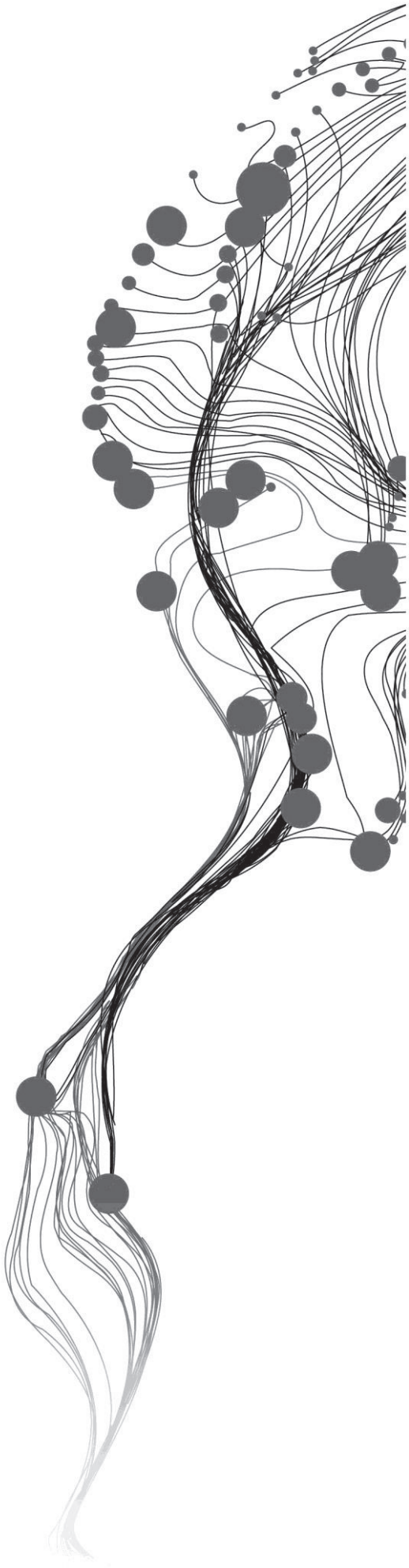
Mining Spatio-temporal Datasets Collected by Volunteers: A Case Study Based on US Lilac Data

SANA BATARSEH
February, 2014

SUPERVISORS:
Dr. R. Zurita-Milla
Prof. Dr. M.J. Kraak



Mining Spatio-temporal Datasets Collected by Volunteers: A Case Study Based on US Lilac Data

SANA BATARSEH
Enschede, The Netherlands, February, 2014

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-informatics

SUPERVISORS:
Dr. R. Zurita-Milla
Prof. Dr. M.J. Kraak

THESIS ASSESSMENT BOARD:
Dr. A.A. Voinov (Chair)
Dr. Ir. A.J.H. van Vliet (External Examiner, Wageningen UR, The Netherlands)
Dr. R. Zurita-Milla (Member)
Prof. Dr. M.J. Kraak (Member)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

In recent years, the rapid advance of Web 2.0 technologies and the integration of location sensing technologies to mobile devices made Volunteered Geographic Information (VGI) a flourishing phenomenon. With VGI, spatio-temporal datasets are collected by large number of volunteers without following a prescribed "scientific experiment" approach and sometimes without any previous planning. This characteristic added challenges to the process of discovering knowledge from such spatio-temporal datasets. In this context, this research is motivated toward developing a workflow for discovering knowledge from spatio-temporal datasets collected by volunteers. In order to achieve the research's main aim, we addressed a relatively massive spatio-temporal dataset in the field of phenology that is the historical Lilac dataset. Lilac dataset contains observations about the location and date of Lilac first leaf and first flower in the US. The observations were made by large number of volunteers during almost 50 years of monitoring. Lilac first flower event is used as an indicator to monitor the variability in the onset of spring. Nevertheless, Lilac dataset represents a typical VGI dataset, which was collected by volunteers without robust planning. It suffers from unequal spatial distribution and many missing values. A comprehensive workflow, which includes a set of interactive exploratory and computational approaches, was developed to find trends in the onset of spring from first flower events. The workflow started by a data understanding stage, in which investigating the data linage, exploring the data content and verifying the data quality were performed. Next, a data preparation stage was implemented to solve the problem of missing values within the dataset. At this stage the latest version of the Spring Index model (SI-x) was used to simulate first flower missing values. The data preparation stage was followed by a data mining stage, in which Self Organizing Maps (SOM) was applied for finding trends in first flower events and clustering them. SOM could accommodate to the complexity of space and time dimensions in Lilac dataset and it could explicitly address the task assigned to it. The final stage of the workflow included presenting the results of the knowledge discovery process. The results were presented in data space, geographic space and time space. Eventually, the workflow developed for discovering knowledge from Lilac dataset is adaptable to other spatio-temporal datasets collected by volunteers. The main stages of the workflow and many of the applied techniques can be generalized for similar cases.

Key words: Spatio-temporal data mining, knowledge discovery process, SOM, VGI, Phenology, Lilac dataset, Spring Index Model.

ACKNOWLEDGEMENTS

First, I want to express my gratitude to my Supervisors Dr. Raul Zurita-Milla and Prof. Dr. Menno-Jan Kraak for their support, guidance and advice throughout the research. I have learned a lot from you. Thank you.

I want to thank the Netherlands Fellowship Program (NFP) for sponsoring me to complete this MSc programme. The chance they offered me is highly appreciated.

I want to thank all my friends at ITC, who surrounded me with their kindness during my stay in Enschede. Siddhi and Gustavo you made me feel like being at home, away from home.

I want to thank my father Tumeh, my mother Amal, my sisters Nora and Mai, my brothers Sanad and Khaled and my special aunts Ghada and Falak for always being there for me despite the distance. I'm very grateful to every one of you.

Finally, I want to thank my husband Markus, who made this possible. This won't be done without you. Thank you for all the support and the encouragement you gave me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background and Problem Description

Nowadays we are experiencing a rapid advance in geographic data acquisition technologies, which could be explained by the progress of sensors, positioning satellites and tracking devices. This growth has led to enormous geographic data volumes that encompass both spatial and temporal dimensions: everything happens somewhere and occurs at some point in time, so-called spatio-temporal data. This emerging type of data brought new challenges to data analysis domain, especially as spatio-temporal data requires the application of proper analytical techniques for discovering the implicit knowledge contained in them.

In fact, spatio-temporal datasets are considered to be complex when it comes to the application of analytical techniques. These datasets are distinguished by the existence of spatial and temporal dependencies. For instance, spatial dependency can be expressed by Tobler's "first law of geography" that "everything is related to everything else but near things are more related than distant things" (Chou, 1995). This law implies that characteristics at different locations tend to be related. Similar types of temporal dependency could exist in relationships with time as well. The existence of these dependencies in spatio-temporal datasets impose certain constrains on the analytical techniques used to analyse them (Andrienko et al., 2010).

Another issue that might add complexity to spatio-temporal datasets is the process of data collection. In recent times, this process is not exclusive to professionals and researchers but non-professionals also participate in data collection. Volunteered Geographic Information (VGI) is a term used for the participation of typically non-professional individuals in creating geographic data (Goodchild, 2007). With this paradigm, geographic data is collected by large numbers of volunteers. Usually, this is done without following a prescribed "scientific experiment" approach or robust planning. Such a data collection mechanism negatively affects the quality of the collected datasets (Flanagin & Metzger, 2008; Goodchild & Li, 2012; Yanenko & Schlieder, 2012). This situation added more challenges to the process of discovering knowledge from geographic data collected by volunteers.

Nevertheless, VGI has proven to be effective at obtaining timely geographic information in many different fields at almost no cost (Goodchild & Li, 2012). One of the successful applications of VGI is phenology. Phenology is defined as the study of periodically occurring events in plant and animal life, influenced by the characteristics of the environment (Lechowicz, 2002). Many studies have confirmed the role of phenology as an important indicator to climatic variability and change (Cleland et al., 2007; Schwartz, 1994). There are many environmental monitoring networks all over the world that focus on collecting phenological ground observations by researchers, students and volunteers (Catlin-Groves, 2012). These networks depend on volunteers to monitor the time and location of phenological events, such as: leaf out, flowering, migrations, and egg laying. Consequently, the results of the mentioned observations are spatio-temporal datasets. These datasets should be analysed properly to obtain reliable conclusions about climate change impacts on plants and animals.

During the last decade, our ability to analyse and understand spatio-temporal datasets has not kept pace with our ability to collect and store them (Compieta et al., 2007; Mennis & Guo, 2009; Santucci & Hauser, 2010). Therefore, spatio-temporal data mining has emerged as a research priority. Spatio-temporal data mining is about the application of analytical computational techniques for discovering patterns, trends, clusters and many kinds of findings in massive spatio-temporal datasets (Compieta et al., 2007).

Spatio-temporal data mining is done in a knowledge discovery process that involves a set of interactive exploratory and computational approaches.

In summary, we are in front of two facts. First, at many different fields, there is an ultimate need to analyse spatio-temporal datasets to be able to understand the underlying phenomena. Second, VGI data collection mechanism adds more challenges to the process of discovering knowledge from spatio-temporal datasets collected by volunteers. In this context, this research is motivated toward developing a workflow for discovering knowledge from spatio-temporal datasets collected by volunteers. The following sections describe the case study, objectives and questions characterizing the presented research.

1.2. Case Study

As it was mentioned in the previous section, this research is motivated toward developing a workflow for discovering knowledge from spatio-temporal datasets collected by volunteers. In order to achieve the research's main aim, we addressed a relatively massive spatio-temporal dataset in the field of phenology. The addressed dataset contains observations about the location and date of first leaf and first flower of two types of Lilac: Syringa Chinensis and Syringa Vulgaris, Figure 1. Lilac dataset was collected by volunteers during almost 50 years of monitoring, in between 1956 and 2003. It has 15,072 observations monitored by 1,126 stations. The stations are mainly distributed over the US, only 9 stations are located in Canada. The historical Lilac dataset was downloaded from NASA's Global Change Master Directory (GCMD) website (Schwartz & Caprio, 2003).

Figure 1: Lilac flower.

The occurrence of first flower event of early spring species, especially Lilac is used as a spring indicator to monitor the variability in the onset of spring (Burton et al., 2013; Schwartz, 1990). Proper analysis for the Lilac dataset can reveal interesting information about trends in the onset of North American spring for the last fifty years. Nevertheless, Lilac dataset represent a typical VGI dataset, which was collected by volunteers without robust planning. It suffers from unequal spatial distribution and lots of missing values. In fact, this is the case of many real life VGI datasets. Therefore, the main steps adopted in this research could be used as guidance for analysing similar spatio-temporal datasets.

1.3. Research Objectives

The main research objective is to develop and implement a workflow for discovering knowledge from spatio-temporal datasets collected by volunteers, through addressing a real life VGI dataset in the field of phenology. This objective can be achieved by fulfilling the following sub-objectives, which represent the main stages of the workflow:

- To explore the dataset for identifying its characteristics and the quality issues that emerged from the data collection process.
- To handle the quality issues of the dataset especially the missing values issue in an early stage of the knowledge discovery process to ensure the quality of the results.
- To apply a data mining technique suited for deriving patterns and trends from spatio-temporal datasets. The analytical technique should be able to consider the spatial and temporal dimensions of the dataset.
- To present the results of the computational data mining technique in a way that can convey the discovered information.

1.4. Research Questions

In order to achieve the above objectives, the research is motivated to answer the following questions:

- What are the main stages of the knowledge discovery process that addresses spatio-temporal datasets collected by volunteers?
- Which methods and techniques are best suited to explore spatio-temporal datasets?
- Which characteristics can be revealed during the exploration phase of a spatio-temporal dataset?
- What approaches could be used to handle the quality issues within VGI datasets without losing part of the information?
- What data mining techniques could be used to identify spatial patterns and trends in massive spatio-temporal datasets?
- How to visually present the results of spatio-temporal data mining technique?

1.5. Innovation

The innovation in this research comes from: developing a workflow with all the steps required for discovering knowledge from spatio-temporal datasets collected by volunteers. Spatio-temporal data mining has emerged as a research priority, as there is an ultimate need for mining spatio-temporal datasets in many different fields (Camossi et al., 2008; Compieta et al., 2007). Providing better understanding about the implications of uneven volunteers participation in VGI data collection process is one of the main challenges that faces geographic information science nowadays (Goodchild, 2009).

1.6. Thesis Structure

This thesis contains six chapters. The first chapter provides an introduction to the background of the problem, the case study, the research objectives and questions that are aimed to be solved by the end of the research. The main concepts and methods used in this research are presented in the second chapter, which is titled literature review. The third chapter describes the data used in the research. It also describes the Spring Index model that has been used to simulate the missing values within the Lilac dataset. The fourth chapter describes the developed workflow. It also discusses the functionality of the applied methods and techniques used for discovering knowledge from Lilac dataset. The fifth chapter presents the results of the applied workflow. In addition it presents discussions about the produced results. The final chapter presents the conclusions of the research along with recommendation for future work.

2. LITERATURE REVIEW

This chapter reviews the main concepts and methods used in this research project. The first section provides an overview of Volunteered Geographic Information (VGI) characteristics, in addition to the impact of the VGI data collection process on the quality of the collected data. The next section introduces data mining and the knowledge discovery process. The role of visualization in the knowledge discovery process is discussed in the third section. The last section describes Self Organizing Maps (SOM) as one of the available spatio-temporal data mining techniques.

2.1. VGI characteristics

The idea of people gathering information about their surrounding environment that encompasses spatial component is not new. Brown (2011) at his article in The Guardian mentioned one of the oldest attempts in this domain. It was 1736 when Robert Mashram started monitoring the arrival of first swallows in his village as a spring indicator. After doing this for 50 years and including other 26 spring indicators, he delivered his records to the British Royal Society. Many other country gentlemen started following his example. Yet in recent years, it became tremendously popular for people to collect geographic information and share it through the web, especially with the rapid advance of Web 2.0 technologies and integrating location sensing technologies to mobile devices. This made VGI a phenomenon and allowed it to be recognised as a new source of geographic information (Goodchild, 2007, 2009). VGI is even called as "Neogeography", which means new geography (Turner, 2006).

In fact, VGI made a spectacular change in the contents, characteristics and ways of creating, publishing and using geographic information (Elwood, Goodchild, & Sui, 2012). Elwood et al. (2012) described that with VGI, the creation of geographic information is not limited to mapping agencies and professional cartographers. This is true, since non-professional individuals can create and publish digital geographic information on-line without cost. The authors clarified that individual community members can create geographic information about their surroundings more effectively than cartographic experts from distant mapping agencies. For instance, in the occurrence of natural disasters, volunteers proved to acquire timely and detailed geographic information better than any other source (Longueville, Smith, & Luraschi, 2009). Longueville et al. (2009) have demonstrated VGI role in supporting emergency planning, risk assessment and damage assessment activities. Specifically, through analysing publicly available Twitter messages published during a forest fire occurred near the French city Marseille in July 2009.

In spite of all the VGI benefits, quality has been reported as a main and pending issue (Flanagin & Metzger, 2008; Goodchild & Li, 2012; Yanenko & Schlieder, 2012). Generally, the quality of geographic information has the following major standards: positional, attribute and temporal accuracy, logical consistency, completeness, and lineage. These standards are used by the geographic information community to assess the quality of geographic datasets (Goodchild & Li, 2012). Goodchild and Li (2012) described the quality of VGI datasets as "highly variable and undocumented, it fails to follow scientific principles of sampling design, and its coverage is incomplete". This description summarizes how VGI datasets violates the above mentioned quality standards.

Most of the time, the VGI data collection process does not follow the "scientific experiment" approach, which influences the quality of the acquired datasets. In ideal scientific research, everything is planned before starting with data collection: objectives are stated, the measurements to be done and the techniques to perform them are known, the statistical models to be used for analysis are designed and

even the possible patterns of results are contemplated (Szczepańska, 2011). Brunsdon and Comber (2012) have clarified that VGI data collection mechanism is different from this. Even with trained volunteers it's usually hard to control the spatial distribution of the observations. This depends on the locations of volunteers acquiring the information. Therefore, VGI datasets usually suffer from lacking of sampling design. The authors advised to take this issue into consideration when performing data analysis, calibrating statistical models or testing hypotheses.

In researching the data collection approaches adopted in environmental monitoring programs that depend on volunteers, Wintle, Runge, and Bekessy (2010) classified monitoring programs as targeted monitoring or surveillance monitoring. The authors clarified that targeted monitoring is based on clearly explained objectives and it aims to process among prior defined hypotheses. In addition, these programs provide the volunteers with certain protocols and procedures to be followed during data collection. On the other hand, surveillance monitoring programs lack previously defined objectives, in which a series of observations are produced so they can be used later to generate post hoc hypothesis. The authors argued that surveillance monitoring programs can have value as well. As these programs can help in revealing "unknown unknowns" and assist in generating new unexpected hypotheses.

Many researchers and scientists support targeted monitoring approach. Since it guarantees the production of datasets, which are suitable for scientific research and effective decision making (Wiersma, 2010). In addition, efforts time and money could be saved through prior robust planning (Francis, Blancher, & Phoenix, 2009). However, Wiersma (2010) declared that although targeted monitoring has lots of supporters, surveillance monitoring programs are not going to disappear and most probably are going to increase. Especially because of the rapid advancement in Web 2.0 technologies that made it possible to generate large quantities of VGI datasets in a relatively spontaneous manner. The author added that data mining techniques for analysing such kind of datasets are needed.

2.2. Data mining

As it was discussed at the previous section, there are huge volumes of geographic datasets that were collected by volunteers, most of the time without robust planning. These datasets are characterized by: massiveness (have large number of observations), high dimensionality (include many variables) and complexity (they are dynamic in space and time, have explicit and implicit relations and lack to sampling design). In this context, Mennis and Guo (2009) mentioned that traditional spatial analysis methods such as cartography and spatial statistics could not process the increasing volumes of geographic datasets efficiently neither datasets created by volunteers nor the other geographic data acquisition technologies. That is because traditional spatial analysis methods were developed when geographic datasets were not as huge and complex as today. Additionally, the available computational capabilities at that time were not as powerful as today. To address this challenge spatial and spatio-temporal data mining has emerged as an alternative research field (Compieta et al., 2007; Mennis & Guo, 2009; Santucci & Hauser, 2010).

Data mining is about the application of methods and techniques for extracting knowledgeable information from large and complex datasets (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Fayyad et al. (1996) clarified that data mining techniques have evolved from intersecting research fields such as: statistics, pattern recognition, machine learning, databases, artificial intelligence, data visualization and high-performance computing. Statistics in particular is one of these fundamental fields, because it was traditionally used to obtain patterns from datasets and it has a defined framework for inferring samples results to an overall population. The author added that it was around the 1960s when statisticians started adapt their methods for data analysis to computer-based techniques. This made it possible to handle and analyse huge datasets more efficiently. However, data mining has emerged as an independent research area just in the 1990s (Vesanto, 2002).

Jiawei and Harvey (2009) explained that data mining is able to carry out several tasks such as classification, clustering, association and finding patterns and trends. Decision trees, artificial neural networks, association rules and k-means clustering are examples of the available data mining techniques. Data mining is done through a multi-step process, which is named as knowledge discovery in database (KDD) (Fayyad et al., 1996; Jiawei & Harvey, 2009; Mennis & Guo, 2009). The process generally includes three main stages, which are: data pre-processing, data mining and post processing. Data pre-processing is about data exploration, cleaning, selection and any other activity to prepare the dataset for mining algorithms. Data mining is the step in which mining methods and techniques are applied to extract the implicit knowledge within the dataset. Post processing is about evaluating the results and presenting them in an intelligible manner. Overall, data mining and KDD aims to provide methods and techniques to automate the entire process of data analysis to the most possible degree (Jiawei & Harvey, 2009).

Data mining with respect to geographic data involves the application of data mining computational techniques to find out: patterns and trends, reveal hidden relationships, generate clusters of homogenous objects and generally discover implicit relationships in spatial and spatio-temporal datasets (Compieta et al., 2007). Spatio-temporal data mining in particular is about the application of data mining techniques that consider the implicit relations of the spatial and temporal dimensions, in addition to the non-spatial attributes describing the dataset. Similar to traditional data mining, spatio-temporal data mining is a part knowledge discovery process that called Geographic Knowledge Discovery process (GKD) (Jiawei & Harvey, 2009).

2.3. Visualization and knowledge discovery

Visualization is known to be powerful in acquiring information from massive data. Employing visualization in the knowledge discovery process can take advantage of human abilities to perceive patterns, establish links and make conclusions. Keim et al. (2010) have discussed that Knowledge can be acquired from data through applying visual data exploration approach or automated data analysis approach. The authors argued that in both approaches, visualization plays a significant role, as shown in Figure 2.

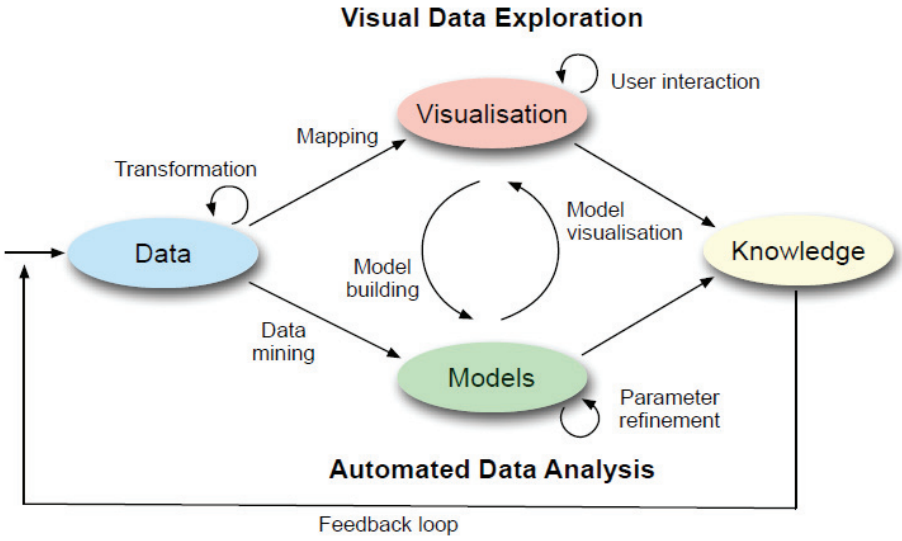


Figure 2: Visualization role in knowledge discovery process. Source: Keim et al., 2010, p.10.

Visual data exploration has a long tradition of discovering knowledge from massive datasets. For instance, Exploratory Data Analysis (EDA) is an analytical approach for exploring data through combining statistics and visual graphic representations. The philosophy behind EDA is to let the datasets speak for themselves and impose the structure among them (Gorunescu, 2011). EDA was prompted by John Tukey in 1977 and since then it has been used to explore and discover relations between variables in the knowledge discovery process (Croarkin, 2003). Histograms, bar charts, boxplots, and scatterplots are examples of EDA visual statistical representations.

Geographic information scientists have their own visual exploration techniques. These include geographic maps, map animations, space time cube and time wave. Andrienko et al. (2003) have demonstrated the role of map animation and space time cube in exploring spatio-temporal datasets. These techniques consider both the spatial and temporal dimensions while visually exploring the dataset. Additionally, they allow an interactive visualization through dynamically changing their appearance upon the analyst's request. Kraak and Li (2012) introduced the time wave for visualizing trends from a temporal angle. They demonstrated how time wave could be used to expose trends that were not obvious using other visual representations. However, for effective exploration, the interactive combination between the different representations is required. This is because each representation conveys only part of the overall information (Andrienko et al., 2010; Compieta et al., 2007).

Automated data analysis is done through applying data mining techniques. This usually involves generating models for the original data. These models need to be refined, calibrated and evaluated. Visualization can allow the analyst to interact with the created models through modifying the model's parameters and evaluating the model's findings. Visualisation can also be helpful in discovering wrong results at an early stage. Santucci and Hauser (2010) have recommended the integration of visualization into data mining techniques. The authors stated that the computational data mining techniques usually act as a black-box, since they do not take into account the analyst's knowledge. Therefore, visual interaction can bring back the analyst knowledge and allow for more effective data mining. Self Organizing Maps (SOM) is one of the data mining techniques that allow visual interaction between the analyst and the built data models (Andrienko et al., 2010).

2.4. Self Organizing Maps (SOM)

Clustering is widely used for discovering knowledge from geographic data. It is about grouping data objects into groups based on some similarity. The goal is to emphasize on similarity within a group of objects and dissimilarity between different groups, a group thus built is called a cluster. Clustering allows to identify interesting structures in the underlying geographic dataset (Camossi et al., 2008). Clustering techniques can be generally classified into two main categories: partitioning clustering and hierarchical clustering (Jiawei & Harvey, 2009). Partitioning clustering divides data objects into a number of non-overlapping clusters, in which a data object is designated to a certain cluster based on proximity or dissimilarity measure. While hierarchical clustering organizes data objects into a hierarchy with a concatenation of nested clusters. In other words, hierarchal clustering groups data objects into clusters and sub-clusters.

Self organizing maps (SOM) is a partitioning clustering technique (Jiawei & Harvey, 2009). This technique makes it possible to find clusters in a dataset through mapping high dimensional data on a regular low dimensional grid, usually two-dimensional. The SOM algorithm is able to transform complex and nonlinear relationships within a dataset into simple geometric relationships on the low dimensional grid. Through reducing data dimensionality, SOM enables visualizing high dimensional data, since humans cannot visualize such data as it is. Accordingly, the SOM algorithm can: perform clustering, reduce the dimensionality of complex data and provide a mean for visualizing high dimensional data.

SOM was developed by Tuevo Kohonen around 1981-82 as a new nonlinear projecting mapping technique (Kohonen, 2013). Kohonen (2013) mentioned that since then and up until now there are more than ten thousand publications including ten books that discuss SOM algorithm and its various applications. However, the briefing presented below for SOM basic algorithm is based on Zurita-Milla et al. (2013), Vesanto (2002) and Vesanto et al. (2000).

SOM algorithm starts with a regular grid of neurons presented on a map. The neurons could be arranged in a rectangular or hexagonal lattice, as shown in Figure 3. The global shape of the map can be represented as a sheet, cylinder or toroid. If two sides of the map are connected, then the shape of the map is cylinder and if all the sides of the map are connected, then the shape is toroid, this is shown in Figure 4. Furthermore, the neurons are linked to each other by a neighbourhood relation. The function of this relation determines how strongly the neurons are linked. However, number of neurons, dimensions of the grid, grid shape and the global shape of the map are defined by the analyst.

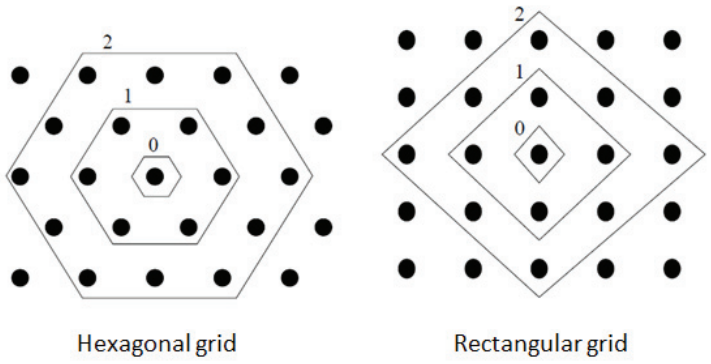


Figure 3: Different grid shapes. Source: Vesanto (2002), p.12.

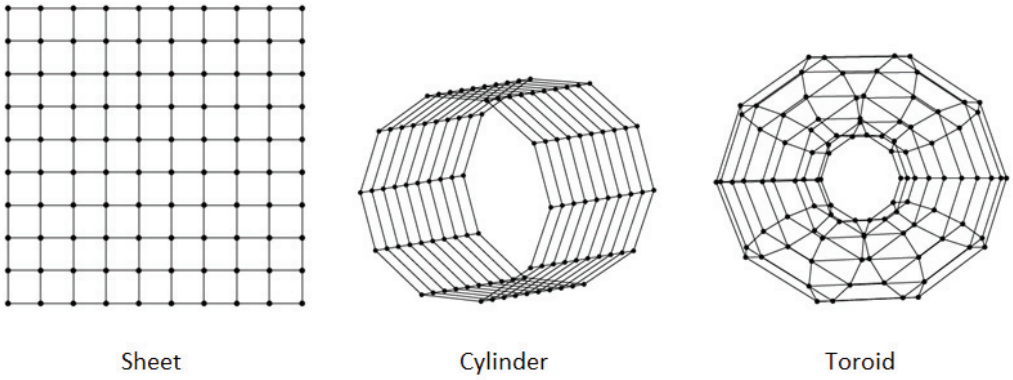


Figure 4: Different global map shapes. Source: Vesanto et al. (2000), p.8.

The input dataset is usually a numerical table. In this table, the attributes represent the variables describing the dataset and the records represent the data objects. The input dataset structure is shown in Figure 5. The SOM algorithm aims to project such a dataset into the previously described grid of neurons. Accordingly, each neuron in the map is assigned to a d-dimensional weight vector, where d is equal to the number of variables at the input dataset. The initial weights of the vectors represent the values of data objects. Such vectors are called prototype vectors.

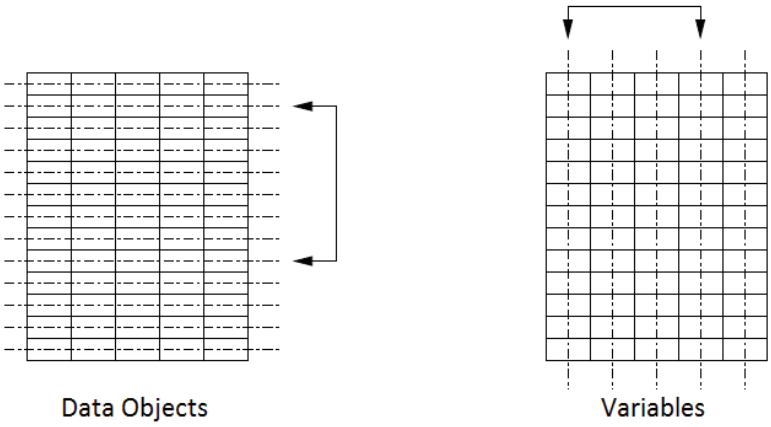


Figure 5: SOM input table structure. Source: Vesanto (2002), p.9.

The SOM algorithm trains the neurons in an iterative process. During the training process, one prototype vector is chosen randomly in each training step, this vector is named as the sample vector. The differences between the sample vector and all the other weight vectors are calculated using Euclidian Distance. The neuron that includes the vector, which shares the minimum distance with the sample vector, is defined to be the Best Matching Unit (BMU). With each training step, the defined BMU moves closer to its associated sample vector. Moreover, the neurons that share a neighbourhood relation with the BMU are moved as well. This process stretches the BMU and its topological neighbours toward the sample vector as shown in Figure 6. A special property of SOM algorithm is that the area of the neighbourhood gets smaller over time. This could be done by reducing the radius of the neighbourhood during the training process. The training process keeps iterating till it reaches a certain number of iterations, which should be defined by the analyst before stating the training. This process produces a grid, in which neurons that are near each other are more similar than neurons that are further apart.

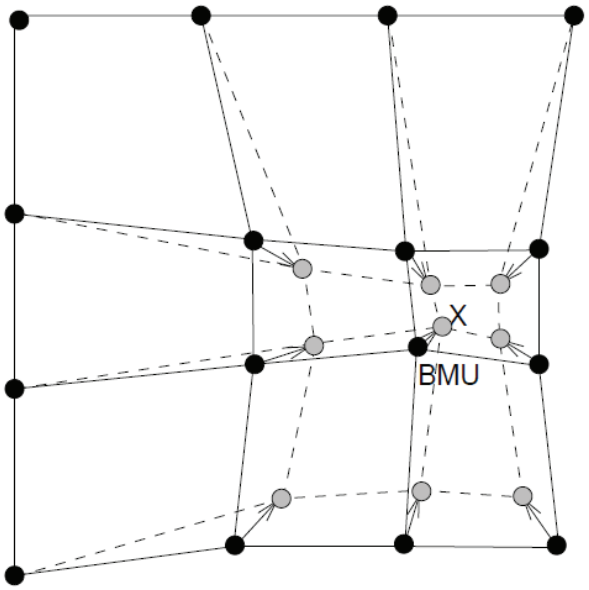


Figure 6: Updating the BMU location and its neighbours toward the sample vector marked with x. Source: Vesanto (2002), p.12.

Regarding the parameters that are defined by the analyst, Kohonen (2013) has stated some guidelines for defining these parameters, which are as the following:
Grid shape: As it was discussed there are usually two common grid shapes, rectangular and hexagonal. The hexagonal grids are recommended, since they are more illustrative in visual terms and give better accuracy.
Dimensions of grid: It is not possible to guess the exact dimensions of the grid before starting SOM algorithm. Trial and error can lead to the most suitable dimensions of gird. Anyways, it is recommended that the dimensions should be roughly corresponding to the two largest principal components in the input data. Also, the oblong shape proved to have faster performance than the square one at the learning process.
Global shape of the map: In the sheet shape, the neighbourhood function at the edges is not regular as in the middle due to discontinuity. Cylinder and toroid shapes are better choices when the input data has a cyclic structure, since these shapes don't suffer from discontinuity at the edges.

The unified distance matrix (U-matrix) is a common way to identifying clusters in SOM results (Andrienko et al., 2010; Vesanto, 2002). The U-matrix allows the visualization of distances between each neuron and its neighbours by representing distances with different colours. Vesanto (2002) clarified that cluster borders can be defined in U-matrix as high distance separated by low distance. Figure 7 shows an example of U-matrix and the clusters identified.

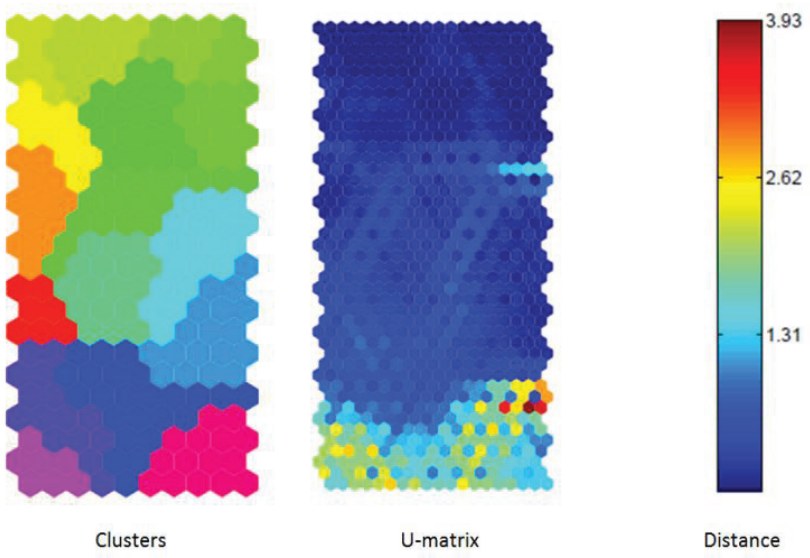


Figure 7: An example of the U-matrix and the clusters identified. Source: Vesanto (2002), p.41.

SOM algorithm has been successfully applied in geographic data analysis. Andrienko et al. (2010) applied SOM to group and arrange spatial distributions and temporal variations based on their similarities. The authors presented how SOM succeeded in analysing spatio-temporal datasets at two levels:

- Analysing the change of the spatial situation over time, which can be called analysing space in time. In this case, the dimensions of the input datasets represent the temporal variation.
- Analysing the distribution of the local temporal variations over different spatial locations, which can be called analysing time in space. In this case, the dimensions of the input datasets represent different spatial locations.

Andrienko et al. (2010) displayed the results of U-matrix on spatial cartographic maps and temporal displays. For displaying the results in geographic space, features in cartographic maps were coloured based on their corresponding positions in the U-matrix. For displaying the results in time space, time graph and time arranger were used. In both temporal displays, segments were coloured based on their corresponding positions in the U-matrix.

Wu, Zurita-Milla, and Kraak (2013) used U-matrix to identify clusters in time series datasets that belong to a number of meteorological stations in the Netherlands. The authors adopted the idea of synchronization in space and time to find out clusters, since synchronization can be identified separately in space and time. Spatial synchronization refers to clusters of stations in geographic space, in which there is a high degree of similarity between their signals along certain time interval. Temporal synchronization clusters years that behaved in a similar way in terms of meteorological conditions. SOM and U-matrix were used in this research to discover spatial and temporal synchronization, as shown in Figure 8.

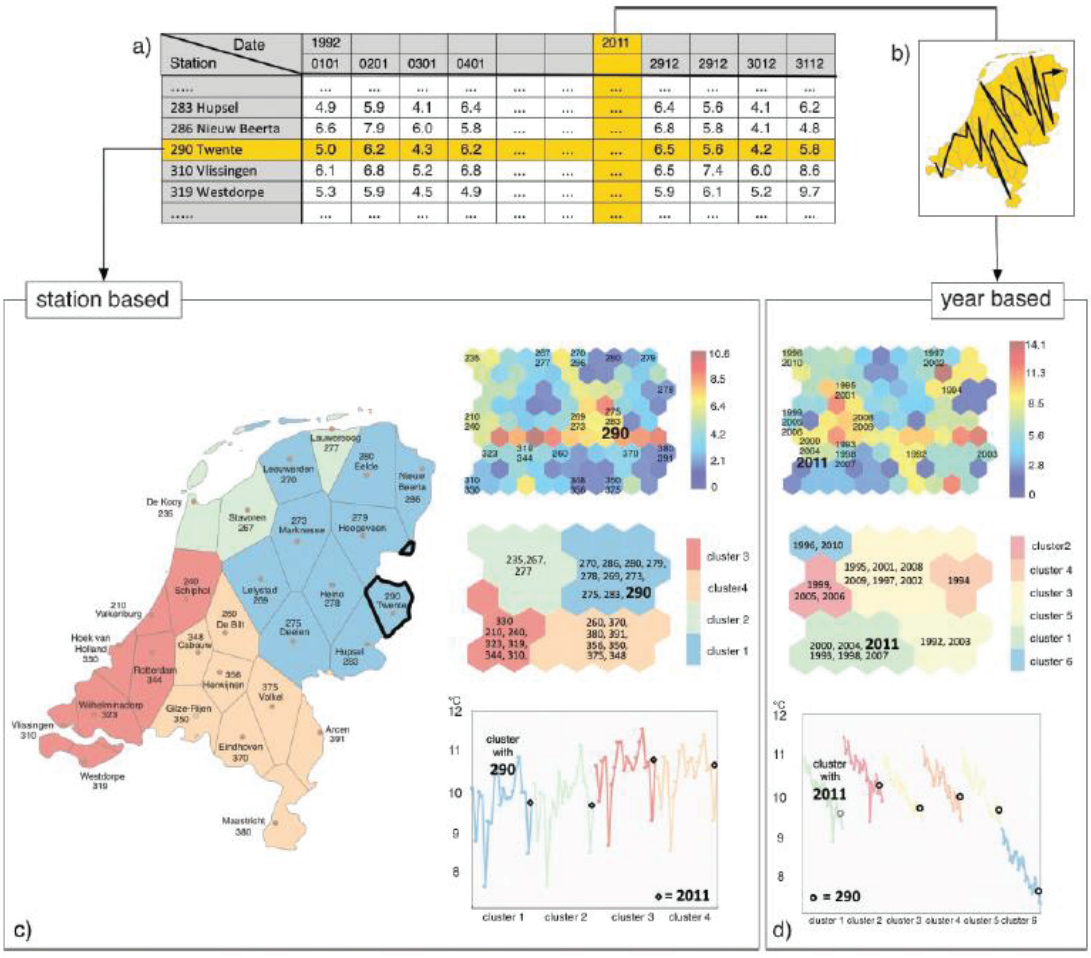


Figure 8: Discovering synchronization in space and time using SOM and U-matrix. Source: Wu et al. (2013).

3. DATA AND MODEL USED

This chapter describes the data and the model used for Lilac case study. In a certain stage at our research, Lilac dataset was found to have a serious issue with missing values. Therefore, we needed to simulate first flower events using the Spring Index model. For this purpose, daily temperatures acquired from the Daymet website were fed to the model used. The first section describes Lilac dataset and its attributes. Daymet website and the services it provides are discussed at the second section. While the Spring Index model applied in our research is discussed at the third section.

3.1. Lilac data

Lilac dataset was downloaded from the NASA's Global Change Master Directory (GCMD) website. The downloaded dataset was structured in two text format tables. The tables and their attributes are shown in Figure 9. The first table contained the first flower and first leaf observations, it had 15,072 records. The plant type attribute gives information about the monitored Lilac type at each station (Syringa Chinensis or Syringa Vulgaris). The dates of first flower and first leaf observations were given as day of the year (DOY), which represent the number of days since the beginning of the year. While the second table contained information about the Lilac monitoring stations, it had 1,126 records. The location information of the stations was given in WGS 84 geographic coordinates. To facilitate handling the data, a join between the two tables, which is based on the unique Station ID attributes, was done.

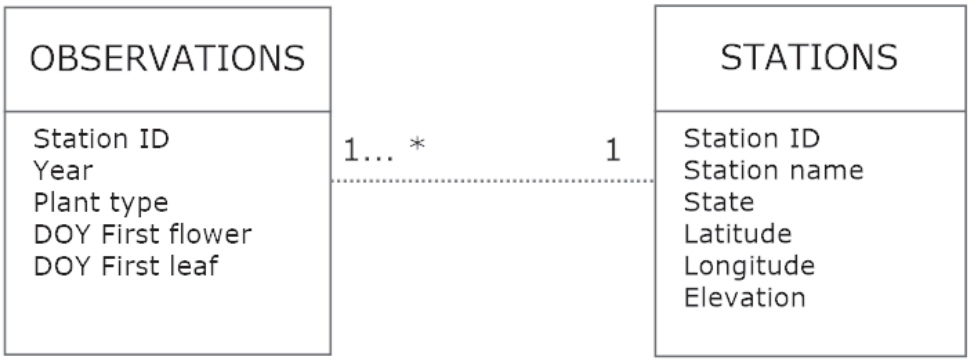


Figure 9: Lilac dataset tables and their attributes.

3.2. Daymet data

The Daymet website provides daily meteorological data presented on a continuous 1km × 1km gridded surface (Daymet, 2012). The surface coverage includes three countries: the United States, Canada and Mexico. The daily gridded surfaces are made for a number of meteorological measurements, which are: minimum and maximum temperature, humidity, precipitation and radiation. These surfaces are available for every day since 1980 to 2012. Daymet website is a part of a research project that is sponsored by NASA. It aims to fulfil the need for continuous surfaces of daily weather data, which are required by plant growth models and many other research fields.

Daymet daily gridded surfaces are interpolated from daily observations that belong to ground meteorological stations. These observations were collected from different sources within the three countries. The ground observations for the United States were obtained from the National Climate Data Centre (NCDC) and the Natural Resources Conservation Service (NRCS). The meteorological observations for Canada were obtained from the Government of Canada and the GHCN-Daily. The Mexican meteorological observations were provided by the Servicio Meteorológico Nacional. Figure 10 shows a map for all the active stations during 2005, which were used for interpolating the 1km x 1km gridded surface.

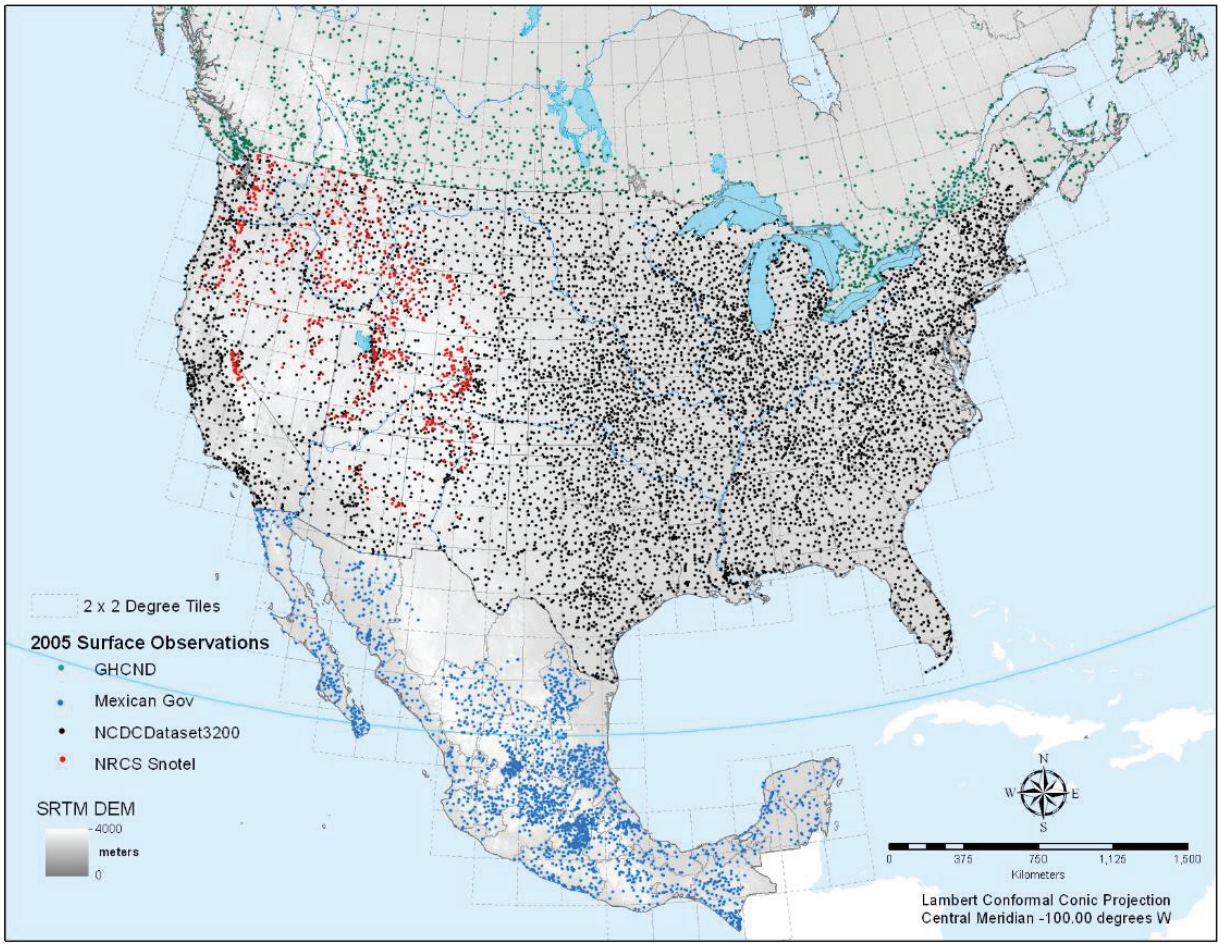


Figure 10: The active stations for 2005 and their distribution within Daymet coverage. Source: (Daymet, 2012).

Daymet applies a system of 2degree × 2degree tiles to be able to process the large number of input data and the large study area. In this system, each tile is processed individually. The interpolation at each prediction point is done using the spatial convolution of a truncated Gaussian filter. In this filter, the search radius of stations depends on the number of stations around the prediction point. The search radius is small in data-rich areas and it gets bigger in data-poor areas. This is achieved through defining an average number of observations to be covered at each prediction point. The final result of this process is a gridded daily data for the coverage area.

Daymet data can be accessed by the user through different ways. One of them is single pixel extraction tool, in which the user can enter a certain location (longitude, latitude) within Daymet spatial coverage to download the required meteorological data for that location. To fulfil this request, data from the nearest 1km × 1km grid are extracted from Daymet database and downloaded in table format. Each column in the table represents a meteorological measurement and each row represents a day. For multiple coordinates extraction, Daymet offers a Java tool that can automatically process a number of coordinates saved in a text file. This tool has one minor restriction: it can just process 30 locations per hour.

3.3. Spring Index model

The Spring Index (SI) model is a regression-based model designed to simulate phenological events from meteorological data. The work with SI model was initiated by M.D. Schwartz around the late 1980s (Schwartz, 1990). The need for such models arose due to the lack of complete period-of-record for plants phenological data in North America (Schwartz, 1994). This situation prevented exploring trends that characterize the transition from winter into spring. Therefore, the idea of developing models, which can simulate phenological events, was proposed and implemented for estimating the missing data.

The SI model passed through many modifications since it was developed and up until now. The history of these modifications was presented by Ault, Zurita-Milla, and Schwartz (2013). The most recent version of Spring Index models is called the extended Spring Index and abbreviated as SI-x (Schwartz, Ault, & Betancourt, 2013). SI-x simulates the first leaf and first flower events for three different plants, which are: Lilac, Arnold Red, and Zabeli. For simulating these events at a certain location, the model requires daily minimum and maximum temperatures along with the latitude of that location. The daily temperatures should be available since the 1st of January to simulate the plants response for that year. The previous versions of SI models needed to accumulate chilling hours for calculating first leaf. With SI-x version, this is not a requirement.

For calibrating and evaluating SI-x model, meteorological data along with plants response data were used. The meteorological data came from observation stations that record standard daily maximum and minimum temperatures across the US (Schwartz et al., 2013). These stations belonged to the National Climatic Data Centre. The plants response data used was the first leaf and first flower for Lilac and Honeysuckle, which was obtained from sites throughout the north-central and north-eastern of the US (Wolfe et al., 2005). The model adopts a multivariate stepwise regression for relating meteorological measurements to plant response. During this process certain parameters are identified for each plant type: Lilac, Arnold Red, and Zabeli.

Ault et al. (2013) have published a Matlab toolbox for simulating spring phenological events. The toolbox calculates SI-x version of Spring Index models. It consists of six core functions that calculate the day of the year for first leafing and first flowering events. The toolbox simulates these events for Lilac, Arnold Red, and Zabeli at any location in the Northern Hemisphere, from maximum and minimum daily temperatures.

The Matlab toolbox was developed from the original Fortran code designed by M.D. Schwartz. The developing process included translating the Fortran routines as identical as possible. By the end the Matlab code was similar in structure and syntax to the original Fortran code. Figure 11 presents the simulation of first leaf and first flower events for different six stations, using the Matlab toolbox and the original Fortran code. These six stations where used to evaluate the results of the Matlab toolbox and how similar they are to the results of the Fortran code.

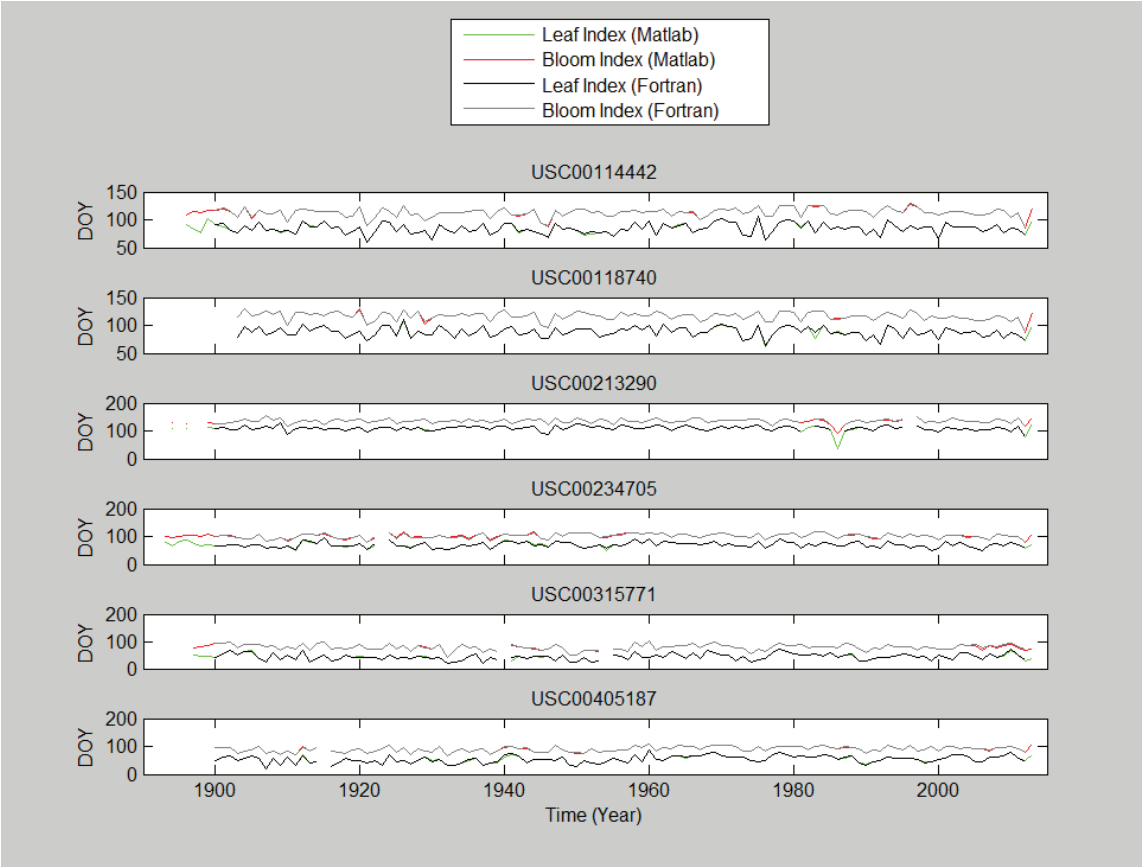


Figure 11: Simulated phenological events for different 6 stations using the Matlab toolbox and the Fortran code.

4. SETUP OF THE WORKFLOW

This chapter starts by describing a comprehensive workflow for discovering knowledge from spatio-temporal datasets collected by volunteers. A further specialized workflow for Lilac case study is presented later. Each section of this chapter discusses the methods applied at a certain stage of the workflow developed for discovering knowledge from Lilac dataset.

The workflow developed for discovering knowledge from spatio-temporal datasets collected by volunteers is based on the knowledge discovery process discussed in section 2.3. It takes into account the special characteristics of spatio-temporal datasets. It also considers the characteristics of VGI and the consequences of VGI data collection process. The workflow contains five main stages, as shown in Figure 12. Below is a description for each stage.

Figure 12: The five main stages of the developed workflow.

Objectives definition

At the beginning of the workflow, the objectives of the whole knowledge discovery process should be defined. The analyst should have a vision about what kind of knowledge is expected to be revealed from the addressed dataset. The analyst should also be familiar with the domain knowledge of the proposed problem. Without having an adequate background and defined objectives, it's hard to make decisions about what to ignore and what to pursue more during the knowledge discovery process (Vesanto, 2002).

Data understanding

Data understanding is the stage that enables the analyst to make sense of the available data, get familiar with it and check how reliable it is. This stage includes understanding the origin, content and quality of the dataset. It is advisable to execute the data understanding early in the knowledge discovery process, since its findings must strongly affect the entire process (Compieta et al., 2007; Vesanto, 2002). Good understanding of the raw data will help in performing proper data preparation and selecting the appropriate data mining techniques. Accordingly, Data understanding involves three tasks, which are:

- *Investigating the lineage of the dataset.* It's important to understand the data's origin and how it was evolved over time. Especially with VGI datasets, as it was discussed in the literature review chapter (section 2.1), the data collection process strongly affects the characteristics and the quality of the collected data. The information acquired with this task will help the analyst to form the first impression about the addressed dataset.
- *Exploring the contents of the dataset.* With this task the analyst can experience and recognize the characteristics of the addressed spatio-temporal dataset, such as: the spatial and temporal distribution of certain variables, the existence of spatial dependency and the relations between variables. This allows the analyst to discover first insights into the data and form initial hypotheses regarding the hidden information.

- *Verifying the quality of the dataset.* This task aims to verify the quality of the raw data, such as: the completeness of the dataset in space and time, the occurrence of missing values and the existence of erroneous values. This task is important because we are addressing VGI datasets, in which quality is a serious issue.

The last two tasks in this stage, exploring the contents and verifying the quality of the dataset are mainly done through applying visual exploration techniques and simple analysis techniques.

Data preparation

Usually VGI datasets have quality problems. This stage is advised to be done for solving these problems before reaching to data mining. To ensure the quality of the knowledge discovery process results, the quality of the raw dataset should be first considered and properly handled (Deng et al., 2011). For instance, if the dataset have problems with erroneous values, then data cleaning is a task that should be done during data preparation. In addition, data preparation stage includes all the activities required to construct the final dataset which will be processed by the data mining techniques. This includes data selection, data transformation and data projection.

Data mining

The two previous stages of the workflow are mainly done to prepare the dataset and the analyst for this stage. Data mining is the stage in which the analyst applies the appropriate data mining technique for achieving the main objectives of the knowledge discovery process. Typically, there are several data mining techniques for the same task. For instance, clusters can be found in a dataset using different data mining techniques. Each technique has its specification, advantages and disadvantages. The selection of the appropriate data mining technique is based on the main objectives of the knowledge discovery process and the characteristics of the dataset. Some techniques have certain requirements regarding the structure of the data. Therefore, going back to the data preparation stage is often necessary.

Results presentation

The aim at this stage is to present the results of the knowledge discovery process in a way that can convey the discovered information. The way of presenting the results should explicitly address the main objectives of the knowledge discovery process. Visualization plays a major role in presenting the results, especially with spatio-temporal datasets. Displaying the results in geographic space and in time space will help in perceiving relations that creates better understanding for the results.

Lilac knowledge discovery workflow

Lilac dataset includes a number of stations spatially distributed over the US. Each station has a signal of first flower events along a certain number of years. Lilac first flower event is used as an indicator to monitor the variability in the onset of spring (Schwartz, 1990; Schwartz et al., 2013). Therefore, clustering stations that have synchronized first flower signals over the same period of time can allow identifying regions that share similar trends in the onset of spring. This has been defined to be the main objective of the Lilac knowledge discovery process.

Lilac dataset has passed through all the discussed stages for revealing information about trends in the onset of spring. Further discussions about the methods and techniques applied for data understanding, data preparation and data mining stages are presented at the following sections. Figure 13 shows an overview for the applied workflow.

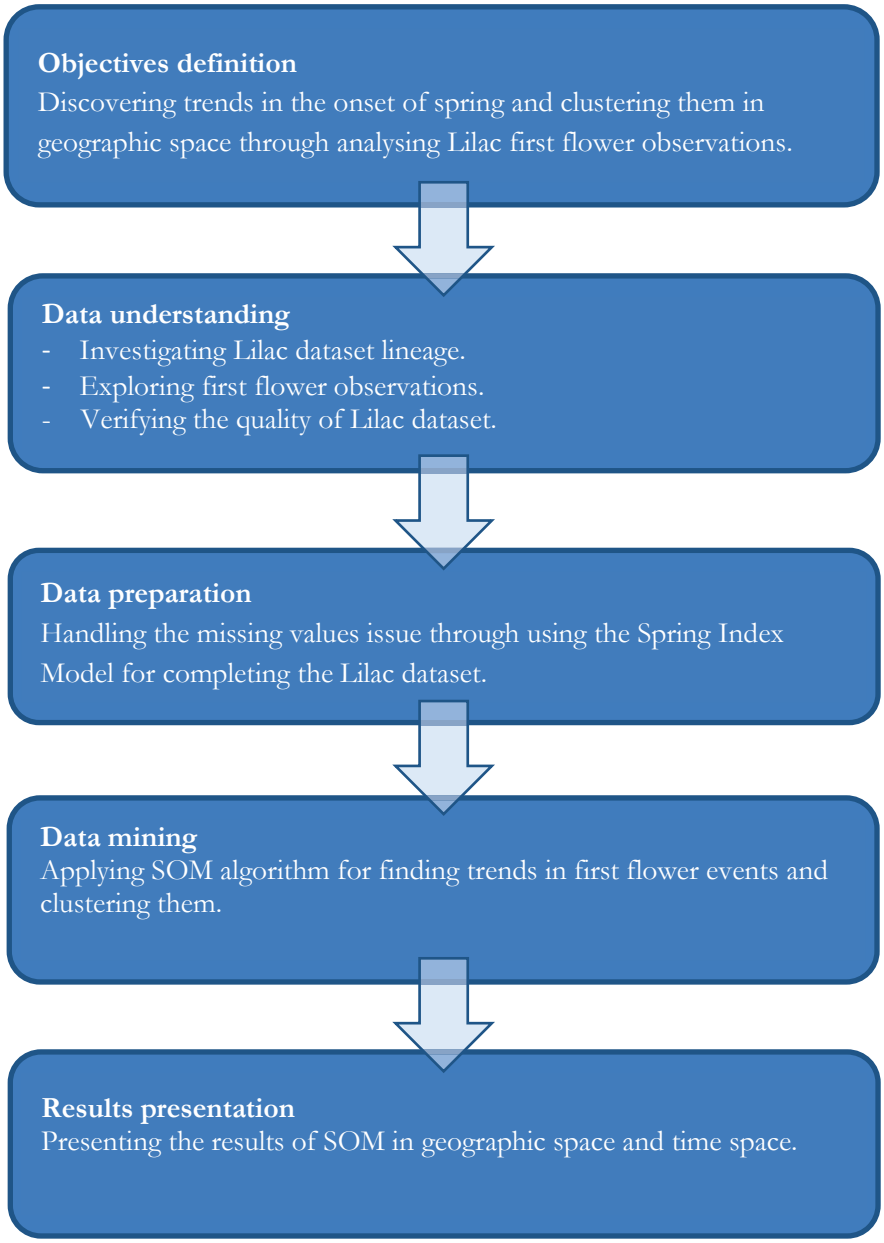


Figure 13: The developed workflow for discovering knowledge from Lilac data.

4.1. Lilac data understanding

Lilac data understanding stage included investigating the lineage of the historical lilac dataset, understanding its content and verifying its quality. A series of methods and techniques were applied to achieve the presented tasks. This section discusses these techniques and the reason behind applying each of them.

Lilac data lineage

We started this stage with investigating the lineage of the dataset, as we didn't have an adequate background about the data history. For achieving this task, many related publications were reviewed to find information describing the data's origin and how it was evolved overtime. The issues that were investigated are: the main purpose for collecting the Lilac dataset, the official programs managed Lilac data collection, the type of volunteers participated in the data collection (e.g. professionals or non-professionals, trained or untrained) and the approaches applied for collecting the dataset.

Visual Exploration

Maps, space time cube, EDA statistical representations were used for visually exploring the dataset. Each technique was used for a reason:

- Mapping the stations in geographic space was used to explore their spatial distribution and coverage along the US. This was done using ArcGIS software.
- Exploring the first flower observations using space time cube was done to investigate their spatial and temporal distributions, in addition to examine their completeness. This was done using R statistical software.
- Creating a set of EDA statistical representations was done to examine the values of first flower attribute and explore their relation with the other variables. This was done using R statistical software.

Moran's I index

The spatial autocorrelation between first flower observations was investigated. Spatial autocorrelation is a statistical representation of Tobler's "first law of geography" that "everything is related to everything else but near things are more related than distant things" (Chou, 1995). The existence of spatial autocorrelation means that nearby observations are more likely to be similar than further ones, which confirms the spatial dependency between the observations. The absence or existence of spatial autocorrelation has a notable influence on selecting the proper analytical techniques. When spatial autocorrelation is found between the observations, then it's improper to use standard statistical techniques for analysis. These techniques assume randomness and independency between the observations (Dale & Fortin, 2009).

To examine the spatial autocorrelation between first flower observations, Moran's I test was applied to the observations of each year separately. Moran's I test measures the global spatial autocorrelation between the observations based on their location and values simultaneously. The algorithm of Moran's I involves the generation of inverse distance weights matrix, in which entries for pairs of points that are close to each other have higher weights than the pairs that are far apart. For more practical distance calculations, we had to project the observations coordinates from geographic WGS 84 longitude and latitude into a projected coordinate system. The new coordinate system has to preserve distances. North America Equidistant Conic was the projected coordinate system used. The Moran's I Index value and both z-score and p-value to evaluate the significance of the Index were calculated. Moran's I test was performed through using Spatial Autocorrelation (Moran's I) tool in ArcGIS.

Regression analysis

For performing initial analysis to explore the trends in first flower observations, we used regression analysis. Linear regression was performed for each station that has 30 years or more of monitoring. First the stations that have 30 years or more of monitoring were selected. Then a regression analysis was performed to analyse the observations of each station independently, in which the day of first flower was the dependent variable and the year was the independent variable. The resulted slopes of such a regression can hold information about the trend in the onset of spring for each station location. The stations that give negative slope indicate that spring is happening earlier over time at their areas. While the stations that give positive slope indicate that spring is retreating over time at their areas. Regression analysis was performed using the R statistical software.

4.2. Lilac data preparation

During exploring the Lilac dataset, first flower observations were found to be incomplete and to have many missing values. This situation precludes proper analysis of the dataset and reaching conclusions about trends in the onset of spring. Therefore during this stage an attempt to complete Lilac first flower observations was performed. The SI-x version of Spring Index model, which was discussed in section 3.3, was used for simulating first flower events. The model needs temperature information for each Lilac station location in order to simulate first flower events at that location. The temperature daily information for each station location was downloaded from Daymet website and fed to the model. The steps applied to complete this task are shown in Figure 14.

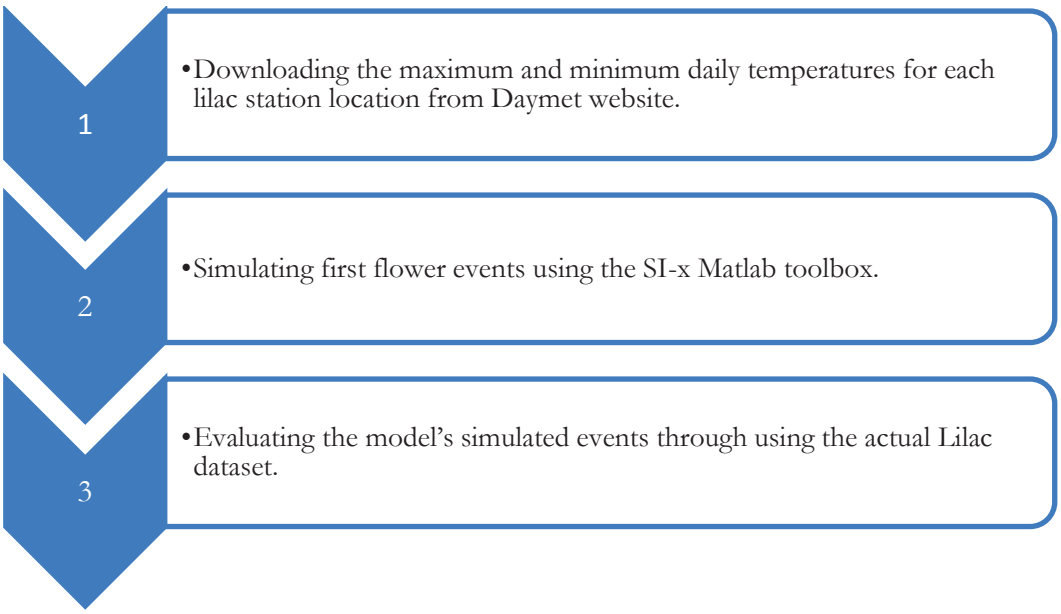


Figure 14: The steps applied for simulating Lilac first flower events.

The description of each step is as the following:

1. Downloading the maximum and minimum temperatures for each lilac station location from Daymet website. The multiple coordinates extraction tool available with Daymet website was called from a Matlab script. This script was done to read lilac stations locations from a text file and write the corresponding temperatures acquired from Daymet to another text file. The Matlab script was applied to acquire daily maximum and minimum temperatures since 1980 until 2012 for each Lilac station location.
2. Simulating first flower events using the SI-x Matlab toolbox. For completing this step, another Matlab script was prepared to enable the SI-x toolbox to read the temperatures text files acquired from Daymet website. The toolbox was used to simulate a complete first flower record for each station for the period in between 1980 and 2013.
3. Evaluating the model's simulated events through using the actual Lilac dataset. Route Mean Square Error (RMSE) was calculated for the simulated first flower values, as well the average error at each station was calculated and mapped in geographic space.

4.3. Lilac data mining unsing SOM

The previous stages that started by understanding the Lilac dataset and ended by completing the missing values were preparation stages for Lilac data mining. In this stage, SOM was applied to find trends in the onset of spring and cluster them. This was set to be the main objective in the Lilac data knowledge discovery process. SOM was applied for Lilac data mining because of these reasons:

- SOM has the ability to process high dimensional datasets and project complex relations into two dimensional surface.
- SOM has been successfully applied to find out clusters in time and space from spatio-temporal datasets. Especially, what has been defined by Andrienko et al. (2010) as analysing space in time.
- SOM is a data mining technique that allows visual interaction between the analyst and the data models prepared, which leads to more verified and reliable results (Santucci & Hauser, 2010).

SOM was applied to find clusters in Syringa Chinensis first flower values that were simulated using SI-x model. Syringa Chinensis first flower values were simulated for 193 stations over a 24 years period, in between1980 and 2003. For processing this dataset with SOM, first flower values were structured as shown in Table 1. Where: (St_1… St_{139}) represent the 193 stations, (Y_1… Y_{24}) represent the 24 years and X represent the first flower simulated value for a certain station in a certain year.

Table 1: The structure of Syringa Chinensis first flower dataset processed by SOM.

		Years			
		Y_1	Y_2	…	Y_{24}
Stations	St_1	$X_{1,1}$	$X_{1,2}$	…	$X_{1,24}$
	St_2	$X_{1,2}$	$X_{2,2}$	…	$X_{2,24}$
	...	…	…	…	…
	St_{139}	$X_{139,1}$	$X_{139,2}$	…	$X_{139,24}$

SOM training parameters

For training SOM, R statistical software including the "kohonen" package was used. SOM training parameters were identified as the following:

Grid shape: The grid shape of the map was chosen to be hexagonal. This shape is recommended for better accuracy. At the hexagonal shape, all the neighbours of the neuron are in the same distance, while this is not the case with the rectangular shape (Kohonen, 2013).

Grid dimension: The map size should be proportional to the number of data objects, in order to be able to detect the deviation of the data. If the map size is too large, then it is possible to overfit the data (Vesanto, 2002) . Therefore, the grid dimension was set to be 12 × 15. This makes the total number of neurons 180. This is close to the number of Lilac stations, which is 193. As a result of the chosen dimensions, every neuron represented a single station expect for 13 neurons, which represented 2 stations.

Neighbourhood radius: The neighbourhood radius is recommended to be 2/3 unit to unit distances (Vesanto, 2002). So we applied the same criteria.

Number of iterations: Number of iterations should be large to achieve better mapping accuracy. It is recommended to be at least 500 times the number of neurons (Vesanto, 2002). In our case we had 180 neurons, so the number of iterations was set to 90,000.

SOM results

After training the dataset with the presented parameters, the results of SOM were visually inspected. The U-matrix plot, which represents the distances between each neuron and its neighbours, was checked. The plot for the number of stations per neuron after training the dataset was checked too. Additionally, we examined the plot that presents the mean distance between objects mapped to a neuron and the codebook vector of that neuron. This map should show small distances as an indicator of good SOM results.

Identifying clusters in SOM results

For identifying clusters in SOM results, usually visualizing distances between each map unit and its neighbours at U-matrix is used. Unfortunately, identifying clusters in this way may lead to different results when performed by different people (Vesanto & Sulkava, 2002). Therefore, finding clusters in an automated way using SOM findings has been suggested. Vesanto and Sulkava (2002) proved that clustering algorithms for SOM findings based on distance matrix, produced good results.

For identifying clusters in SOM results at our case study, we adopted an automated clustering technique that is based on the distance matrix. This was done through applying an automotive clustering option in R “kohanen” package. The number of clusters was set to six. We tried different numbers of clusters but we find that six clusters is a proper number compared to the number of stations. The resulted clusters were later projected into geographic space. These maps were inspected to assess the relationship between SOM clusters and their geographic location. The geographic maps were prepared using ArcGIS.

5. IMPLEMENTATION OF THE WORKFLOW

This chapter presents the results of each stage in the workflow developed for revealing information about the onset of spring from Lilac dataset. The sequence of the sections presenting the results is the same sequence followed at the methodology chapter. The first section presents the results of data understanding stage. The second section presents the results of data preparation stage. While the third section presents the results of data mining using SOM. Each section includes a discussion about the results of the addressed stage.

5.1. Lilac data understanding

This section presents the results of the tasks implemented during understanding Lilac dataset. The first sub-section presents the results of investigating Lilac data lineage. The second sub-section presents the results of Lilac data exploration and quality verification.

5.1.1. Lilac data lineage

The historical Lilac dataset is the outcome of several phenological networks in the US. These phenological networks started at the 1950s with a series of regional agricultural experiment station projects, which were directed by the US Department of Agriculture (USDA). The main aim of these projects was to characterize seasonal trends in the US through using phenology (Schwartz, 1994; USA National Phenology Network, 2011). Therefore, Lilac and Honeysuckle were monitored at USDA projects to indicate the onset of spring. These two species were chosen for this mission because of their special properties, which are: relatively good insect and disease resistance, phenological stages that are easy to observe, cold hardiness, resistance to heat and drought, broad distributional range and adaptability to a variety of soil types (Schwartz, 1990).

The regional agricultural experiment station projects that were involved in collecting the Lilac dataset were motivated by the success of the program established in 1956. This program was initiated by Joseph Caprio (Montana State University) under project W-48, "Climate and Phenological Patterns for Agriculture in the Western Region" in the western US. Later, two projects NC-26 "Weather Information for Agriculture" and NE-35, "Climate of the Northeast-Analysis and Relationships to Crop Response" started in the central and north-eastern states in 1961 and 1965, respectively. In 1970, phenological programs conducted under NC-26 and NE-35 were combined in a new regional project, NE-69, "Atmospheric Influences on Ecosystems and Satellite Sensing". However, observations in the western network continued until 1994, while the eastern and central networks were terminated in 1986 because of losing funding (Schwartz, 1994; Schwartz, Betancourt, & Weltzin, 2012). The eastern network continued collecting phenological information even after the decommissioning because M.D. Schwartz corresponded with the network supervisors to continue participating in an interim network from 1986 until 2004 (USA National Phenology Network, 2011).

One important detail to be recalled here is about the observed Lilac types. The western region network was observing Syringa Vulgaris Lilac besides two species of Honeysuckle, while the central and north-eastern regions networks were observing Syringa Chinensis Lilac beside the same two Honeysuckle species. Furthermore, all the plants of Syringa Chinensis Lilac were of the same genetic clone (Schwartz, 1994).

All the mentioned phenological networks depended on volunteer observers. The volunteers included a few thousand cooperative weather service observers, scientists and technicians at agricultural stations and garden club members. The volunteers provided phenological information about leafing and flowering of Lilac and Honeysuckle by US mail (Schwartz et al., 2012).

Discussion

From investigating the lineage of Lilac data, we realized that Lilac data provides spatially extensive phenological information along the US for almost 50 years. The dataset was collected with the aim of characterizing seasonal trends, which is the same objective of the Lilac knowledge discovery process. Nevertheless, the dataset was found to have some issues, which are:

- Lilac dataset is the output of different observation networks. These networks had different periods of monitoring: the late 1950s in the west of the US and late 1960s in the east of the US.
- The changing character of the networks collected Lilac observations. The monitoring networks had some major changes such as combing the central and north-eastern networks.
- Each network was monitoring different type of Lilac. The western network was monitoring Syringa Vulgaris, while the eastern network was monitoring Syringa Chinensis.
- The Lilac dataset was collected by a large number of professional volunteers but from different backgrounds.

The consequences of the presented issues on the characteristics of Lilac dataset were recommended to be examined during performing the data exploration and quality verification tasks.

5.1.2. Lilac data exploration and quality verification

A series of methods and techniques have been used during Lilac data exploration and quality verification. The findings of these methods and techniques are presented below.

Visual exploration

The visual exploration approach started by exploring the spatial distribution of the monitoring stations. The locations of the 1,126 stations were mapped on a backdrop of the US states boundaries, as shown in Figure 15. The symbology of the stations represents the type of Lilac monitored by each station. From the produced map, we could notice:

- The boundaries between the two main networks: the eastern and western could be recognized, hence each network observed different type of Lilac. The eastern network observed Syringa Chinensis while the western network observed Syringa Vulgaris.
- The irregular spatial distribution of the stations along the US. The stations density is high in the western and north-eastern regions of the US while it's less in the central region.
- The south-eastern region has no Lilac stations. This was found to happen because of the weather conditions at that region, as Lilac don't receive sufficient chilling to grow there (USA National Phenology Network, 2011).

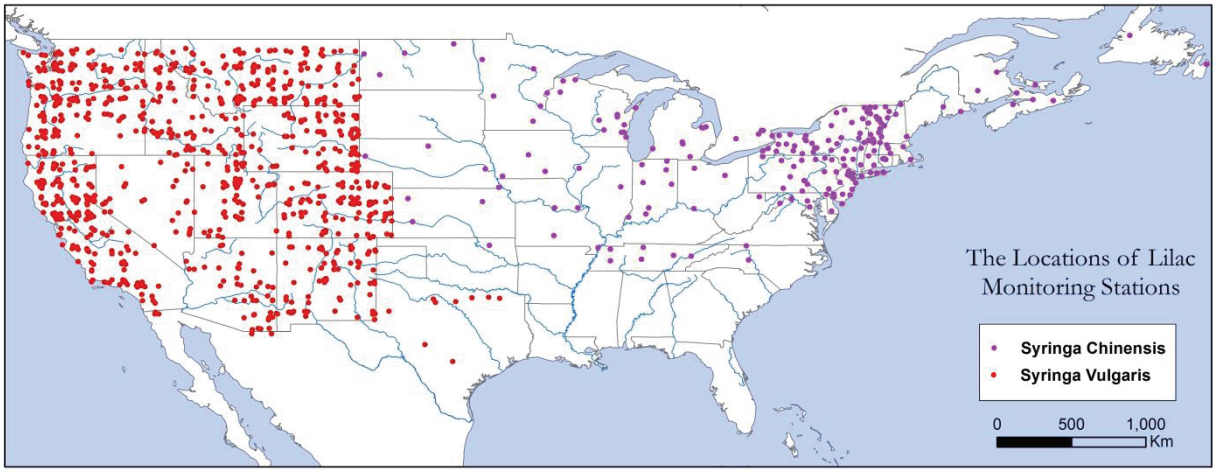


Figure 15: The locations of Lilac monitoring stations symbolized according to the observed Lilac type.

For interactively explore first flower observations, we used space time cube. The two planar dimensions of the space time cube represented the geographic coordinates of the observations, while the third vertical dimension represented the years in which the observations were made. A snapshot for the used space time cube is shown in Figure 16. By using this technique, we could notice:

- The incompleteness of first flower observations. The temporal gaps in the records of the stations were clear, especially during the last ten years 1990-2000.
- The different time coverage between observations of the eastern and the western networks. As we could observe that the western network (monitoring Syringa Vulgaris) was active since the end of the 50s, while the eastern network (monitoring Syringa Chinensis) started to be active by the end of the 60s.

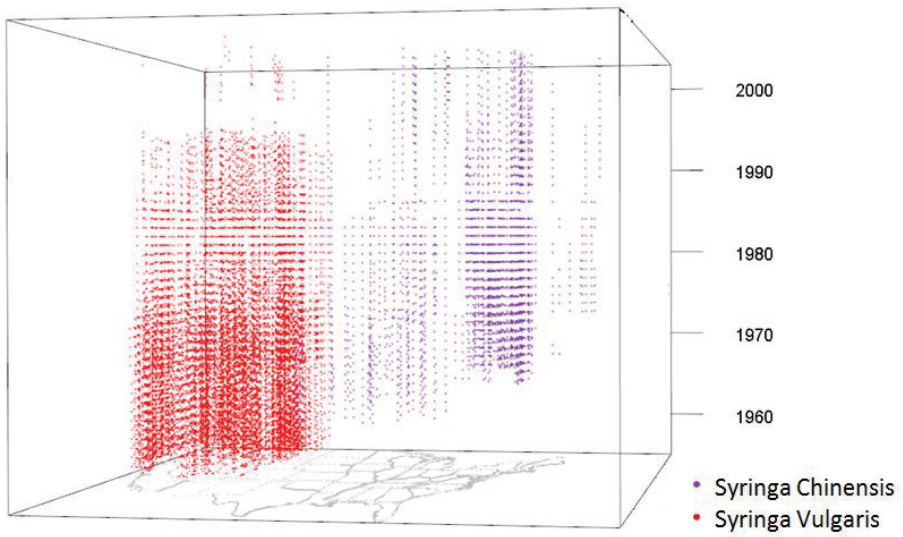


Figure 16: First flower observations represented in space time cube.

EDA statistical approach was applied to examine the values of first flower observations. We started the EDA approach by checking the values of first flower attribute. We found that 14,265 records out of 15,072 records had first flower values, while the other 807 records were shown as missing values. The 14,265 observations were monitored by 1,125 stations, since there is one station that had first leaf observations but not any first flower observation. The way, in which Lilac first flower observations and the stations monitored them were divided between the two Lilac types, is presented in Table 2.

Table 2: The way Lilac first flower observations are divided between the two Lilac types.

	Syringa Vulgaris	Syringa Chinensis	Total
Number of observations	11,357	2,908	14,265
Number of stations	931	194	1,125

Different statistical representations were produced for first flower observations. Histograms and boxplots results are described below, as the other produced representations gave almost the same conclusions. Two histograms were created, one for each Lilac type, as shown in Figure 17. The purpose of preparing histograms was to explore the distribution of first flower values and as well to check for outliers. At the created histograms, the difference between the number of Syringa Chinensis and Syringa Vulgaris observations was clear. We didn't notice the existence of significant outliers.

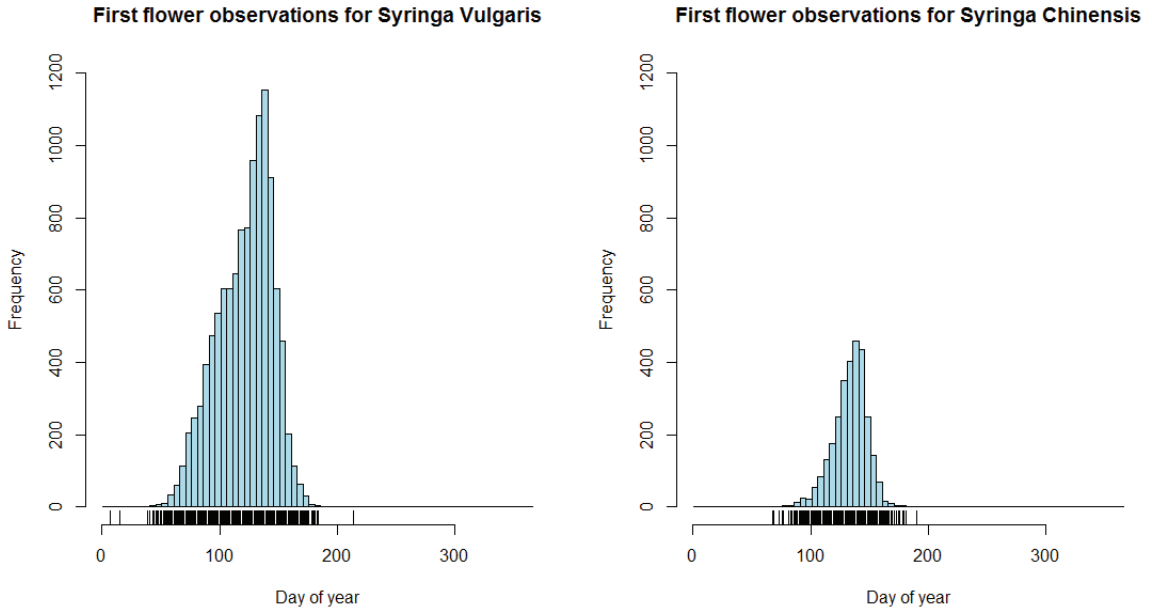


Figure 17: Histograms for first flower observations, one for each Lilac type.

In a trial to explore trends in first flower observations, we used boxplots, in which observations of first flower were plotted versus their corresponding year. Two boxplots were prepared, one for each Lilac type, as shown in Figure 18. It was hard to get any conclusion about trends over time by just depending on these boxplots. However, we noticed the bigger boxes for Syringa Vulgaris observations compared to Syringa Chinensis observations. This means that Syringa Vulgaris observations have a

relatively broader range of first flower dates. This could be dependent on the terrain properties of the west region of the US, where Syringa Vulgaris is monitored. The west region of the US is distinguished by the existence of high mountains and significant diversity in the terrain. This is not the case at the east region of the US, where Syringa Chinensis is monitored.

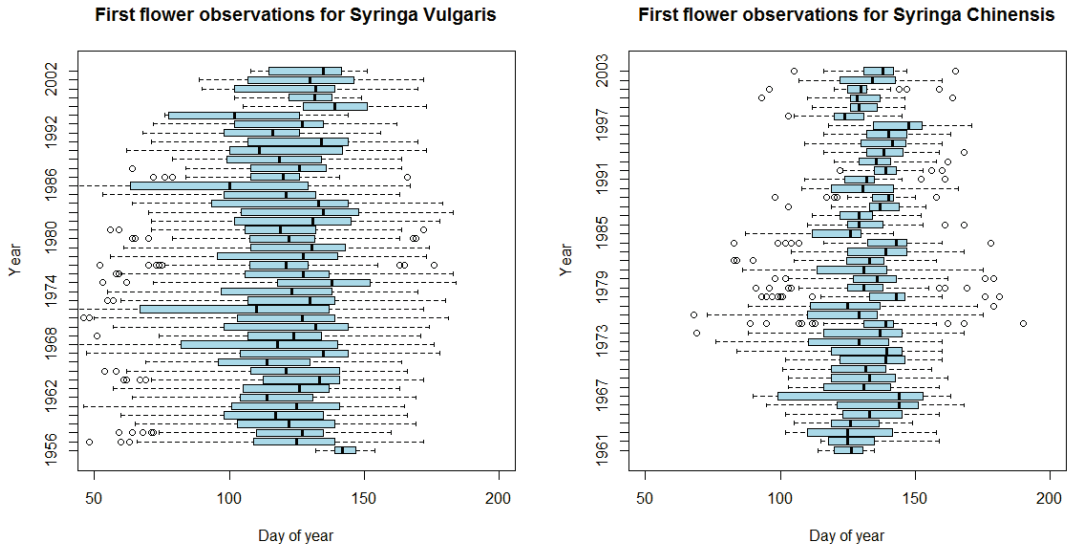


Figure 18: Boxplots for first flower observations per year, one for each Lilac type.

Through using the space time cube, we noticed the different time coverage between the observations of the eastern and the western networks. For additional exploration regarding this issue, we prepared boxplots for the coverage period of each network, as shown in Figure 19. We noticed that the western network (monitoring Syringa Vulgaris) had more observations between the mid of 60s till the mid of 70s, while the eastern network (monitoring Syringa Chinensis) had more observations between the mid of 70s till the mid of 80s. This confirmed our previous note that each network has different time coverage.

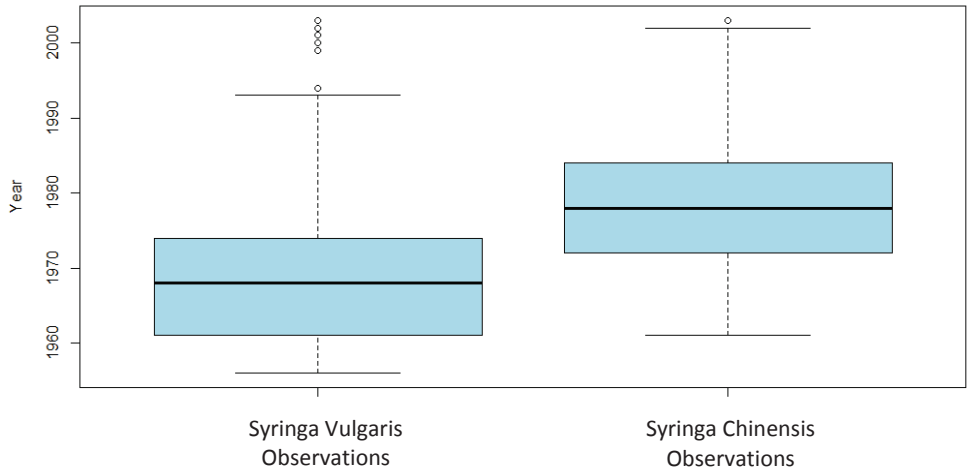


Figure 19: Boxplots for the coverage period of each network.

Moran's I test

Moran's I test was performed to investigate the existence of spatial autocorrelation between first flower observations. The test was applied to the observations of each year separately. The results showed that first flower observations are spatially correlated, whereas the results for 47 years showed positive spatial autocorrelation, while just a single year didn't show any spatial autocorrelation between its observations. This year is 1956, which is the first year of monitoring. A map that shows the locations of the observations at 1956 was visually inspected. This year was characterized by having a low number of observations, which were spatially scattered. This could explain the absence of spatial autocorrelation between the observations of that year. The values of Moran's I index, z-score and p-value for the observations of each year are presented in Appendix A.

Regression analysis

Linear regression analysis was performed to analyse the trend of each station that has 30 years or more of monitoring. The selection of the targeted stations resulted in 43 stations. Regression analysis was performed for each station separately, in which the day of first flower observation was the dependent variable and the year of the observation was the independent variable. Out of 43 stations, 39 stations showed negative regression slopes. This means that spring was happening earlier over time at the areas of these stations. While the other 4 stations showed insignificant positive regression slopes.

Although the results of the regressions indicated that spring is happening earlier over time in most of the analysed locations, this analysis had a major problem. The stations were found to have different time coverage. In addition, some of the stations were found to have gaps in their records. This situation prevented us from using the above mentioned results to make conclusions about trends in the onset of spring. Nevertheless, this analysis confirmed that there is a serious problem of missing observations within Lilac dataset that would prevent analysing trends. The regression results along with the time coverage of each station are shown in Appendix B.

Discussion

By the end of data exploration and quality verification process, we could define the main characteristics of Lilac dataset and its major quality problems. Using different visual representations such as maps, space time cube and statistical plots allowed us to form a comprehensive vision about the dataset and especially first flower observations.

The main issue discovered during this process is the incompleteness of first flower observations. In fact, the dataset has 1,126 stations and 48 years of monitoring, between 1956 and 2003. If we assumed that each station has a complete record of first flower observations, there should be more than fifty thousand observations but the collected observations were just 14,265. This large number of missing values could be explained by the changing character of the phenological projects that monitored Lilac first flower event. As a result, most of the stations didn't have a complete record of Lilac first flower observations.

A summery for different periods of monitoring Lilac first flower and their corresponding number of stations is shown in Table 3. This summery confirms the issue of missing values. The table shows that almost half of the stations have 1 to 10 years of monitoring. Moreover, there isn't any single station with a complete record of first flower observations, which is 48 years. The numbers of stations, which have more than 30 years of monitoring, were just 33 stations. Even the stations, which have a relatively high number of observations, were found to have different time coverage.

Table 3: Different periods of monitoring with the corresponding number of stations for each period.

Period of monitoring (years)	Number of stations
1 - 10	553
11 - 20	408
21 - 30	132
31 - 40	33

For solving the problem of missing values, we suggested using the Spring Index model, which is discussed in section 3.3, for simulating first flower events. The Spring Index model was originally designed for this purpose. Schwartz (1994) stated that "When lacking complete period-of-record in phenological information, one way of exploring trends on a regional basis is to develop analog models". The author clarified that although the model's simulated values can't be as accurate as real observations, such a model would be the only way for solving the problem of missing values in phenological records.

Spring Index model has been used to simulate missing values in Syringa Chinensis observations. The use of Spring Index model was discussed in a series of publications such as Schwartz and Reiter (2000), Schwartz et al. (2006) and Schwartz et al. (2013). But none of the reviewed publications have used Spring Index model to complete Syringa Vulgaris observations as well. In addition, the last version of Spring Index model SI-x hasn't been used yet for the purpose of completing Syringa Chinensis or Syringa Vulgaris observations. Accordingly, the last version of Spring Index model SI-x was recommended to be used for simulating the missing values in first flower observations to be able to analyse trends properly.

Another important issue that was revealed during Lilac data exploration is the existence of spatial autocorrelation between first flowering observations. Actually, this result was expected. Ecological data are usually characterized by the existence of spatial autocorrelation. It is considered to be the geographic result of the interaction of geologic, climatic, topographic, and biological variables. Though the result was expected, it's important to perform the autocorrelation check in such datasets. The existence of spatial autocorrelation influences the selection of the appropriate analytical techniques. Many statistical analytical techniques assume independency between observations, such as linear regression. Accordingly, it's improper to apply these techniques for analysing spatial dependent datasets.

Brunsdon and Comber (2012) discussed the consequences of ignoring spatial autocorrelation between Lilac first flower observations. The authors have showed that performing linear regression analysis for the observations pooled from all the stations gave misleading results, as shown in Figure 20. The result of this regression was positive slope, which means that first flowering is happening later over time. This result doesn't conform to the results acquired from the regression analysis applied to the stations one by one. The authors clarified that if location is ignored, the average first flower will retreat over time. This is because the eastern network stations, which are relatively in a colder region, are dominating the observations at the end of the study period. That was clear in the boxplots presented in Figure 19. The presented example explains the importance of investigating the spatial and temporal dependency in spatio-temporal datasets. This can prevent applying improper analytical techniques that lead to wrong results.

As it was discussed at the previous chapter, the findings of data understanding stage would affect the next stages in the knowledge discovery process. This is true, as after completing the Lilac data understanding stage we had two main recommendations for the next stages. First, the problem with missing values should be solved before going to data mining. Second, the selected data mining technique to analyse first flower observations should consider the spatial and the temporal dependencies between the observations. Therefore, data preparation stage was performed to solve the missing values problem.

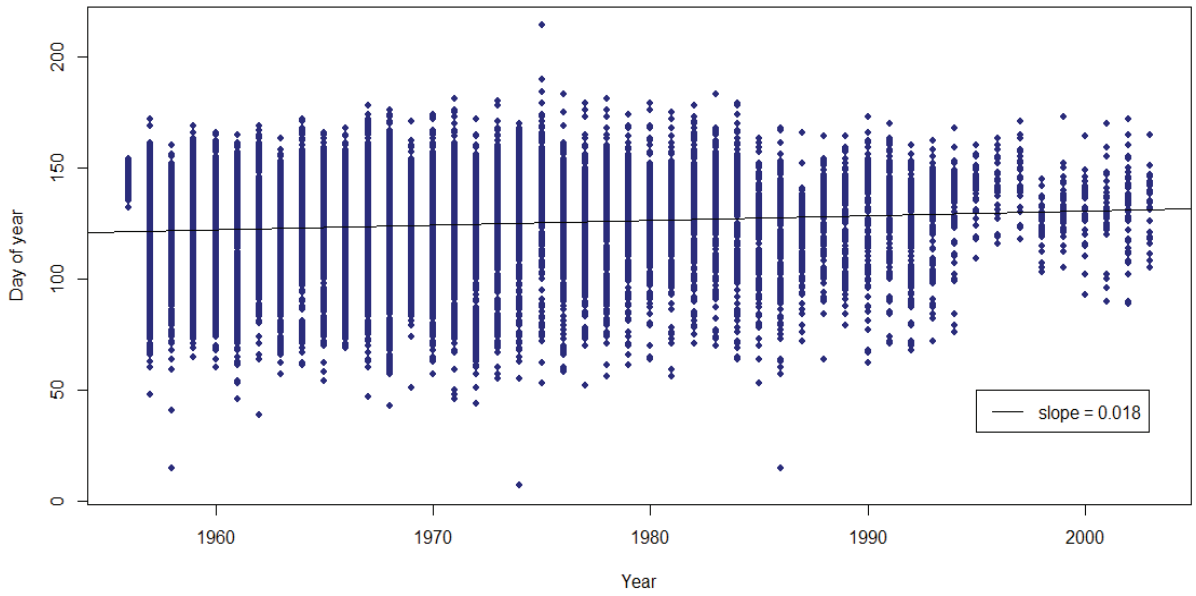


Figure 20: Linear regression analysis for first flower observations pooled from all the stations.

5.2. Lilac data preparation

This section presents and discusses the results of the three steps applied to simulate a complete record for Syringa Chinensis and Syringa Vulgaris first flower dates. The first step was downloading the daily minimum and maximum temperatures from Daymet website. The daily temperatures provided by Daymet website were available for the years in between 1980 and 2012. We succeeded in downloading the temperatures for 1113 stations locations. The temperatures for 13 stations locations couldn't be downloaded. Daymet website kept giving an error for these 13 locations in particular that we couldn't solve. Therefore, the result of this step was 1113 text files; each file includes daily minimum and maximum temperatures since 1980 to 2012 for a certain monitoring station.

The second step was simulating first flower events using the Matlab toolbox, which is based on SI-x version of Spring Index model. For this purpose, the downloaded daily temperatures from Daymet website along with the latitude of each station were fed to the toolbox. The result of this step was complete first flower simulated values for the 1113 stations over 24 years period, starting on 1980 and ending on 2003. The results were structured in a table that had the following attributes: Station ID, year and DOY for first flower. The table had 26,712 records (24 years × 1113 stations).

The third step was evaluating the simulated first flower events. For this purpose, the simulated values were compared to their corresponding actual Lilac first flower observations. Out of 26,712 simulated values only 2,980 values had corresponding actual observations. In which, 1,255 values belong to Syringa Chinensis and 1,725 belong to Syringa Vulgaris. Root Mean Square Error (RMSE) was calculated according to the following equation:

$$\text{RMSE} = \sqrt{\frac{\sum^{n}(\hat{y}_i - y_i)^2}{n}}$$

Where:

n = The number of years.

$\hat{y}_i$ = First flower simulated value for the year i .

y_i = First flower observed value for the year i .

RMSE was calculated three times: for all the simulated values, only for Syringa Chinensis simulated values and only to Syringa Vulgaris simulated values. The RMSE values were as the following:
RMSE for all the simulated values = 17.7 days
RMSE for Syringa Chinensis simulated values (East network) = 5.5 days
RMSE for Syringa Vulgaris simulated values (West network) = 22.8 days

Plots for actual observations versus simulated values were done for the three results. Figure 21 shows a plot for the actual observations versus all the simulated values. The regression of actual observations versus simulated values and 1:1 line are shown in the plot too. Figure 22 shows two plots for the actual observations versus the simulated values for Syringa Chinensis and Syringa Vulgaris separately. The regression of actual observations versus simulated values and 1:1 line are shown in theses plots too.

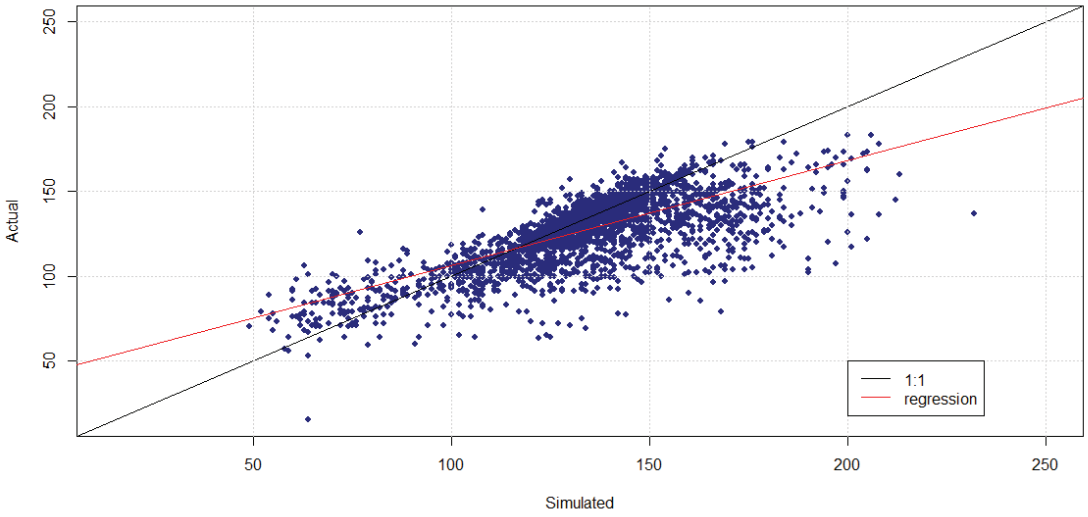


Figure 21: Actual first flower observations plotted versus simulated values, for both Lilac type.

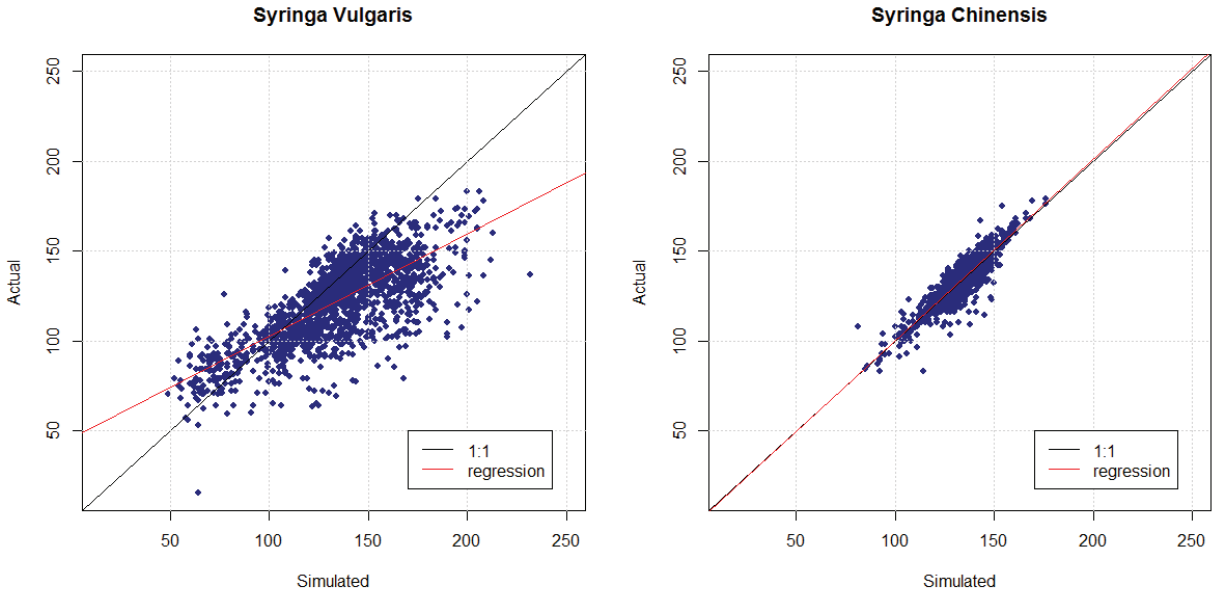


Figure 22: Actual first flower observations plotted versus simulated values, for each Lilac type.

Discussion

The results of this stage showed two main obstacles that affected the plan of having complete first flower records for all Lilac stations, starting 1956 and ending 2003. These obstacles are:

- The daily temperatures provided by Daymet website were available for the years in between 1980 and 2012. This issue limited our plan with downloading daily temperatures since 1956, in which monitoring Lilac first flower started.
- The RMSE value for Syringa Vulgaris was found to be around 23 days. Such a value is not acceptable to achieve the objective of the knowledge discovery process, which is finding trends in the onset of spring. This issue limited our plan with having a complete first flower values for both Syringa Chinensis Lilac and Syringa Vulgaris Lilac.

For the sake of completing the research, we used Syringa Chinensis first flower simulated values for the available period starting 1980 until 2003 as an input for data mining.

However, the problem of having high RMSE for Syringa Vulgaris simulated values was further checked. One possible reason for this high RMSE is the diversity of the terrain in the west region of the US, where Syringa Vulgaris Lilac is monitored. The changing terrain could have caused the inaccurate SI-x model's results at these regions. Especially that the temperature measurements fed to the model were obtained from 1km × 1km gridded surface. The resolution of this surface could be coarse for diverse terrain areas. For checking this hypothesis, we calculated the average error at each station then mapped the errors in geographic space.

The Average Error (AE) for each station was calculated according to the following equation:

$$\mathrm{AE} = \frac{1}{n} \sum^{n} (\hat{y}_i - y_i)$$

Where:

n = The number of years. It could be different from a station to another, as this depends on finding corresponding observed value to the simulated value for a certain station at a certain year.

$\hat{y}_i$ = First flower simulated value for the year i.

y_i = First flower observed value for the year i.

A map for the stations symbolized according to their average error values is shown in Figure 23. The errors caused by overestimated or underestimated simulated values are distinguished by different symbols. An overestimated value is the one that was simulated later than it should be. While an underestimated value is the one that was simulated earlier than it should be. We could notice from this map that the simulated values for Syringa vulgaris stations located in the middle of the US are more accurate than the simulated values for the stations located in the west of the US. The locations of the average errors were visually compared with a Digital Elevation Model (DEM) for the US. The DEM used is shown in Figure 24. Through comparing the average error with the elevations, we could notice that the errors increases in the diverse terrain areas. Therefore, the inaccurate SI-x model's results for Syringa Vulgaris values could be referred to the diverse terrain in the west of the US.

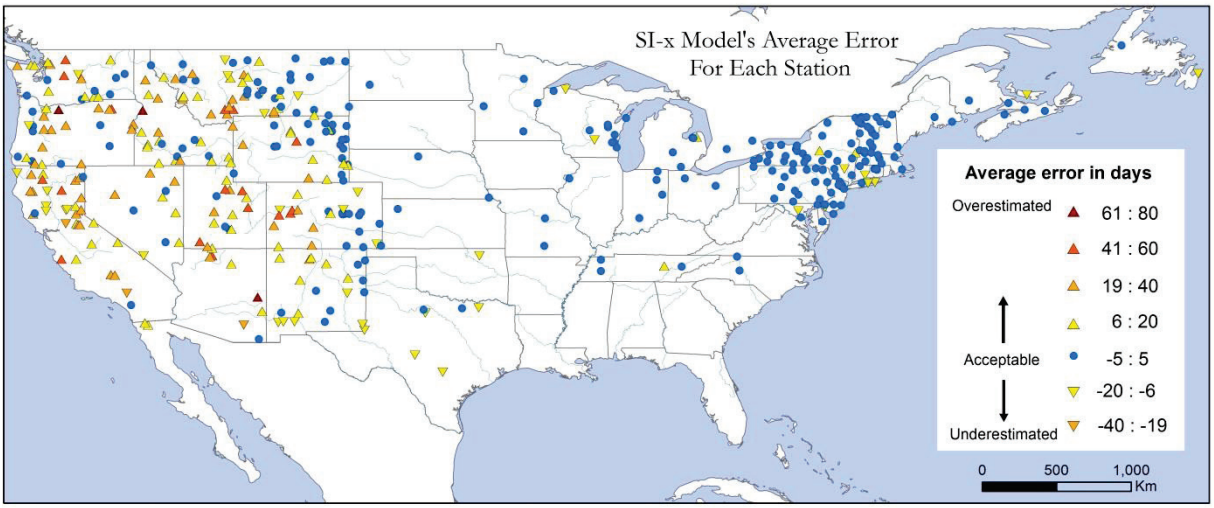


Figure 23: SI-x model's average error for each station.

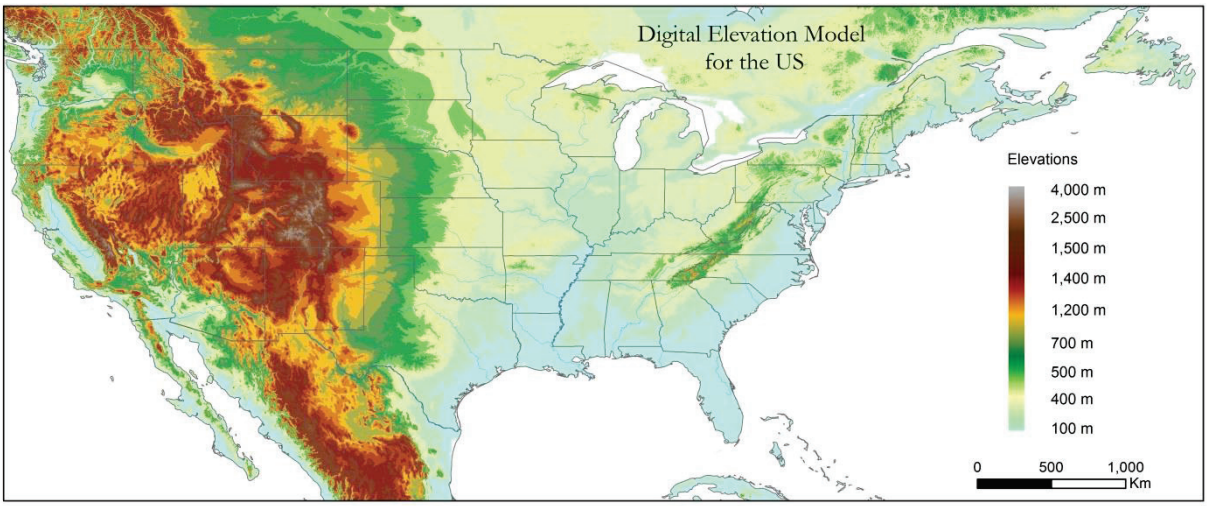


Figure 24: Digital Elevation Model for the US.

5.3. Lilac data mining using SOM

This section presents and discusses SOM results. SOM was used to find clusters in Syringa Chinensis first flower simulated events. Syringa Chinensis first flower events were simulated for 193 stations over a 24 years period that extends from 1980 to 2003. For training this dataset using SOM, it was structured as shown in Table 1, while the applied training parameters were as described in section 4.3. The resulted counts plot and distance plot are presented first. Then, the resulted U-matrix is presented along with the clusters derived using the automated clustering technique. Finally, the projection of the derived clusters into geographic space and time space is presented.

The counts plot, which illustrates the number of stations per neuron after training the dataset using SOM, is shown in Figure 25. The number of stations per neuron is symbolized by different colours. Before training the dataset every neuron represented a single station expect for 13 neurons, which represented 2 stations. After training the dataset, we noticed that the arrangement of stations within the neurons has changed so there became some neurons that represented three stations. This is the result of training the dataset so stations that have similar signals of first flower events over the years are clustered.

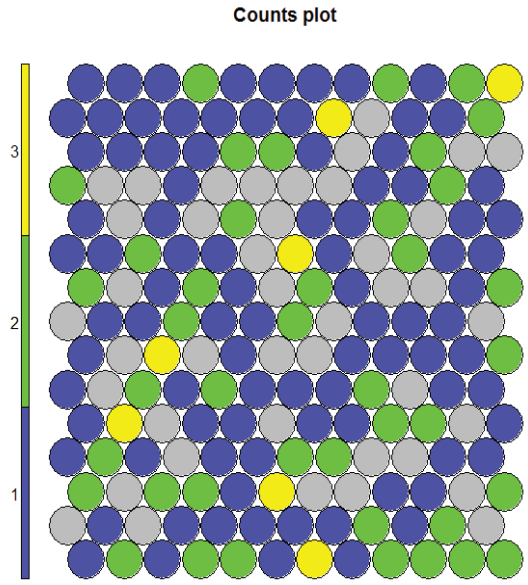


Figure 25: SOM counts plot.

The distance plot that shows the mean distance between objects mapped to a neuron and the codebook vector of that neuron is shown in Figure 26. This plot should show small distances as an indicator of good SOM results. Almost all of the neurons in our case showed very low distances, expect two neurons for which we observed relatively bigger distances. This plot can point out the good quality of SOM results in our case.

The U-matrix plot, which represents the distances between each neuron and its neighbours, is shown in Figure 26 to the left. For clustering the neurons, we used an automotive clustering technique that is based on the distance matrix. The result of clustering the neurons to six clusters is shown at the same figure to the right. The black lines represent the clusters borders.

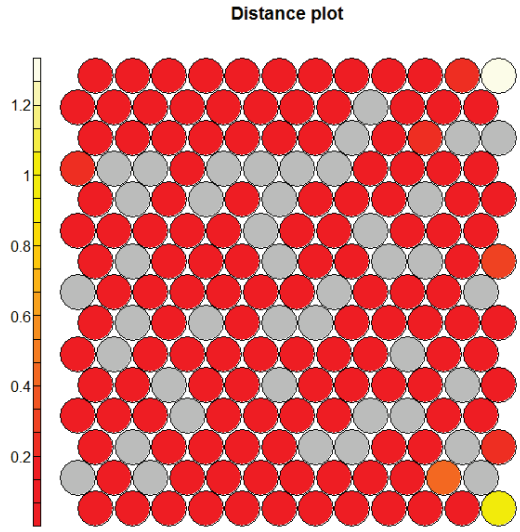


Figure 26: SOM distance plot.

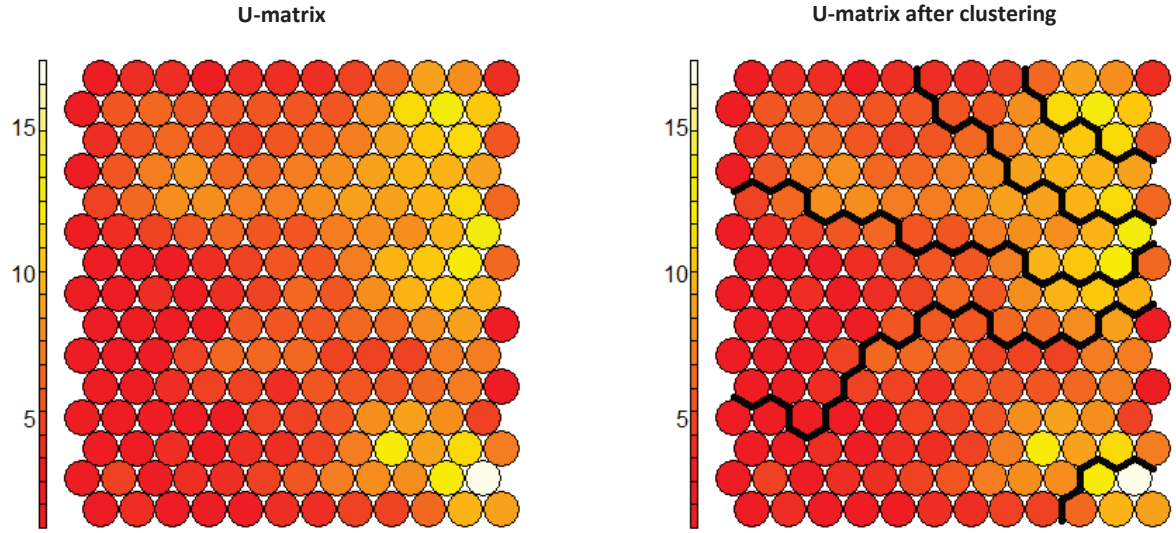


Figure 27: U-matrix resulted of SOM training (left) and the clustered U-matrix (right).

Stations grouped in one SOM cluster are the stations that have similar trends in first flower events along the 24 years period. If we recalled that Lilac first flower event is used as a spring indicator, then projecting SOM clusters in geographic space can allow identifying regions that have similar trends in the onset of spring. The projection of SOM derived clusters into geographic space is shown in Figure 28. The clusters in data space are shown at the top while the same clusters at geographic space are shown at the bottom. From this figure, we could notice that each cluster in data space was projected to almost certain latitude in geographic space. Nevertheless, in the north east of the US, we could notice that cluster four and cluster five are interfering. This could be explained by topography. The region, where the two clusters are interfering, is distinguished by the existence of diverse terrain. The topography of this region can be viewed at the DEM presented in Figure 24. This shows that beside latitude, elevation plays a role in the occurrence of Lilac first flower.

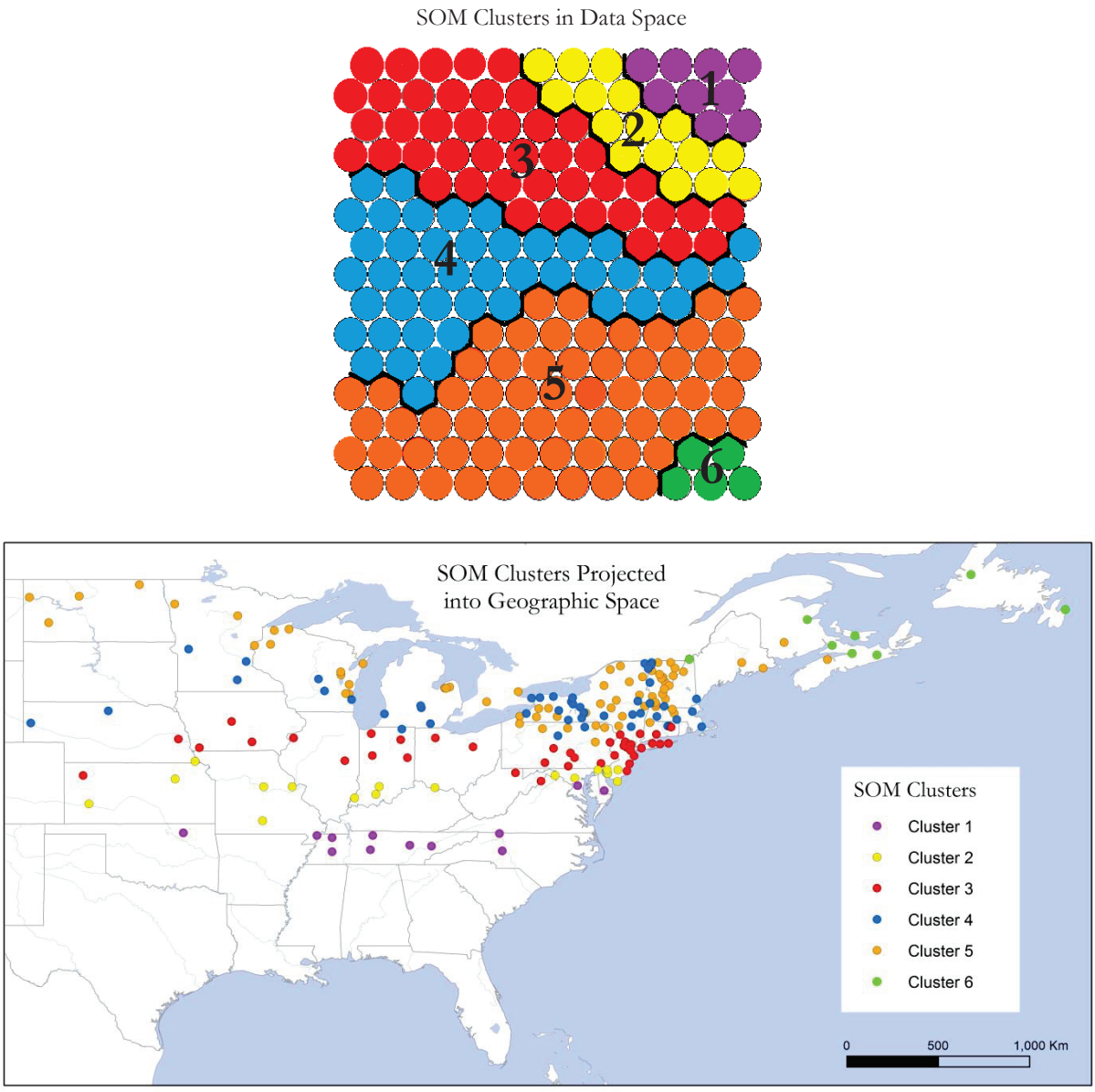


Figure 28: SOM clusters projected into geographic space.

For exploring the signals of the clusters in time space, first flower values of each cluster were plotted versus the year. From these plots, we could notice the temporal match between the signals of the stations at the same cluster. This shows that SOM succeeded in clustering the stations that has synchronised signals of first flower values over time. For exploring the trends of the clusters, regression analysis was performed for the first flower values of each cluster. At this regression, day of first flower event was the dependent variable and the year was the independent variable. The regression analysis showed that five clusters have negative slopes. This indicates that spring is happening earlier over the time at the regions of these clusters. Just one cluster, which is cluster 5, showed an insignificant positive slope. The results of the regression analysis along with the number of stations in each cluster are shown in Table 4. The plots of the clusters along with the regression analysis are shown in Appendix C.

Table 4: Regression analysis results for SOM clusters.

	Number of stations	Slope
Cluster 1	12	-0.20
Cluster 2	17	-0.21
Cluster 3	40	-0.11
Cluster 4	46	-0.02
Cluster 5	70	0.07
Cluster 6	8	-0.41

Another way for exploring SOM clusters in time space was preparing boxplots that shows the dates of first flower events at each cluster, as shown in Figure 27. This figure can illustrate the dominant first flower values for each cluster. If we relate the locations of the clusters in geographic space to these boxplots, then we can conclude that the more north the stations are located the later the onset of spring is happening. Such a result is reasonable and it conforms to the phenomenon we are researching.

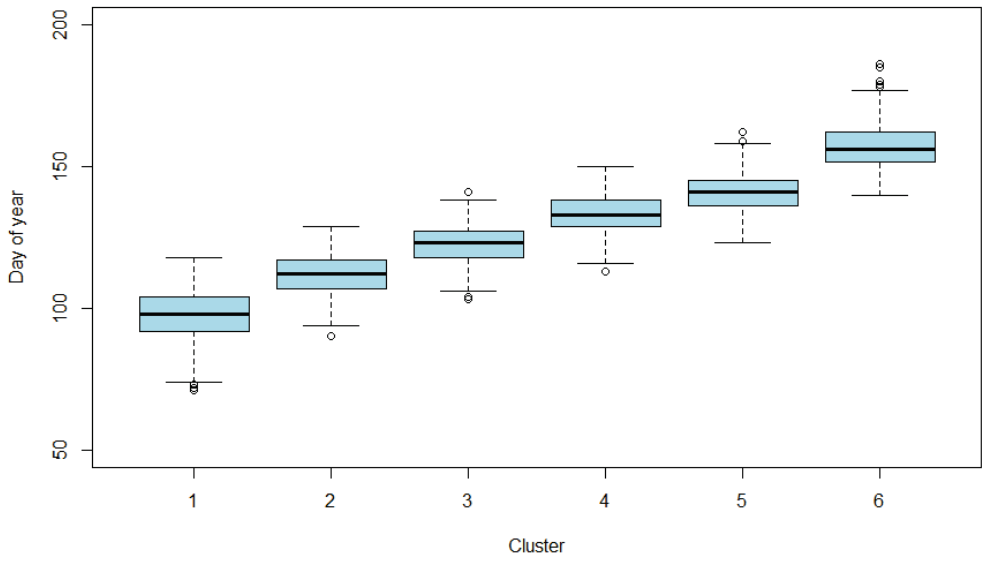


Figure 29: Boxplots for first flower values at each SOM cluster.

Discussion

SOM was successfully applied to find clusters in first flower values. This data mining technique could accommodate to the complexity of space and time dimensions in Lilac dataset. SOM was able to cluster stations that have synchronized signals of first flower values over a 24 years period. At the same time it could allow us to visually inspect the results of the data mining process in different ways. Counts plot, distance plot, U-matrix and even the clusters projected into geographic space were different visual representations for SOM results, which enabled us to be involved in the data mining process.

The automated clustering technique used for deriving clusters gave reasonable results. Projecting the derived clusters into geographic space showed that they are forming clusters there too. Viewing the derived clusters in time space showed that they are homogeneous. Identifying clusters from the U-matrix

using visualization will not give such informative clusters. In addition, it would give different results from person to another. Therefore, from our experience, we recommend the automated clustering for SOM results that is based on distance matrix.

From SOM results we could find that Lilac first flower event is strongly related to the latitude. SOM clusters could show clearly how the day of first flower is happening later the more north the monitoring station is located. We could notice also that elevation plays a role in the occurrence of Lilac first flower. This was clear with the interfering of clusters four and five in the north-eastern region, where the terrain is diverse. The different representations of SOM results enabled us to perceive relations between data, time and geographic spaces.

It would be more informative if we could apply SOM to cluster the first flower signals for Syringa Chinensis and Syringa Vulgaris stations. This could reveal information about trends in the onset of spring for the entire US. But the problem with missing values at Lilac dataset didn't allow this, as more than half of the values were missing.

6. CONCLUSIONS AND RECOMMENDATION

This chapter presents the conclusions of the research in the first section, while the recommendations for future research are presented in the second section.

6.1. Conclusions

The main objective of our research was to develop and implement a workflow for discovering knowledge from spatio-temporal datasets collected by volunteers. In order to achieve the research's main aim, we addressed the historical Lilac dataset. The workflow developed for discovering knowledge from Lilac dataset is adaptable to other spatio-temporal datasets collected by volunteers. The main stages of the workflow and many of the techniques applied can be generalized for similar cases.

The workflow started by defining the objectives of the knowledge discovery process. The defined objectives helped us in making decisions during all the stages of the knowledge discovery process. Decisions regarding what to peruse more and what to ignore, especially when many variables were interfering. The defined objectives could also assist us in selecting the analytical applied techniques. The selection of the techniques was done with respect to the defined objectives. Therefore, prior defined objectives for the knowledge discovery process are recommended.

Data understanding stage included three main tasks, which are investigating the data lineage, exploring the data contents and verifying the data quality. At Lilac case study, we were lucky to have a number of available publications that discuss the dataset origin and how it evolved overtime. This allowed us to form a vision about the characteristics of the dataset and the possible quality problems. When addressing VGI datasets, it's recommended to investigate the data history. If this wasn't applicable then more attention should be given to data exploration and quality verification tasks.

For exploring the data contents and verifying its quality, a series of exploration techniques were applied. The applied techniques showed good performance in handling the tasks assigned to them. Mapping the data in geographic space enabled us to identify the spatial distribution of the observations. Using space time cube allowed us to explore the spatial and temporal distribution of the observations, in addition to investigate the completeness of the observations. EDA statistical representations allowed us to get insight into the data variables and the relations between them. Using Moran's I test allowed us to reveal the existence of spatial autocorrelation between the observations. Finally applying simple analysis techniques such as regression allowed us to get an idea about the trends in the dataset and at the same time helped us to identify problems that exist in the dataset.

During data understanding stage applying exploration techniques that can consider the spatial and temporal dimensions of the dataset is essential. At Lilac case study, we realised the importance of using the space time cube, as exploring the data in geographic space was not enough to identify the incompleteness of the observations. However, there exist other exploration techniques that consider both space and time, such as map animation and time wave. These techniques can also be used for visually exploring spatio-temporal datasets in this stage.

Data preparation stage is proposed to handle the quality problems with VGI datasets before going to data mining. At Lilac case study, data preparation stage is meant to solve the problem of missing values through using the Spring Index model. The problem of missing values within Lilac dataset was not solved properly. As we just could solve this problem for Syringa Chinensis observations but not Syringa Vulgaris observations. In fact, if we could solve the missing values problem, then we would be able to find trends in the onset of spring for the whole US.

Many publications identified quality to be a major problem in VGI dataset (Flanagin & Metzger, 2008; Goodchild & Li, 2012; Yanenko & Schlieder, 2012). From our experience with Lilac case study we agree with these opinions. This makes data preparation stage one of the most challenging stages that requires the application of innovative techniques to solve the quality problems within the addressed VGI dataset.

Data mining stage included applying the data mining technique that can fulfil the main objective of the knowledge discovery process. With Lilac dataset SOM was successfully applied to discover trends in the onset of spring and to cluster them. When addressing other datasets, there are number of spatio-temporal data mining techniques that can be applied in this stage. The thing that should be considered when selecting the data mining technique is the ability of this technique to handle space and time properly, in addition to its ability to fulfil the objectives of the knowledge discovery process.

The stage of presenting the results included presenting the results of data mining techniques in a way that can convey the discovered information. At Lilac case study, the results of the data mining technique were presented in data space, geographic space and time space. This allowed us to create better understanding for the results. Therefore, we recommend presenting the results of mining spatio-temporal datasets in geographic and time spaces.

From working with Lilac case study, we could notice the importance of visualization in the knowledge discovery process. Visualization played a major rule in every stage. At data exploration the visual exploration techniques were essential for identifying the data characteristic and its quality problems. At data preparation stage visualizing the errors in geographic space helped us identifying the problem with the model's simulated values. At data mining and results presentation stages, visualization was essential too.

6.2. Recommendation

By the end of the research, this section presents our recommendations for future work.

Recommendations regarding Lilac case study:

- Further investigation for the reason behind the inaccurate simulated events for Syringa Vulgaris stations. This could be done through using other source for daily temperatures information. Temperatures from ground meteorological stations that are located near to Lilac monitoring stations would be a good alternative.
- Trying to analyse the dataset through using other data mining technique that is robust to missing values. Actually, this will be a hard task to be done because with Lilac dataset more than half of the values are missing.

Recommendations regarding mining spatiotemporal datasets collected by volunteers:

- Programs that depend on volunteers as a source of information should be well planned. The programs should be based on clearly explained objectives and they should follow certain protocols and procedures during data collection to ensure the quality of the collected datasets. This will save efforts spent in solving the quality problems within the collected datasets.
- There is a need for innovative techniques that can solve the quality problems within VGI datasets. The available exploration techniques are capable to identify the problems within these datasets but solving these problems is another issue.
- There is a need for data mining techniques that can be robust to the quality problems within VGI datasets, such as the lack of sampling design and the occurrence of erroneous and missing values.

LIST OF REFERENCES

Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., von Landesberger, T., Bak, P., & Keim, D. (2010). Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. *Computer Graphics Forum, 29*(3), 913-922.

Andrienko, G., Andrienko, N., Schumann, H., Tominski, C., National, U., Dransch, D., & Kraak, M. J. (2010). Space and time. In D. Keim, J. Kohlhammer, G. Ellis & F. Mansmann (Eds.), *Mastering the Information Age Solving Problems with Visual Analytics* (pp. 57-86). Goslar, Germany: Eurographics Association.

Andrienko, N., Andrienko, G., & Gatalsk, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing, 14*(6), 503-541.

Ault, T. R., Zurita-Milla, R., & Schwartz, M. D. (2013). A Matlab toolbox for calculating spring indices from daily from daily meteorological data.

Brown, P. (2011). Weatherwatch: phenology in the UK, *The Guardian*. Retrieved from http://www.guardian.co.uk/news/2011/apr/11/weatherwatch-phenology

Brunsdon, C., & Comber, L. (2012). Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach. *Geoinformatica, 16*(4), 675-690.

Burton, A., Glenis, V., Jones, M. R., & Kilsby, C. G. (2013). Models of daily rainfall cross-correlation for the United Kingdom. *Environmental Modelling & Software, 49*(0), 22-33.

Camossi, E., Bertolotto, M., & Kechadi, T. (2008). Mining Spatio-Temporal Data at Different Levels of Detail. In L. Bernard, A. Friis-Christensen & H. Pundt (Eds.), *The European Information Society* (pp. 225-240): Springer Berlin Heidelberg.

Catlin-Groves, C. L. (2012). The Citizen Science Landscape: FromVolunteers to Citizen Sensors and Beyond. *International Journal of Zoology, 2012*.

Chou, Y.-H. (1995). Spatial pattern and spatial autocorrelation. In A. Frank & W. Kuhn (Eds.), *Spatial Information Theory A Theoretical Basis for GIS* (Vol. 988, pp. 365-376): Springer Berlin Heidelberg.

Cleland, E. E., Chuine, I., Menzel, A., Mooney, H. A., & Schwartz, M. D. (2007). Shifting plant phenology in response to global change. *Trends in Ecology & Evolution, 22*(7), 357-365.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., & Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages & Computing, 18*(3), 255-279.

Dale, M. T., & Fortin, M.-J. (2009). Spatial autocorrelation and statistical tests: Some solutions. *Journal of Agricultural, Biological, and Environmental Statistics, 14*(2), 188-206.

Daymet: Daily surface weather on a 1 km grid for North America,1980 - 2012. (2012). Retrieved 20/11, 2013, from http://daymet.ornl.gov/

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

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers, 102*(3), 571-590.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In M. F. Usama, P.-S. Gregory, S. Padhraic & U. Ramasamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1-34): American Association for Artificial Intelligence.

Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3-4), 137-148.

Francis, C. M., Blancher, P. J., & Phoenix, R. D. (2009). Bird monitoring programs in Ontario: What have we got and what do we need? *The Forestry Chronicle, 85*(2), 202-217.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography *GeoJournal, 69*(4), 211-221.

Goodchild, M. F. (2009). Geographic information systems and science: today and tomorrow. *Procedia Earth and Planetary Science, 1*(1), 1037-1043.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*(0), 110-120.

Gorunescu, F. (2011). Exploratory Data Analysis *Data Mining* (Vol. 12, pp. 57-157): Springer Berlin Heidelberg.

Jiawei, H., & Harvey, J. M. (2009). Geographic Data Mining and Knowledge Discovery An Overview *Geographic Data Mining and Knowledge Discovery, Second Edition* (pp. 1-26): CRC Press.

Keim, D., Kohlhammer, J., Mansmann, F., May, T., & Wanner, F. (2010). Visual Analytics. In D. Keim, J. Kohlhammer, G. Ellis & F. Mansmann (Eds.), *Mastering the Information Age Solving Problems with Visual Analytics* (pp. 7-18). Goslar, Germany: Eurographics Association.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks, 37*(0), 52-65.

Kraak, M.-J., & Li, X. (2012). Explore Multivariable Spatio-Temporal Data with the Time Wave: Case Study on Meteorological Data. In A. G. O. Yeh, W. Shi, Y. Leung & C. Zhou (Eds.), *Advances in Spatial Data Handling and GIS* (pp. 79-92): Springer Berlin Heidelberg.

Lechowicz, M. (2002). Phenology. *Encyclopedia of Global Environmental Change, 2*, 461-465.

Longueville, B. D., Smith, R. S., & Luraschi, G. (2009). *"OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires*. Paper presented at the Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, Washington.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems, 33*(6), 403-408.

Santucci, G., & Hauser, H. (2010). Data Mining. In D. Keim, J. Kohlhammer, G. Ellis & F. Mansmann (Eds.), *Mastering the Information Age Solving Problems with Visual Analytics* (pp. 39-56). Goslar, Germany: Eurographics Association.

Schwartz, M. D. (1990). Detecting the onset of spring: a possible application of phenological models. *Climate Research, 1*(1), 23-29.

Schwartz, M. D. (1994). Monitoring global change with phenology: The case of the spring green wave. *International Journal of Biometeorology, 38*(1), 18-22.

Schwartz, M. D., Ahas, R., & Aasa, A. (2006). Onset of spring starting earlier across the Northern Hemisphere. *Global Change Biology, 12*(2), 343-351.

Schwartz, M. D., Ault, T. R., & Betancourt, J. L. (2013). Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *International Journal of Climatology, 33*(13), 2917-2922.

Schwartz, M. D., Betancourt, J. L., & Weltzin, J. F. (2012). From Caprio's lilacs to the USA National Phenology Network. *Frontiers in Ecology and the Environment, 10*(6), 324-327.

Schwartz, M. D., & Caprio, J. M. (2003). *North American First Leaf and First Bloom Lilac Phenology Data*. Retrieved from: ftp://ftp.ncdc.noaa.gov/pub/data/paleo/phenology/north_america_lilac.txt

Schwartz, M. D., & Reiter, B. E. (2000). Changes in North American spring. *International Journal of Climatology, 20*(8), 929-932.

Turner, A. (2006). *Introduction to neogeography*: O'Reilly.

USA National Phenology Network. (2011). History of lilac and honeysuckle phenological observations in the USA. from https://www.usanpn.org/?q=node/36

Vesanto, J. (2002). *Data Exploration Process Based on the Self−Organizing Map.* (Phd), Helsinki University of Technology, Finland.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM Toolbox for Matlab 5. Helsinki, Finland.

Vesanto, J., & Sulkava, M. (2002). Distance matrix based clustering of the Self-Organizing Map. In J. R. Dorronsoro (Ed.), *Artificial Neural Networks - Icann 2002* (Vol. 2415, pp. 951-956). Berlin: Springer-Verlag Berlin.

Wiersma, Y. F. (2010). Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. *Avian Conservation and Ecology, 5*(2).

Wintle, B. A., Runge, M. C., & Bekessy, S. A. (2010). Allocating monitoring effort in the face of unknown unknowns. *Ecology Letters, 13*(11), 1325-1337.

Wolfe, D., Schwartz, M., Lakso, A., Otsuki, Y., Pool, R., & Shaulis, N. (2005). Climate change and shifts in spring phenology of three horticultural woody perennials in northeastern USA. *International Journal of Biometeorology, 49*(5), 303-309.

Wu, X. J., Zurita-Milla, R., & Kraak, M. J. (2013). Visual Discovery of Synchronisation in Weather Data at Multiple Temporal Resolutions. *Cartographic Journal, 50*(3), 247-256. doi:

Yanenko, O., & Schlieder, C. (2012). Enhancing the Quality of Volunteered Geographic Information: A Constraint-Based Approach. In J. Gensel, D. Josselin & D. Vandenbroucke (Eds.), *Bridging the Geographic Information Sciences* (pp. 429-446): Springer Berlin Heidelberg.

Zurita-Milla, R., van Gijsel, J. A. E., Hamm, N. A. S., Augustijn, P. W. M., & Vrieling, A. (2013). Exploring Spatiotemporal Phenological Patterns and Trajectories Using Self-Organizing Maps. *Ieee Transactions on Geoscience and Remote Sensing, 51*(4), 1914-1921.

APPENDICES

Appendix A

The results of Moran's I test for first flowering observations. The test was applied to the observations of each year separately.

Year	Moran's I Index	z-score	p-value
2003	0.097	1.739	0.082
2002	0.157	3.032	0.002
2001	0.125	2.495	0.012
2000	0.117	2.526	0.011
1999	0.121	2.864	0.004
1998	0.058	1.668	0.095
1997	0.097	2.781	0.005
1996	0.172	4.673	0.000
1995	0.140	2.900	0.003
1994	0.433	9.040	0.000
1993	0.342	11.267	0.000
1992	0.555	17.116	0.000
1991	0.427	13.580	0.000
1990	0.540	15.722	0.000
1989	0.470	14.708	0.000
1988	0.296	12.612	0.000
1987	0.242	13.914	0.000
1986	0.284	18.634	0.000
1985	0.364	20.233	0.000
1984	0.356	29.584	0.000
1983	0.291	25.870	0.000
1982	0.333	18.623	0.000
1981	0.297	18.837	0.000
1980	0.349	25.222	0.000
1979	0.321	23.547	0.000
1978	0.429	32.190	0.000
1977	0.279	20.981	0.000
1976	0.252	30.334	0.000
1975	0.314	39.201	0.000
1974	0.387	28.645	0.000
1973	0.513	44.012	0.000
1972	0.674	51.474	0.000
1971	0.482	57.120	0.000
1970	0.662	58.488	0.000
1969	0.284	18.634	0.000

Year	Moran's I Index	z-score	p-value
1968	0.492	56.269	0.000
1967	0.781	67.834	0.000
1966	0.537	52.127	0.000
1965	0.679	70.971	0.000
1964	0.692	80.541	0.000
1963	0.575	47.670	0.000
1962	0.514	63.282	0.000
1961	0.622	91.735	0.000
1960	0.656	40.818	0.000
1959	0.831	52.591	0.000
1958	0.697	45.055	0.000
1957	0.622	40.624	0.000
1956	0.035	0.810	0.418

Appendix B

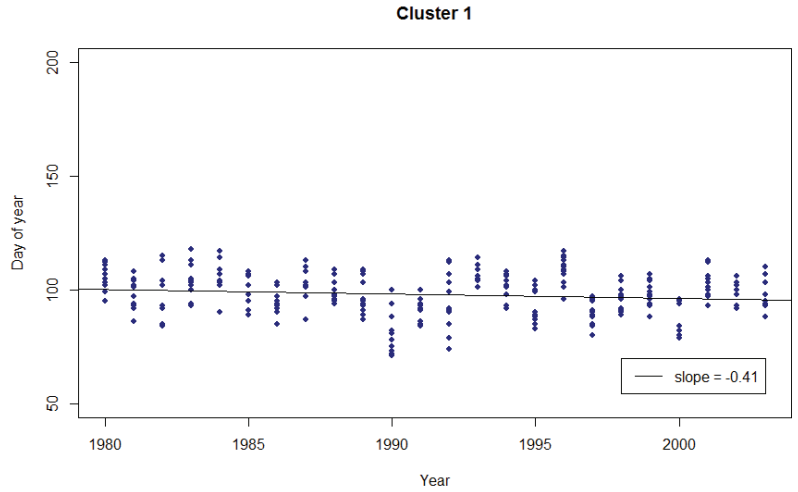
The results of the regression analysis for the stations that have 30 years and more of first flower observations, in which the day of first flower observations was the dependent variable and the year was the independent variable.

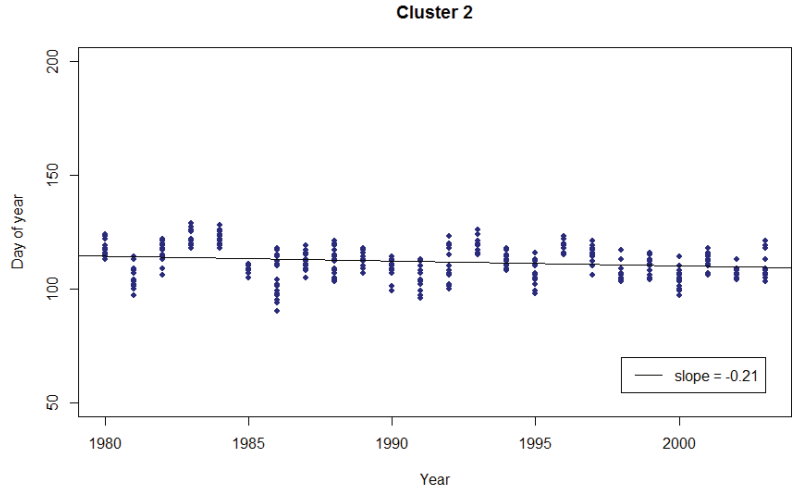
Station	Latitude	Longitude	Number of years	Start	End	Regression Slope
40738	40.45	-123.15	32	1957	1992	-0.22
49122	39.09	-123.12	30	1957	1986	-0.96
51528	39.13	-105.17	30	1957	1992	-0.12
53246	38.40	-108.59	35	1957	1993	-0.30
56559	37.21	-106.30	40	1957	2003	-0.25
65445	41.97	-73.22	35	1967	2003	-0.23
190666	42.63	-72.12	32	1969	2002	-0.15
190998	42.12	-71.90	31	1969	2003	-0.35
194246	42.27	-72.88	30	1969	2003	-0.29
213303	47.23	-93.50	30	1973	2002	0.05
241044	45.40	-111.03	38	1956	1993	-0.33
241408	45.12	-111.41	31	1656	1986	0.20
243013	46.51	-108.19	31	1957	2001	-0.55
244084	48.05	-116.00	39	1956	2003	-0.09
244345	45.55	-108.15	31	1956	1993	-0.20
244506	45.29	-108.58	36	1956	1993	-0.34
244715	45.55	-104.05	33	1956	1993	-0.12
245106	45.18	-107.22	34	1956	1989	-0.45
245285	48.27	-105.56	34	1956	1992	-0.52
245387	46.30	-110.20	35	1956	2001	-0.02
245572	48.29	-104.27	30	1957	1988	-0.10
245761	47.03	-109.57	32	1956	1992	-0.22
246700	47.39	-111.36	35	1956	2003	-0.13
272174	43.13	-70.93	34	1966	2003	-0.24
294009	32.56	-107.34	31	1957	1993	0.09
294089	33.24	-105.18	34	1957	1993	-0.21
296659	32.39	-103.23	32	1957	1993	-0.20
300889	40.95	-72.30	34	1967	2003	-0.46
301401	44.88	-73.47	34	1968	2003	-0.29
306774	41.38	-74.68	36	1967	2003	-0.27
322188	46.88	-102.80	30	1962	1992	-0.60
339312	40.78	-81.92	40	1962	2003	-0.14
354147	45.34	-116.50	30	1957	1987	-0.23
421588	40.55	-111.24	36	1957	1993	-0.09
422798	39.05	-111.08	30	1957	2003	-0.45
423046	38.17	-111.16	30	1957	1993	0.09
426357	39.23	-112.20	31	1957	1994	-0.43

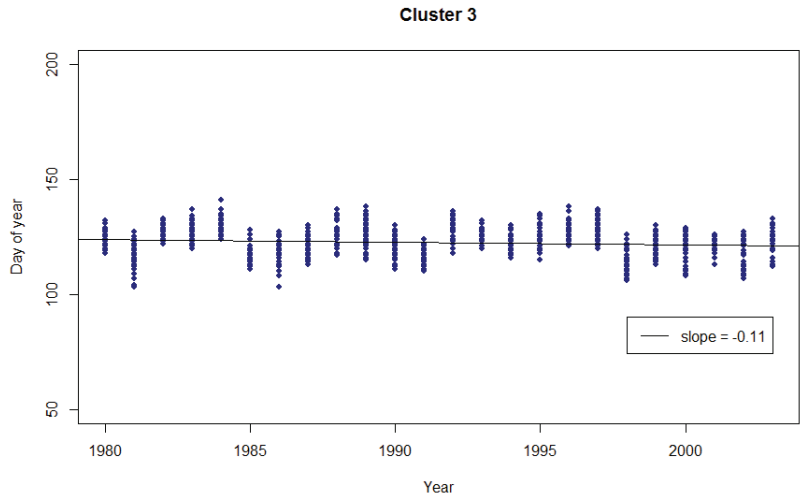
Station	Latitude	Longitude	Number of years	Start	End	Regression Slope
431243	43.38	-72.60	37	1965	2003	-0.18
432843	44.52	-73.12	37	1965	2003	-0.19
438556	43.80	-72.27	34	1966	2003	-0.20
456624	48.07	-123.26	30	1957	2003	-0.47
456974	48.39	-118.44	35	1957	1993	-0.19
489207	43.56	-104.46	32	1957	1994	-0.33

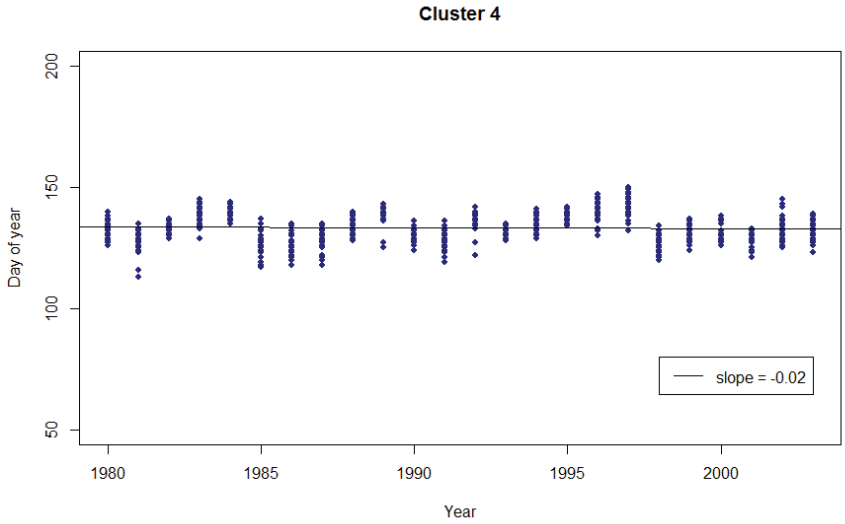
Appendix C

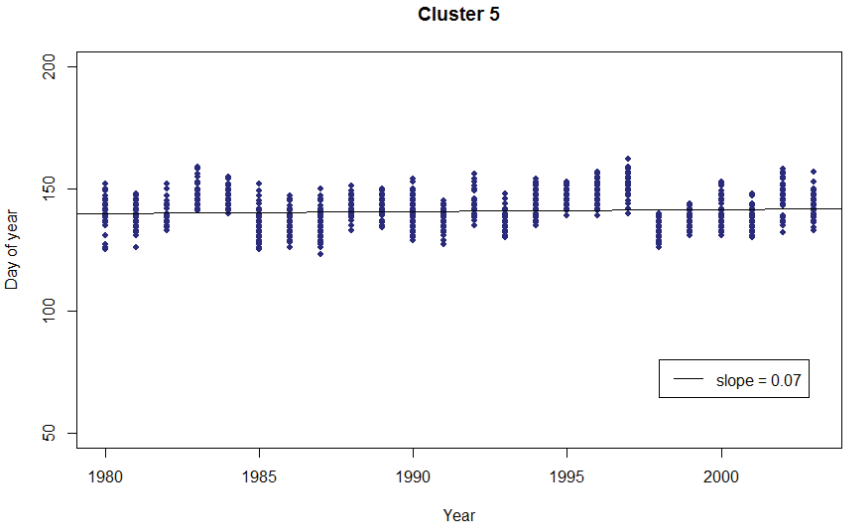
Plots for SOM clusters, the first flower values of each cluster were plotted versus the year.

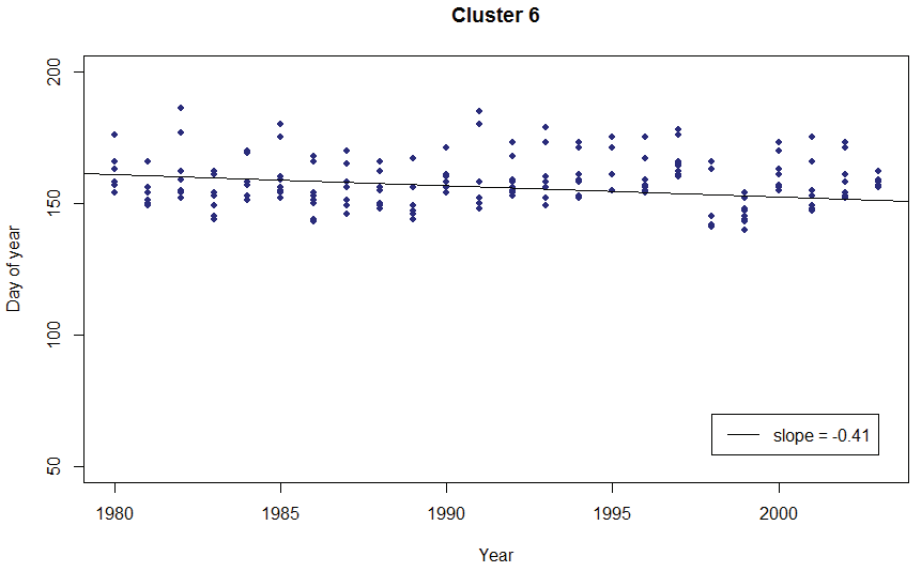







Cluster 4
Day of year
Year
slope = -0.02
50
100
150
200
1980
1985
1990
1995
2000


Cluster 5
Day of year
Year
slope = 0.07
50
100
150
200
1980
1985
1990
1995
2000


Cluster 6
Day of year
Year
slope = -0.41
50
100
150
200
1980
1985
1990
1995
2000