Lecture notes in geoinformation and cartography, part 3.*, ed. A. Y. E. al., 79-92. Heiderlberg: Springer-Verlag Berlin.

Li, Y., J. Han & J. Yang. 2004. Clustering moving objects. In *The tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 617-622.

Lieth, H. 1974. Purposes of a phenology book. In *Phenology and Seasonality Modeling*, ed. H. Lieth, 3-19. Springer Berlin Heidelberg.

Maceachren, A., X. Dai, F. Hardisty, D. Guo & G. Lengerich. 2003. Exploring high-D spaces with multiform matrices and small multiples. In *the International Symposium on Information Visualization*, 31-38. Seattle.

MacEachren, A. M. & I. Brewer (2004) Developing a conceptual framework for visually-enabled geocollaboration. *International Journal of Geographical Information Science,* 18**,** 1-34.

MacEachren, A. M. & M.-J. Kraak (2001) Research challenges in geovisualization. *Cartography and Geographic Information Science,* 28**,** 3-12.

Madeira, S. C. & A. L. Oliveira (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* 1**,** 24-45.

Matsumoto, K., T. Ohta, M. Irasawa & T. Nakamura (2003) Climate change and extension of the Ginkgo biloba L. growing season in Japan. *Global Change Biology,* 9**,** 1634-1642.

Menzel, A. (2000) Trends in phenological phases in Europe between 1951 and 1996. *International Journal of Biometeorology,* 44**,** 76-81.

Menzel, A., T. H. Sparks, N. Estrella, E. Koch & etc (2006) European phenological response to climate change matches the warming pattern. *Global Change Biology,* 12**,** 1969-1976.

Miller, H. J. & J. Han. 2009. Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery - 2nd Edition,* eds. H. J. Miller & J. Han, 1-26. London: Taylor & Francis Group.

Mills, R. T., F. M. Hoffman, J. Kumar & W. W. Hargrove (2011) Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats. *Procedia Computer Science,* 4**,** 1612-1621.

Openshaw, S. & P. J. Taylor (1979) A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences,* 21**,** 127-144.

Pan, G., G. Qi, W. Zhang, S. Li, Z. Wu & L. T. Yang (2013) Trace analysis and mining for smart cities: issues, methods, and applications. *IEEE Communications Magazine,* 121**,** 120-126.

Parmesan, C. & G. Yohe (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature,* 421**,** 37-42.

Pelleg, D. & A. Moore (2000) Xmeans: Extending k-means with efficient estimation of the number of clusters. *ICML,* 1.

Pensa, R. G. & J.-F. o. Boulicaut. 2008. Constrained co-clustering of gene expression data. In *International Conference on Data Mining SDM'08*. Atlanta, USA.

Peñuelas, J., T. Rutishauser & I. Filella (2009) Phenology feedbacks on climate change. *Science,* 324**,** 887.

Polgar, C. A. & R. B. Primack (2011) Leaf‐out phenology of temperate woody plants: from trees to ecosystems. *New Phytologist,* 191**,** 926-941.

Qiu, G. 2004. Image and feature co-clustering. In *The 17th International Conference on Pattern Recognition.*, 991-994. IEEE.

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.

Schwartz, M. D. 1997. Spring index models: An approach to connecting satellite and surface phenology. In *Phenology in Seasonal Climates I,* eds. L. H & D. Schwartz, 23-38. Leiden: Backhuys.

--- (1998) Green-wave phenology. *Nature,* 394**,** 839-840.

Schwartz, M. D., R. Ahas & A. Aasa (2006) Onset of spring starting earlier across the Northern Hemisphere. *Global Change Biology,* 12**,** 343-351.

Schwartz, M. D., T. R. Ault & J. L. Betancourt (2013) Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *International Journal of Climatology,* 33**,** 2917-2922.

Schwartz, M. D. & X. Chen (2002) Examining the onset of spring in China. *Climate Research,* 21**,** 157-164.

Schwartz, M. D., B. C. Reed & M. A. White (2002) Assessing satellite-derived start-of-season measures in the conterminous USA. *International Journal of Climatology,* 22**,** 1793-1805.

Schwartz, M. D. & B. E. Reiter (2000) Changes in North American spring. *International Journal of Climatology,* 20**,** 929-932.

Shekhar, S., R. Vatsavai & S. Chawla (2009) Spatial classification and prediction models for geospatial data mining. *Geographic Data Mining and Knowledge Discovery***,** 117-147.

Sim, K., Z. Aung & V. Gopalkrishnan. 2010. Discovering correlated subspace clusters in 3D continuous-valued data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on,* 471-480.

Sim, K., V. Gopalkrishnan, A. Zimek & G. Cong (2013) A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery,* 26**,** 332-397.

Smith, T. M., R. W. Reynolds, T. C. Peterson & J. Lawrimore (2008) Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate,* 21**,** 2283-2296.

Steinbach, M., G. Karypis & V. Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining,* 525-526. Boston.

Stöckli, R. & P. L. Vidale (2004) European plant phenology and climate as seen in a 20-year AVHRR land-surface parameter dataset. *International Journal of Remote Sensing,* 25**,** 3303-3330.

Studer, S., R. Stockli, C. Appenzeller & P. L. Vidale (2007) A comparative study of satellite and ground-based phenology. *International Journal of Biometeorology,* 51**,** 405-14.

Swayne, D. F., D. T. Lang, A. Buja & D. Cook (2003) GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis,* 43**,** 423-444.

Takamura, H. & Y. Matsumoto. 2002. Two-dimensional clustering for text categorization. In *The 6th Conference on Natural Language Learning,* 1-7. Association for Computational Linguistics.

van Vliet, A. J., R. S. de Groot, Y. Bellens, P. Braun, R. Bruegger, E. Bruns, J. Clevers, C. Estreguil, M. Flechsig & F. Jeanneret (2003) The European phenology network. *International Journal of Biometeorology,* 47**,** 202-212.

Verbesselt, J., R. Hyndman, G. Newnham & D. Culvenor (2010) Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment,* 114**,** 106-115.

Walther, G.-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg & F. Barilein (2002) Ecological responses to recent climate change. *Nature,* 416**,** 389-395.

Wan, L.-H., Y.-J. Li, W.-Y. Liu & D.-Y. Zhang. 2005. Application and study of spatial cluster and customer partitioning. In *International Conference on Machine Learning and Cybernetics*, 1701-1706. IEEE.

White, M. A., K. M. de Beurs, K. Didan, D. W. Inouye, A. D. Richardson, O. P. Jensen, J. O'Keefe, G. Zhang, R. R. Nemani, W. J. D. van Leeuwen, J. F. Brown, A. de Wit, M. Schaepman, X. Lin, M. Dettinger, A. S. Bailey, J. Kimball, M. D. Schwartz, D. D. Baldocchi, J. T. Lee & W. K. Lauenroth (2009) Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982-2006. *Global Change Biology,* 15**,** 2335-2359.

White, M. A., F. Hoffman, W. W. Hargrove & R. R. Nemani (2005) A global framework for monitoring phenological responses to climate change. *Geophysical Research Letters,* 32**,** L04705.

Williams, S. E. & J.-M. Hero (2001) Multiple determinants of Australian tropical frog biodiversity. *Biological Conservation,* 98**,** 1-10.

Williams, S. E. & J. Middleton (2008) Climatic seasonality, resource bottlenecks, and abundance of rainforest birds: implications for global climate change. *Diversity and Distributions,* 14**,** 69-77.

Wu, T., G. Song, X. Ma, K. Xie, X. Gao & X. Jin (2008) Mining geographic episode association patterns of abnormal events in global earth science data. *Science in China Series E: Technological Sciences,* 51**,** 155-164.

Wu, X., R. Zurita-Milla & M.-J. Kraak (2013) Visual discovery of synchronization in weather data at multiple temporal resolutions. *The Cartographic Journal,* 50**,** 247-256.

--- (2015) Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science,* 29**,** 624-642.

--- (2016) A novel analysis of spring phenological patterns over Europe based on co-clustering *Journal of Geophysical Research: Biogeosciences,* under review.

Xu, X., Y. Lu, K.-L. Tan & A. K. Tung. 2009. Finding time-lagged 3D clusters. In *IEEE 25th International Conference on Data Engineering*, 445-456.

Zhang, P., Y. Huang, S. Shekhar & V. Kumar. 2003. Correlation analysis of spatial time series datasets: a filter-and-refine approach. In *Advances in*

Knowledge Discovery and Data Mining, eds. K.-Y. Whang, J. Jeon, K. Shim & J. Srivastava, 532-544. Springer Berlin Heidelberg.

Zhang, X., M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. F. Hodges, F. Gao, B. C. Reed & A. Huete (2003) Monitoring vegetation phenology using MODIS. *Remote Sensing of Environment,* 84**,** 471-475.

Zhang, X., D. Tarpley & J. T. Sullivan (2007) Diverse responses of vegetation phenology to a warming climate. *Geophysical Research Letters,* 34.

Zhang, Y., G. F. Hepner & P. E. Dennison (2012) Delineation of phenoregions in geographically diverse regions using k-means++ clustering: a case study in the Upper Colorado River Basin. *GIScience & remote sensing,* 49**,** 163.

Zhao, J., P. Forer & A. S. Harvey (2008) Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization,* 7**,** 198-209.

Zhao, L. & M. J. Zaki. 2005. TRICLUSTER: An effective algorithm for mining coherent clusters in 3D microarray data. In *Proc. of the 2005 ACM SIGMOD International Conference on Management of Data*, 694--705.

Zhou, J. & K. Ashfaq. 2006. ParRescue: Scalable parallel algorithm and implementation for biclustering over large distributed datasets. In *26th IEEE International Conference on Distributed Computing System.*

Zirlewagen, D. & K. Von Wilpert (2010) Upscaling of environmental information: support of land-use management decisions by spatio-temporal regionalization approaches. *Environmental Manage,* 46**,** 878-93.

Zurita-Milla, R., J. A. E. van Gijsel, N. A. S. Hamm, P. W. M. Augustijn & A. Vrieling (2013) Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. *Geoscience and Remote Sensing, IEEE Transactions on,* 51**,** 1914-1921.

Summary

Large volumes of spatio-temporal data are becoming increasingly available due to advancements in modern data collection and data sharing techniques. Mining patterns from these data helps to provide useful information and this, in turn, helps decision-making in many application areas. This research focuses on the analysis of one important type of spatio-temporal data, geo-referenced time series (GTS). GTS contain time evolving sequences for observed attributes that are measured at fixed locations and typically uniform temporal intervals. Examples of GTS are time series of weather variables collected at meteorological stations or satellite image time series.

As one important task of data mining, clustering provides an overview of data elements at a higher level of abstraction (i.e. the clusters) and also allows the analysis of details when focusing on particular clusters. As such, clustering analysis can be used to extracts patterns from GTS. Yet, the spatial, temporal and attribute information contained in the resulting clusters often makes it difficult to understand and interpret the extracted patterns. In this case, graphic representations provided by geovisualization can be used to stimulate the visual thinking of the results and to help to explore spatial and temporal patterns from GTS.

Despite previous studies on combining clustering and geovisualization, several challenges still exist in the exploration of complex spatial and temporal patterns in GTS. For instance: (1) the exploration of differences in spatial and/or temporal patterns caused by the so-called Modifiable Temporal Unit Problem (MTUP); (2) the exploration of concurrent spatial and temporal patterns; (3) the full exploration of spatio-temporal patterns from complex datasets; and (4) the exploration of patterns from 3D GTS.

This PhD thesis addresses the aforementioned challenges. More precisely, this thesis proposed and developed approaches based on one-way clustering, co-clustering and tri-clustering methods that enable the extraction of complex patterns from GTS, and then employed suitable geovisualization techniques to visualize the results.

To explore differences in the spatial and/or temporal patterns caused by MTUP, an analytical approach was developed to explore spatial and temporal synchronization, a specific pattern, at different temporal resolutions. This analytical approach combines self-organizing maps (SOMs), U-matrix maps, SOMs cluster maps and trend plots. Using Dutch daily average temperatures, the approach independently explored spatial and temporal synchronization in

temperature at daily, weekly and monthly resolutions: identify station-clusters with similar temperature distributions along all years and year-clusters (a calendar year is used as basic unit) with similar temperature behaviour along all stations. Results showed that both spatial and temporal synchronizations were affected by MTUP: variations were found in both elements and numbers of station-clusters and the elements of year-clusters at different temporal resolutions.

To concurrently explore spatial and temporal patterns, the Bregman block average co-clustering algorithm with I-divergence (BBAC_I) was introduced. In particular, BBAC_I was used to identify groups of data (co-clusters) that contain similar values over both the spatial and the temporal dimensions of the data. Then heatmaps, small multiples and ringmaps were used to visualize these co-clusters. Daily average temperatures collected at 28 Dutch meteorological stations from 1992 to 2011 were analyzed at yearly, monthly and daily resolutions. The results pointed out regions as well as subsets of years/months/days that have similar temperatures. Results showed a decreasing temperature pattern from southwest to northeast of the Netherlands and from "cold" to "hot" years/months/days. This analysis also confirmed the effect of different temporal resolutions: the finer the temporal resolution is, the more complex the patterns become.

To fully explore the concurrent spatio-temporal patterns, an analytical approach was developed by combining the BBAC_I and the *k*-means clustering algorithms. The co-clusters identified using BBAC_I were refined using *k*-means by grouping those with similar values. After that, a heatmap and small multiples, and two sets of small multiples and timelines were used to display the refined co-clusters. Together with a temperature-driven phenological model, this approach was applied to a European temperature dataset (1950 to 2011 and a resolution of 0.25 degrees) to explore spring phenological patterns. Results identified four spatial phenological patterns and their temporal dynamics, which indicate that the first years of the study period tend to have very late springs and recent years have early springs. Results also identified twelve main temporal phenological patterns and their spatial dynamics, which indicate that western Turkey has the most intense variable springs. These results indicate that the proposed approach enables the full exploration of spatio-temporal patterns in spring phenology, by allowing the simultaneous exploration of both spatial patterns and their temporal dynamics and of temporal patterns and their spatial dynamics.

To explore patterns from 3D GTS, the Bregman average cuboid tri-clustering algorithm with I-divergence (BCAT_I) was developed as an extension of BBAC_I. By concurrently grouping GTS along three dimensions, BCAT_I

enables the analysis of 3D GTS and identifies tri-clusters. Like with BBAC_I, *k*-means was used to refine the tri-clusters. Applied to daily average temperature collected at 28 Dutch meteorological stations from 1992 to 2011, the tri-clustering analysis identified refined tri-clusters that contain similar temperatures along the spatial (stations) and two nested temporal dimensions (years and days). Both 3D and 2D heatmaps, small multiples and timelines were used to display the patterns of intra-annual temperature variability that can be analyzed from the refined tri-clusters. The results showed six unique spatial patterns of intra-annual temperature variability that were associated to four groups of years. A further analysis of these patterns revealed an intense variability in spring and winter temperatures after 1996 at the northeast & center of the Netherlands. Such a variability also exists in spring temperatures at the southeast of the country. Besides, the tri-clustering analysis also revealed homogeneous summer temperatures across the country.

To conclude, this thesis presents the combination of clustering methods and geovisualization techniques for the exploration of GTS. Depending on the patterns to be extracted, the clustering methods range from a one-way clustering to the newly developed tri-clustering one. By combining appropriate geovisualization techniques, these clustering-based approaches enable the exploration of complex spatial and temporal patterns, which facilitates the extraction of useful information from GTS and contributes to a better understanding of spatio-temporal data.

Samenvatting

Door de snelle ontwikkeling van datainwinningtechnieken en de mogelijkheid gegevens te delen komen steeds meer en grotere hoeveelheden ruimtelijk-temporale gegevens ter beschikking. Het extraheren van patronen uit deze gegevens via data mining kan nuttige informatie naar boven brengen die het nemen van beslissingen in allerlei toepassingen kan ondersteunen. Dit onderzoek concentreert zich op de analyse van een bepaalde type ruimtelijk-temporele gegevens, namelijk de geo-gerefereerde tijdseries (GTS). GTS bevatten een serie attribuutwaarden verzameld op vaste tijdstippen voor een vaste locatie. Voorbeelden van GTS zijn de waarnemingen van weerstations, of de opname van satellieten die steeds op dezelfde moment op een locatie terugkeren.

Een belangrijk onderdeel van data mining is clustering. Clusters geven een overzicht van de gegevens op een hoger abstractieniveau, maar laten ook toe om zich op details te focussen wanneer men zich op bepaalde clusters richt. Zo kan clusteranalyse gebruikt worden om patronen te verkrijgen uit de GTS. Echter de temporele, ruimtelijke en attribuut informatie in de clusters maakt het soms moeilijk om het patroon te begrijpen of interpreteren. Hier kan visualisatie uitkomst bieden. Kaarten en diagrammen stimuleren het visuele denken, en kunnen gebruikt worden om de patronen te exploreren.

Ondanks dat in eerdere studies clustering en visualisatie zijn gecombineerd blijven er nog vele uitdagingen over als het gaat om de exploratie van complexe ruimtelijk-temporele patronen in GTS. Het betreft onder andere: (1) de exploratie van verschillen in ruimtelijke en/of temporele patronen veroorzaakt door het zogenaamde Modificeerbare Temporale Unit Probleem (MTUP); (2) de exploratie van gelijktijdige ruimtelijke en temporele patronen; (3) de volledige exploratie van ruimtelijk-temporele patronen in complexe datasets; (4) de exploratie van patronen van een 3D GTS.

Dit proefschrift behandelt ieder van deze uitdagingen. In de thesis worden enkele benaderingen voorgesteld en uitgewerkt die gebaseerd zijn op eenrichtings-clustering, op co-clustering en op tri-clusteringsmethoden die het mogelijk maken om patronen uit de complexe datasets te extraheren en via visualisatie methoden te tonen.

Om de verschillen in ruimtelijke en/of temporele patronen veroorzaakt door de MTUP te kunnen exploreren is er een analytische benadering ontwikkeld die kijkt naar ruimtelijke en temporele synchronisatie voor een specifiek patroon bij verschillende temporele resoluties. In deze analytische benadering worden zelf-organiserende kaarten (SOM), de U-matrix afbeeldingen en SOM-

clustervisualisaties en trendplots gebruikt. Met twintig jaar aan Nederlandse dagelijkse gemiddelde temperaturen als voorbeeld is er onafhankelijk van elkaar gekeken naar de synchronisatie van ruimtelijke en temporele patronen op basis van dagelijkse, wekelijkse en maandelijkse resoluties. Het doel was om clusters van weerstations te identificeren met vergelijkbare temperatuur waarneming over alle jaren en om jaar-clusters te vinden die een vergelijkbare temperatuur hadden voor alle weerstations. Het resultaat liet zien dat zowel de ruimtelijk als de temporele patronen beïnvloed worden door de MTUP. Diverse variaties van weerstation en jaarclusters op de verschillende temporele resoluties werden gevonden.

Om tegelijkertijd ruimtelijke en temporele patronen de exploreren is het Bregman blok gemiddelde co-clustering algoritme met I-divergentie (BBAC_I) geïntroduceerd. De BBAC_I is met name gebruikt om groepen van gegevens (co-clusters) te identificeren die vergelijkbare waarden over zowel de ruimtelijke als temporele dimensie bevatten. Om dit inzichtelijk te maken zijn deze gevisualiseerd in heatmaps, in kleine kaart series en in ringkaarten. Opnieuw werden de dagelijkse gemiddelde temperaturen van 28 Nederlandse weerstations gemeten tussen 1992 en 2011 gebruikt in de analyse, op dagelijkse, wekelijkse en maandelijks resoluties. De resultaten toonde dat zowel regio's als subsets van jaar/week/dag vergelijkbare patronen hebben. Ook liet het een afnemend temperatuur patroon zien van zuidwest naar noordoost Nederland en van koude naar hete jaar/maand/dagen. Ook hier werd het effect duidelijk van de gebruikte temporel resolutie, hoe kleine de resolutie des te complexer de patronen.

Om nog beter tegelijkertijd de exploratie van ruimtelijke en temporele patronen ter hand te nemen is een analytische methoden ontwikkeld die BBAC_I combineert met de k-means clustering algoritme. De co-clusters die werden gevonden met BBAC_I zijn verder verfijnd via k-means door de clusters met vergelijkbare waarden de groeperen. Ook hier zijn voor de visualisatie van de verfijnde co-clusters heatmaps, en kleine kaart series gebruikt, aangevulde met kaartseries gecombineerd met tijdlijnen. Als voorbeeldstudie is een Europese temperatuurdataset tussen 1950 en 201 met een ruimtelijk resolutie van 0.25 graden gebruikt. Dit in combinatie met de fenologisch temperatuur gestuurd model, met als doel om de start van het fenologisch voorjaar te exploreren. De resultaten laten vier ruimtelijke fenologsiche patronen met hun temporele dynamiek zien, waarbij opvalt dat de eerste jaren een verlaat voorjaar, en de recente jaren een vroeg voorjaar laten zien. Ook kwamen twaalf temporele fenologische patronen naar voren, ieder met hun ruimtelijke dynamiek. Zo blijkt westelijk Turkije de meest variabele voorjaren te hebben. Dit alles toont aan de

geïntroduceerde aanpak de volledige exploratie van zowel ruimtelijk met een temporele dynamiek als temporele patronen met een ruimtelijk dynamiek mogelijk maakt.

Om patronen van 3D GTS te exploreren is het Bregman gemiddelde cuboid tri-clustering algoritme met I-divergentie (BCAT_I) ontwikkeld als extensie op de BBAC_I. Dit maakt het mogelijk de GTS te groeperen langs drie dimensies en tri-cluster te identificeren. Net als bij de BBAC_I is k-means gebruikt om de clusters te verfijnen. Opnieuw is de dataset van de 28 Nederlandse weerstations gebruikt. De tricluster analyse toonde verfijnde triclusters met vergelijkbare temperatuur voor de ruimtelijke dimensie (de weerstations) en voor twee geneste temporale dimensies (jaren en dagen). Voor de visualisatie zijn zowel 3D als gewone heatmaps, en kleine kaart series met tijdlijnen gebruikt om de variabiliteit in temperatuur te laten zien. Het resultaat liet zes unieke ruimtelijk patronen zien van intra-jaarlijkse temperatuur variabiliteit gekoppeld aan vier jaar groepen. Een nadere analyse van deze patronen onthulde een intense variabiltiet in voorjaars en wintertemperaturen na 1996 in het noordoosten en centrum van het land. Een dergelijke variabiliteit in voorjaartemperaturen was ook zichtbaar voor het zuidoosten van het land. Voor de zomer liet de methode zien dat er sprake is van homogene temperaturen in de zomer voor het gehele land.

Concluderend kan gesteld worden dat dit proefschrift een combinatie van clustering methoden en visualisatie technieken voor de exploratie van GTS heeft gepresenteerd. Afhankelijk van de patronen die men wil extraheren kan men gebruikmaken van eenrichtings clustering tot de hier special ontwikkelde tri-clustering. In samenhang met de verschillende visualisatie opties bieden de clusteringsbenaderingen een uitstekende manier om ruimtelijke en temporale patronen in GTS apart of in combinatie te exploreren met als doel de ruimtelijk-temporele gegevens beter te begrijpen.

ITC Dissertation List

http://www.itc.nl/research/phd/phd_graduates.aspx