# Integrating time series mining and fractals to discover patterns and extreme events in climate and remote sensing databases

***Luciana Alvim Santos Romani***

# Integrating time series mining and fractals to discover patterns and extreme events in climate and remote sensing databases [1]

***Luciana Alvim Santos Romani***

**Advisor:** ***Profa. Dra. Agma Juci Machado Traina***

Thesis presented to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, as part of the doctoral exam requisites of the Doctorate Program in Computer Science and Computational Mathematics.

**USP – São Carlos**
**December/2010**

[1] Sponsorship - Embrapa
Additional Support - FAPESP, CNPq, CAPES and Microsoft Research

*To my husband Roberto,*
*and my dear sons Lucas and Rafael.*

*I want to know how God created this world. I am not interested in this or that phenomenon, in the spectrum of this or that element. I want to know His thoughts; the rest are details.*

ALBERT EINSTEIN

# Acknowledgments

*I would like to publicly acknowledge those who contributed and supported me throughout this journey. First of all, I thank the Lord for giving me all the conditions necessary to achieve one of the great dreams of my life putting hope in my heart, clarity in my mind and mostly wonderful people who aided me during those four years of hard work.*

*I would like to thank my dear husband Roberto and my lovely sons Lucas and Rafael by their affection and fortitude during difficult times, by their comprehension during my absences, by the incentive during the hours that discouragement arose and for making my life much happier and fulfilled. I consider myself a lucky person to have been born in a wonderful family. My parents Edson and Benedita taught me to respect the world and people, to never give up my dreams and to have the courage to face all obstacles. I thank them and my sister Cris for their attention, encouragement words, for hugging and cheering me all the time.*

*My special thanks to my advisor, Prof. Agma, who guided me during these past years with dedication, wisdom and affection. I learned a lot observing her manner to talk to people, to explain her ideas and to disseminate her knowledge through numerous published papers. Prof. Agma is a kind of special person with a rare goodness who is very difficult to meet nowadays. I feel honored to have been her student and I am grateful that she made these four years milder and memorable.*

*My gratitude to Prof. Caetano and Prof. Elaine at ICMC - USP by the time and effort invested on my thesis. I also thank by ideas, collaborative work, the reviews of my papers and in particular for introducing me to the fractal theory that made me to immerse in a world of fabulous discoveries and knowledge.*

*I dedicate my sincere thanks to Marcela, Mônica and Silvana due to their contributions to my work, by the several reviews and also by incentive words in moments of despair and anguish. I have very much enjoyed being a member of the Database and Image Group (GBdI) of USP in São Carlos, especially I am grateful to my colleagues Junior, André, Humberto, Camila, Ana Lúcia, Pedro, Carol, Daniel Kaster, Robson, Sergio, Marcelo, Willian and Letricia. I am also grateful to Daniel Chino, Santiago and Bruno, who I had the pleasure to help in their projects of Undergraduate Research, by their help implementing many ideas that were motivated by this thesis. My gratitude is also to the staff and Professors of ICMC – USP by attention and politeness they always had to answer my requests.*

*My thanks to the group of specialists in Meteorology and Remote Sensing as Ana Avila, Jurandir, Renata, Priscilla and Marina who evaluated my work, made interesting suggestions and validated the methods proposed in this thesis. I would also like to thank the staff of Cepagri - Unicamp and my colleagues for welcoming me so warmly over the last four years.*

*Special acknowledgment is due to my workmates Silvio (academic advisor), Eduardo Assad, Fernanda, Stanley, Adriano, João Francisco, Julio, Marcos, Nanci, Goretti, and Carla Osawa who helped me in different situations with great attention and goodwill.*

*I also acknowledge Embrapa (Brazilian Agricultural Research Corporation) for the financial support and for the four-year leave of absence to improve my education. I also thank the funding agencies FAPESP, Microsoft Research, CNPq and CAPES that support the research undergoing at the GBdI laboratory.*

# Abstract

This thesis presents new methods based on fractal theory and data mining techniques to support agricultural monitoring in regional scale, specifically regions with sugar cane fields. This commodity greatly contributes to the Brazilian economy since it is a viable alternative to replace fossil fuels. Since climate influences the national agricultural production, researchers use climate data associated to agrometeorological indexes, and recently they also employed data from satellites to support decision making processes. In this context, we proposed a method that uses the fractal dimension to identify trend changes in climate series jointly with a statistical analysis module to define which attributes are responsible for the behavior alteration in the series.

Moreover, we also proposed two methods of similarity measure to allow comparisons among different agricultural regions represented by multiples variables from meteorological data and remote sensing images. Given the importance of studying the extreme weather events, which could increase in intensity, duration and frequency according to different scenarios indicated by climate forecasting models, we proposed the CLIPSMiner algorithm to identify relevant patterns and extremes in climate series. CLIPSMiner also detects correlations among multiple time series considering time lag and finds patterns according to parameters, which can be calibrated by the users.

We applied two distinct approaches in order to discover association patterns on time series. The first one is the Apriori-FD method that integrates an algorithm to perform attribute selection through applying the correlation fractal dimension, an algorithm of discretization to convert continuous values of series into discrete intervals, and a well-known association rules algorithm (Apriori). Although Apriori-FD has identified interesting patterns related to temperature, this method failed to appropriately deal with time lag. As a solution, we proposed CLEARMiner that is an unsupervised algorithm in order to mine the association patterns in one time series relating them to patterns in other series considering the possibility of time lag.

The proposed methods were compared with similar techniques as well as assessed by a group of meteorologists, and specialists in agrometeorology and remote sensing. The experiments showed that applying data mining techniques and fractal theory can contribute to improve the analyses of agrometeorological and satellite data. These new techniques can aid researchers in their work on decision making and become important tools to support decision making in agribusiness.

# *Resumo*

*Esta tese apresenta novos métodos baseados na teoria dos fractais e em técnicas de mineração de dados para dar suporte ao monitoramento agrícola em escala regional, mais especificamente áreas com plantações de cana-de-açúcar que tem um papel importante na economia brasileira como uma alternativa viável para a substituição de combustíveis fósseis. Uma vez que o clima tem um grande impacto na agricultura, os agrometeorologistas utilizam dados climáticos associados a índices agrometeorológicos e mais recentemente dados provenientes de satélites para apoiar a tomada de decisão. Neste sentido, foi proposto um método que utiliza a dimensão fractal para identificar mudanças de tendências nas séries climáticas juntamente com um módulo de análise estatística para definir quais atributos são responsáveis por essas alterações de comportamento.*

*Além disso, foram propostos dois métodos de medidas de similaridade para auxiliar na comparação de diferentes regiões agrícolas representadas por múltiplas variáveis provenientes de dados meteorológicos e imagens de sensoriamento remoto. Diante da importância de se estudar os extremos climáticos que podem se intensificar dado os cenários que preveem mudanças globais no clima, foi proposto o algoritmo CLIPSMiner que identifica padrões relevantes e extremos em séries climáticas. CLIPSMiner também permite a identificação de correlação de múltiplas séries considerando defasagem de tempo e encontra padrões de acordo com parâmetros que podem ser calibrados pelos usuários.*

*A busca por padrões de associação entre séries foi alcançada por meio de duas abordagens distintas. A primeira delas integrou o cálculo da correlação de dimensão fractal com uma técnica para tornar os valores contínuos das séries em intervalos discretos e um algoritmo de regras de associação gerando o método Apriori-FD. Embora tenha identificado padrões interessantes em relação à temperatura, este método não conseguiu lidar de forma apropriada com defasagem temporal. Foi proposto então o algoritmo CLEARMiner que de forma não-supervisionada minera padrões em uma série associando-os a padrões em outras séries considerando a possibilidade de defasagem temporal.*

*Os métodos propostos foram comparados a técnicas similares e avaliados por um grupo composto por meteorologistas, agrometeorologistas e especialistas em sensoriamento remoto. Os experimentos realizados mostraram que a aplicação de técnicas de mineração de dados e fractais contribui para melhorar a análise dos dados agrometeorológicos e de satélite auxiliando no trabalho de pesquisadores, além de se configurar como uma ferramenta importante para apoiar a tomada de decisão no agronegócio.*

# Contents

# List of Figures

# List of Tables

# List of Symbols

| | |
|---|---|
| $A = a_1, a_2, \ldots, a_E$ | Definition of dataset $A$, composed of attributes $a_i$ |
| $C_{r,i}$ | Count of points in the i-th grid cell of side size $r$ |
| $D$ | Fractal dimension |
| $D_2$ | Correlation fractal dimension (for finite fractal sets) |
| $E$ | Embedding dimension |
| $iC()$ | Maximum individual contribution of an attribute |
| $M_j(C) \rightarrow a_j$ | Mapping of attributes $C \subset A$ restricting the values of $a_j$ |
| $pD()$ | Partial Intrinsic Dimension |
| $r$ | The side of the cells in a (hyper) cubic grid |
| $\mathbb{R}$ | Domain of real numbers |
| $\xi$ | Strength threshold of correlations to be retrieved |
| $\xi C$ | Attribute set core |
| $\xi B_p$ | Correlation base of a correlation group $p$ |
| $\xi G_p$ | Correlation group $p$ |
| $S$ | Time series |
| $v_i$ | Value in a time series |
| $t_i$ | Time value in a time series |
| $e_i$ | Events of type $(v_i, t_i)$ |
| $d_i$ | Difference of events $d_i = v_{i+1} - v_i$ |
| $S_e$ | Event sequence |
| $S_{ea}$ | Ascending event sequence |
| $S_{ed}$ | Descending event sequence |
| $S_{es}$ | Stable event sequence |
| $V$ | Pattern of type Valley (negative peak) |
| $M$ | Pattern of type Mountain (positive peak) |
| $P$ | Pattern of type Plateau (small variation) |
| $y$ | Time series amplitude |
| $\delta$ | Minimum variation between two consecutive events |
| $\rho$ | Relevance Factor |
| $\lambda$ | Plateau Length |
| $\tau$ | Time delay |
| $n$ | Number of elements in a time series |
| $p$ | Number of time windows (pieces of time series) |
| $q$ | Minimum number of events in an event sequence |

# Glossary of Terms

**APCA** *Adaptive Piecewise Constant Approximation.*, 52
**AVHRR** *Advanced Very High Resolution Radiometer*, 24

**CCAR** *Colorado Center for Astrodynamics Research*, 25

**DFT** *Discrete Fourier Transform.*, 50
**DM** *Data Mining is the central stage in the KDD process, where computational techniques are applied to extract unknown and useful patterns from the data.*, 43
**DTW** *Dynamic Time Warping.*, 62
**DWT** *Discrete Wavelet Transform.*, 50

**ENSO** *El Niño Southern Oscillation*, 16
**ETM** *Maximun Evapotranspiration*, 19
**ETP** *Potential Evapotranspiration*, 19
**ETR** *Real Evapotranspiration*, 19
**EVI** *Enhanced Vegetation Index*, 23

**GSP** *Generalized Sequential Patterns.*, 59

**HRPT** *High Resolution Picture Transmission*, 24

**IPCC** *Intergovernmental Panel on Climate Change*, 17

**Kc** *Crop coefficient*, 19
**KDD** *Knowledge discovery in databases is an interactive sequence of steps with the purpose of discovering useful information and knowledge from data.*, 42

**LAI** *Leaf Area Index*, 19

**MCC** *Maximum Cross Correlation*, 26
**MODIS** *Moderate Resolution Imaging Spectroradiometer*, 23
**MVC** *Maximum Value Composite*, 27

**MVQ** *Multi-resolution Vector Quantized.*, 52

**Nadir** *Intersection point between vertical line that is perpendicular to the horizontal plane and the celestial sphere, but in the opposite hemisphere to that one where observer is located.*, 24

**NAVPRO** *A set of scripts that call the subroutines of NAV system*, 25

**NDVI** *Normalized Difference Vegetation Index*, 23

**NOAA** *National Oceanic and Atmospheric Administration*, 24

**OMEGA** *Supervised algorithm to feature selection and discretization.*, 47

**PAA** *Piecewise Aggregate Approximation.*, 51

**PLA** *Piecewise Linear Approximation.*, 50

**SAX** *Symbolic Aggregate approXimation.*, 52

**SPADE** *Sequential Pattern Discovery using Equivalence classes.*, 59

**SPIRIT** *Sequential Pattern Mining Regular Expression Constraints.*, 60

**SST** *Sea Surface Temperature*, 16

**SVD** *Singular Value Decomposition.*, 50

**WMO** *World Meteorological Organization*, 13

**WRSI** *Water Requirements Satisfaction Index*, 20

# Chapter 1

# Introduction

According to official data (IBGE, 2007), the Brazilian agribusiness segment contributed with 23.3% of the national Gross Domestic Product (GDP), 42% of exports and 37% of jobs in 2007. Significant advances have been made in determining suitable areas for agricultural crop development through agricultural zoning program, undertaken by the Brazilian Ministry of Agriculture (Rosseti, 2001).

The Brazilian agricultural zoning program aims at reducing agriculture losses caused by two climatic-associated risks: dry spells during the reproductive stage and excessive rainfall during the harvesting periods. This official program defines planting calendars for the main crops in the country, which have been calculated to achieve risk rates lower than 20% regarding climate problems, based on climate data and agrometeorological methods.

Since agriculture is affected by the climate conditions, it is strategic to the governments to be able to forecast climate trends in order to generate or modify public policies and to act when necessary to reduce negative impacts on the economy. In this context, the scientific community has a particular concern in relation to the climate since researchers in whole world have no doubt about global warming, being necessary to understand the main causes of this phenomenon and its consequences for Earth's ecosystems. Since the global warming affects the whole planet, the Intergovernmental Panel on Climate Change (IPCC) was created, aimed at evaluating and analyzing the data concerning such changes and proposing ways to deal with the issues derived from the climate changes.

One of the causes for the global warming is the increase of greenhouse gases emissions. Thus, one alternative to mitigate this problem is the replacement of fossil fuels with renewable energy sources, such as ethanol. In Brazil, sugar cane is the main agricultural crop used to produce ethanol. The country has a privileged position to support the growing international demand for sugar and anhydrous ethanol for fuel. With two main producing regions and alternate crops, Brazil is able to maintain its worldwide market presence throughout the year. In fact, this agricultural commodity has a strategic

importance to the national economy. Therefore, there is an evident need for accurate crop prediction techniques that would help the production planning and the marketing strategy for domestic and foreign markets.

In this context, remote sensing data can be used to improve traditional agrometeorological methods for harvest monitoring or forecasting. Nowadays, these data are more accessible and there are appropriate technology (software and hardware) to receive, distribute, manipulate and process long time series of satellite images. Remote sensing images have contributed to advances in several areas, such as Geography, Meteorology and more recently Agriculture. These technologies can improve agrometeorological monitoring enabling a reduction in losses of agricultural crops by extreme weather events, such as drought and frost.

Reliable estimates of agricultural production are powerful tools to guide producers on issues related to planting and also to assist agribusiness in operating and marketing sectors. They may generate trustworthy data to support the government in the decision making process, aimed at reducing negative impacts on the economy or to take advantage of favorable situations in the climate and in the agricultural market.

Considering this scenario, the development of computational models to filter, transform, merge and analyze data from many different areas is complex and challenging. The complexity increases whenever it is necessary to combine several climatic and agrometeorological variables, and moreover when using climate and agriculture models together. Therefore, the application of statistical models, as well as the development of new computational methods become very important to aid in the analyses of climate data from ground-based stations, and outputs of forecasting models and remote sensing imagery.

## 1.1 Motivation

Recently, improvements in data collection methods and sensor technology have provided an increase in the acquisition of spatial data. Consequently, data related to agrometeorology gathered from ground-based stations and those obtained from remote sensing imagery have increased fast and continuously. This huge amount of data has been stored by several research institutions and universities, once Climatology and Agrometeorology research need historical series at least 30 years long (Zhai et al., 2005).

Historical series of agrometeorological data from several private and governmental institutions in Brazil have been integrated in the Agritempo system [1] developed in partnership between Embrapa Agriculture Informatics and the Center for Meteorological and

[1] www.agritempo.gov.br

Climatic Research Applied to Agriculture (Cepagri) at the University of Campinas (Unicamp) since 2001. Agritempo organizes and manages over 1,000 ground-based stations (mechanical and automatic).

Although the meteorological stations network in the country is continuously increasing, it is already insufficient to gather information from all regions of Brazil. Moreover, the accurate information about agricultural commodities is strategic to an organized economy and to resources management of a nation. In order to obtain this valuable information, the process of crop yield estimation should be precise and objective. In Brazil, the responsible companies for crop season forecasting are CONAB[2] and IBGE[3] that still use conventional methods based on surveys with producers, cooperatives, seed and fertilizer suppliers and other productive chain sectors. These surveys are applied six times a year in several Brazilian regions. However, in order to be more accurate about the extension of agricultural crops, it should be important to carry out a crop inventory more frequently. Nevertheless, these periodic inventories are not viable because of the great extension of Brazilian's territory.

The equation of crop season forecasting system is relatively simple, since it consists on multiplying the estimation of cultivated area by the average productivity expected for the crop. Thus, the difficulty of this equation is to precisely determine the information regarding two variables: cultivated area and productivity. There are crops with several varieties and kinds of management in the country. Moreover, there are many kinds of soil and pluviometric regimes that lead to productivity levels completely different for each region in Brazil.

These facts strengthen the importance of applying remote sensing in assessment, monitoring and forecasting of agriculture crops in the country. In this context, Cepagri/Unicamp has stored remote sensing images from NOAA-AVHRR since April, 1995. At present, there are over six terabytes of images, which have been used in Agrometeorology and Remote Sensing research. In general, information extracted from remote sensing imagery combined with climate data can reveal useful information, which can help researchers to monitor and to estimate the production of agricultural crops.

Employing information obtained by remote sensing images makes the process more robust and trustful, but also more time consuming. Meteorologists and agrometeorologists use well-known statistical methods, such as principal component analysis, cluster analysis, frequency distribution, geostatistics, Fourier transform, non-parametric statistics and so on, for analyzing and finding patterns in Earth science. They are interested in defining the climatic behavior of each region in order to identify their anomalies, such as long periods of

[2]http://www.conab.gov.br/
[3]http://www.ibge.gov.br/

drought (days with rainfall bellow 10 mm or without rainfall), days with extreme rainfall, periods of drought in the Winter associated to high temperatures and other phenomena. Considering all these aspects, the necessity of developing methods and techniques to aid agrometeorologists to analyze and to extract relationships and patterns from this large amount of climate data, as well as remote sensing images and outputs of climate forecasting models is evident.

In fact, the diversity, complexity and volume of data to be processed bring interesting challenges to the computer science field. For example, data mining techniques can be employed to extract relevant patterns and correlations from climate data (Ganguly & Steinhaeuser, 2008). Outliers and exception analyses (Luo et al., 2008) can help finding critical points in the climate changes. Temporal association rules (Yoo & Shekhar, 2009) are important tools in data mining and can be employed to relate climate data to interpretation from the meteorologist. Beyond that, visual data mining (Simoff et al., 2008) is a valuable resource to explore and to help understanding the data.

Finally, as the doctorate candidate is a researcher with Embrapa Agriculture Informatics, which is partner of Cepagri/Unicamp, the access to climate and remote sensing databases was granted. According to all aspects described in this section, we define the goals of this doctorate thesis in the next section.

## 1.2 Goals

The main goal of this thesis is the definition and development of methods to support agrometeorologists in tasks related to agricultural monitoring and research on climate change. The hypothesis is that the development of methods based on fractal theory and time series mining allows the correlation analysis and knowledge discovery from time series of climate data and remote sensing images.

A numerous quantity of time series is available in real agroclimatic applications. Additionally, climate data distribution usually changes over time. Thus, a specific goal of this work was the proposition of methods to track the behavior of evolving data, pointing out the attributes that are responsible for trend changes.

Agricultural crop fields can be characterized by different variables involving physiologic aspects of plants, agrometeorological indexes, climatic conditions and vegetation indexes obtained from operations between satellite channels. Thus, another specific goal of this thesis was the development of methods to similarity search of agricultural crop fields considering them as multidimensional objects.

Finally, the most important objective of this work was the definition of methods to

detect relevant patterns and extreme events in time series and association rules between heterogeneous time series of climate data and remote sensing images.

## 1.3 Challenges

A challenge in the time series mining field is the development of well-suited techniques to mine patterns in time series of continuous values without losing the information about the time of occurrences. Specifically, quantizing time series to retain the temporal meaning of the patterns is a problem to be considered.

Other great challenge for data miners is the mining process of heterogeneous time series. When we have to consider time delay to discover association patterns between heterogeneous time series, the problem becomes even more complex. However, this is an important issue to be tackled once the influence of meteorological events, such as occurrence of rainfall or not, during the growing of plants happens after a certain period of time.

One of the main effects of climate change is the increasing in the frequency, duration and intensity of extreme phenomena (IPCC, 2007). Thus, another challenge is how to detect these extremes with enough details about the conditions of occurrence analyzing historical series of climate data.

Summarizing, some questions must be answered to overcome these challenges:

- How to discover higher and extreme events in time series maintaining the semantic information?
- How to detect correlation between time series considering time delay?
- How to compare regions represented as multidimensional objects composed of heterogeneous time series?
- How to associate heterogeneous time series of continuous data in order to discover new patterns retaining temporal meaning?

To deal with these problems we proposed the use of the fractal theory associated to time series mining techniques, which is not commonly employed in Climatology, Agrometeorology and Remote Sensing fields.

## 1.4 Contributions

This thesis brings contributions to different areas of knowledge: time series mining, climatology, agrometeorology and remote sensing. The major contributions are:

- A new method to mine association rules from heterogeneous time series (Apriori-FD). The Apriori-FD method combines techniques of feature selection, discretization and association rules to discover patterns and knowledge from climate data and remote sensing images. This method was applied on datasets of sugar cane crop fields and detected interesting rules related to temperature conditions appropriate to the best sugar cane production, which varies from one region to another (Romani et al., 2008).

- A data stream monitoring method combining a fractal-based approach with a statistical analysis module to monitor data highlighting trend changes and the attributes that influence the changes. Results showed that monitoring the fractal dimension is a suitable technique to monitor climate and remote sensing streams spotting the regions of interest making the agrometeorologists' work easier (Romani et al., 2009a; Nunes et al., 2010).

- Two new similarity measures for multidimensional objects composed of heterogeneous time series. Both methods weight the well-known DTW distance function with a correlation factor. The CV-DTW uses correlation between two variables such as Pearson's correlation and FD-DTW takes advantage of correlation fractal dimension. Experiments showed that the methods are appropriate to compare regions of sugar cane fields (Romani et al., 2009d, 2010a).

- A new unsupervised algorithm (CLIPSMiner) for discovering relevant and extreme patterns in heterogeneous climate and remote sensing time series of continuous data. The algorithm also discovers high and extreme phenomena according to parameters tuned by users in entire time series or by time windows, which allows comparisons between different periods of time. Results showed that CLIPSMiner finds climatic episodes along the historical series retaining the semantic of the data (Romani et al., 2009c, 2010d). Experiments with real data reached results according to climatologists' expectations (Romani et al., 2009b).

- A new unsupervised algorithm (CLEARMiner) to mine rules that associate patterns in a time series to patterns in other series considering time lag. CLEARMiner first converts time series into a symbolic representation and afterwards discovers association patterns between series. The algorithm considers a time-window constraint to reduce the search space. Results indicated that CLEARMiner finds rules known and unknown by specialists and can be used to support decision making processes (Romani et al., 2010c).

## 1.5 Organization of this work

This thesis is organized in 9 chapters as follows:

- Chapter 1 presented the introduction, motivation and objectives of this thesis.
- Chapter 2 describes concepts related to Agrometeorology and Remote Sensing that are important to better comprehend the problems tackled in this work. The chapter presents the geo-processing process employed to correct distortions in satellite images and the aspects concerning sugar cane crops.
- Chapter 3 details the fractal theory that is used to detect correlation between climate and remote sensing data and to support some methods proposed in this thesis.
- Chapter 4 reviews concepts regarding data mining and time series mining. The chapter also presents the main data mining tasks highlighting association rules, which is the focus of this thesis. Moreover, this chapter briefly introduces some similarity measures.
- Chapter 5 shows how we employ the fractal theory in time series analysis. This chapter details the Apriori-FD method, which combines the correlation fractal dimension and the Apriori algorithm to association rules mining. In addition, this chapter presents a data stream monitoring method.
- Chapter 6 reports details about two similarity measures developed to deal with multidimensional objects, weighting the well-known DTW distance function by correlation factors.
- Chapter 7 describes the CLIPSMiner, which is an algorithm proposed to detect relevant and extreme events in time series.
- Chapter 8 details the CLEARMiner algorithm that was proposed to find association patterns between time series, which was the main contribution of this thesis.
- Finally, chapter 9 presents conclusions, the major contributions of this thesis and ideas for further works.

There are also two appendixes that describe the *SatImagExplorer* system and an improvement in the fractal-based analysis method described in chapter 5. Both implementations were performed by undergraduate students on the scope of this doctorate project.

# Part I

# Concepts and Related Work

# Chapter 2

# Agrometeorology and Remote Sensing

## 2.1 Introduction

The branch of science that studies physical phenomena of the atmosphere called meteors (in Greek) is defined as *Meteorology*. Specifically, meteorologists study atmospheric conditions in a given time. These conditions result from the air motion that is originated from the spatial variation of forces acting on the air mass. Another important aspect of this atmospheric movement is its statistical description in terms of sequential average values. Based on these values, meteorologists can describe annual rate of atmospheric phenomena occurrence. This average sequencing defines the climate of a given region and determines which activities can be performed in that place. Thus, this average characterization defines the *Climatology*. Therefore, Meteorology works with instantaneous values while Climatology uses average values (of long periods).

Meteorology is divided in different specialized parts with specific objectives. One of them is denominated *Agrometeorology*, which is devoted to the atmospheric conditions and its consequences in the rural environment. The climatic conditions indicate the most viable agricultural activity to a given region and the meteorological conditions determine the productivity level for that activity in a certain period (Pereira et al., 2002).

Agrometeorologists defined a planting calendar through the balance between rainfall and evapotranspiration, which depends on surface conditions (land use and soil) and atmospheric demand (energy availability, air humidity and wind speed). One of the most important crops in Brazil is the sugar cane, which is used to produce sugar, ethanol, and energy. Brazil occupies the top position in the world ranking of the sugar cane production and has an important role to attend the world's sugar and ethanol needs. Then, this

agricultural commodity is strategic to the economy of the country. The sugar cane crops are cultivated in large and contiguous fields, which allows the use of low resolution satellite images.

Thus, remote sensing data can be an alternative to more conventional methods, because the sensors have an excellent spatial and temporal coverage. These sensors also make it possible to obtain continuous information from the country land, with spatial resolution of a few kilometers and temporal data in order of minutes. However, measurements obtained from remote sensors are indirect and, therefore, it is necessary to develop models that relate the features available in the satellite spectral channels to parameters associated with the required information.

In this scenario, several satellites are being used to assist in land monitoring and climate forecasting. In this chapter, some concepts of Agrometeorology, Remote Sensing and Sugar Cane crops are presented to provide a theoretical foundation in the application field of this work. Thus, climate and weather are defined in Section 2.2. Sugar cane characteristics and related work are presented in Section 2.3. Basic concepts of remote sensing are defined in Section 2.4, as well as vegetation indexes and the sensor and the satellite used in this thesis. Finally, some works about remote sensing images for sugar cane monitoring are discussed in Section 2.4.3.

## 2.2 Weather and Climate

Atmosphere is a mass in continuous movement that induces variations in the meteorological conditions predominant in a region. The description of atmosphere status can be both in terms of instant (current condition) or statistical (average condition). It introduces a time scale to describe the atmospheric conditions. Thus, weather refers to instantaneous description and climate is related to average description.

**Definition 2.1** ***Weather*** *is the state of the atmosphere, mainly with respect to its effects upon life and human activities. As distinguished from climate, weather consists of the short-term (minutes to days) variations in the atmosphere. Popularly, weather is thought in terms of temperature, humidity, precipitation, cloudiness, visibility, and wind (Pereira et al., 2002; Vianello & Alves, 1991).*

**Definition 2.2** ***Climate*** *is a statistical description that expresses the average conditions (30 years or more) of the sequencing of time at a place, i.e. the slowly varying aspects of the atmosphere–hydrosphere–land surface system. The pace of seasonal variations in temperature, rainfall, humidity, etc. characterizes the climate of a region (Pereira et al., 2002; Vianello & Alves, 1991).*

Based on climatic descriptions, it is possible to know in advance what weather conditions are predominant in the region, and hence what agricultural activities have a higher chance of success. According to the World Meteorological Organization (WMO), a uniform period of at least three consecutive ten-years is called *Climatological Normal*. This 30-year period is defined by WMO as sufficient and satisfactory to cancel outliers in the series, but must also take into consideration the ever-changing climate. WMO advises choosing periods such as 1901-1930, 1931-1960, 1961-1990, that define the succession of Climatological Normal patterns to make comparisons among stations of data gathering, regions and periods of 30 years (Zhai et al., 2005).

Figure 2.1 presents the annual variation of air temperature and rainfall for the São Paulo state. In fact, one year similar to the Climatological Normal probably never occurred. However, the Normal represents the most likely climate conditions in a given region. As it can be seen in Figure 2.1(a), temperature of São Paulo state varies between a minimum 11°C in July and maximum 28°C in February. Analyzing the graph of rainfall (Figure 2.1(b)), it can be observed that the majority of the annual rainfall occurs during Spring and Summer in the southeast of Brazil. During the Winter, months are less rainy. Accordingly, the climate of São Paulo state is characterized by rainy and warm Summers, while Winters are mild and dry.

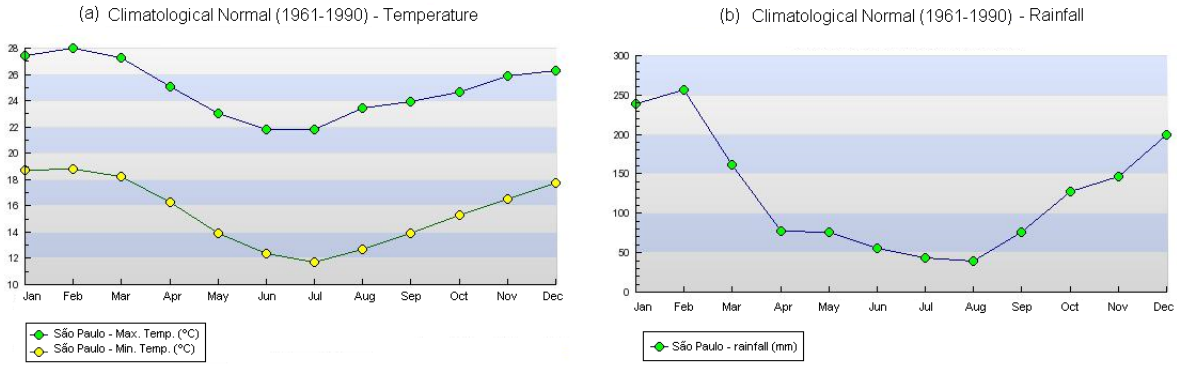



Figure 2.1: Graphs of the last Climatological Normal for São Paulo state corresponding to the period from 1961 to 1990. (a) Air temperature varying from 11°C to 29°C (b) Rainfall distribution along the year (adapted from National Institute of Meteorology - INMET).

One region has a daily variation in the meteorological conditions (rainfall, temperature, humidity, etc.) due to the Earth's rotation movement. This variation is a natural phenomenon that occurs in all locations, with greater or lesser intensity. There is also annual variation of meteorological conditions that generates the seasons owing to the yearly revolution of the Earth around the Sun and the tilt of the Earth's axis relative to the plane of revolution.

Air temperature and rainfall are the most used variables that represent the meteorological conditions. *Temperature* is one of the effects of sun radiation. The atmosphere warming near the surface occurs mainly by transport of heat from surface heating by solar rays.

The height of sensor for measuring the temperature is between 1.5 to 2.0 m above the surface in a meteorological shelter to allow the free passage of air, but prevents the incidence of solar radiation on equipment.

For meteorological and climatological purposes, the air temperature is measured under a reference condition, which allows comparison between different locations. Considering this reference condition, the default pattern of daily variation for temperature during the Winter in the São Paulo state is similar to the graph presented in Figure 2.2(a). During Summer, the average temperature is higher than during the Winter, as it can be seen in Figure 2.2(b). The maximum temperature occurs with a lag of two or three hours in relation to the peak of solar irradiance (12 o'clock on days without clouds), whereas the minimum temperature occurs just before sunrise due to the night cooling (Pereira et al., 2002).

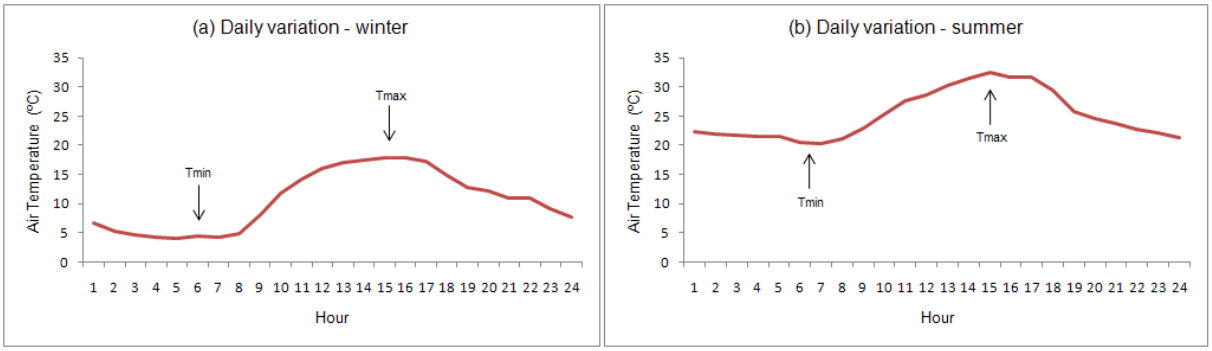



Figure 2.2: Typical daily variation of air temperature in the Sao Paulo state: (a) During Winter, temperatures range from 3°C to 20°C (b) In the Summer, temperature is higher than in other seasons (from 18 to 35).

Rainfall is another important variable that is used in different analyses by climatology and agrometeorology. In tropical regions, *rainfall* or *pluviometric precipitation* is the main way that water returns from the atmosphere to the Earth's surface after the evaporation and condensation process. The volume and distribution of rainfall that precipitates annually in a given region determine the possible kinds of vegetation and agricultural exploration.

Rainfall may be characterized by their source, such as frontal passage, local convection and orographic effects (mountains). Rainfall is measured by the pluviometric height, which is the height of precipitated water and is expressed in millimeters (mm). This pluviometric height is equivalent to the height of the precipitated volume by unit of

horizontal area.

Usually, the data gathering is made everyday at 12 UTC or 9:00 a.m. LT (Local Time) in the conventional agrometeorological post. The record of rainfall value is continuous in an automatic ground-based station, obtaining values of intensity and total height from 0 to 24 hours.

In the São Paulo state, the distribution of rainfall along the year follows the pattern of the Climatological Normal. In other words, Summers/Springs are rainy and Fall/Winters are dry as exemplified in Figure 2.3.

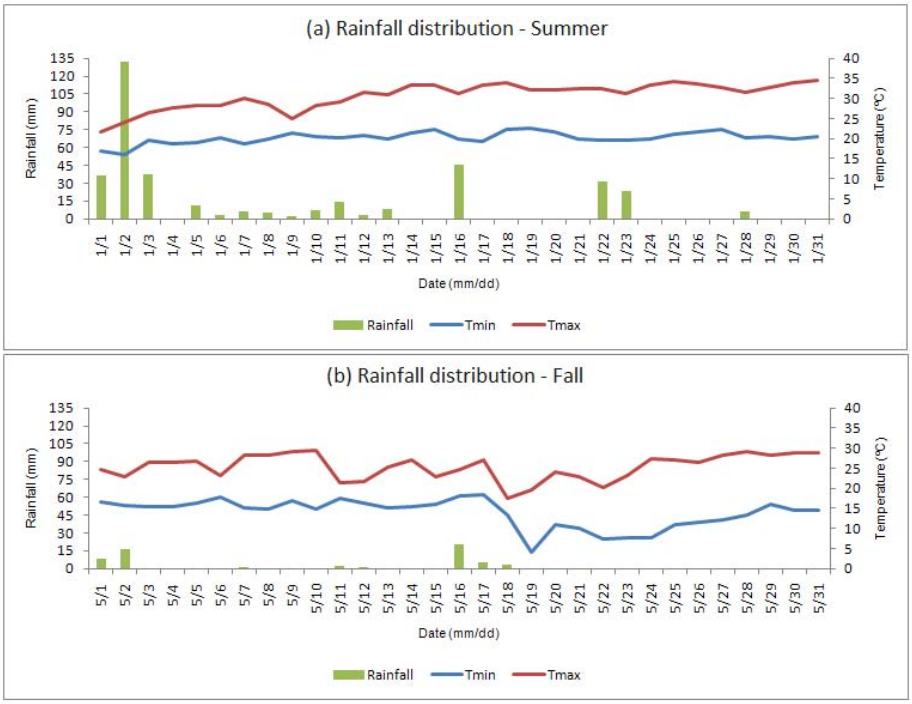



Figure 2.3: Graphs of rainfall distribution in two different seasons: Summer and Fall. (a) The first one presents more occurrence of rainfall along the month (January) with extreme rainfall and high values of maximum temperatures. (b) The second one shows lack of rainfall during the month (May) and minimum temperatures reaching lower values.

São Paulo is an important state in Brazil, responsible for approximately 30% of the Brazilian gross domestic product (GDP). This state has many activities in the industry and agriculture segments that depend on water resources. Thus, any change in the climate conditions can impact on the social and economy of the state.

In order to verify if there are alterations in the total and extreme rainfall in São Paulo, Dufek & Ambrizzi (2008) assessed some climate change indexes derived from daily precipitation data. They discovered that the number of days with very heavy precipitation increased as well as the number of consecutive dry days. Additionally, they observed that intense precipitation was concentrated in a few days in the studied period (1950 to 1999). These results indicate an important change in São Paulo's climate reinforcing the

requirement of more researches on climate change and anomalies.

### 2.2.1 Anomalies and Climate Change

Meteorologists are interested in defining the climatic behavior of each region by identifying anomalies, which are meteorological and climatological events with large deviations from the mean. Some examples of anomalies are long periods of drought, unusual floods, heat waves, days with extreme rainfall, increasing in the number of hurricanes, etc.

Experts have tried to explain these anomalies through phenomena, which occur in the ocean and affect the nature, such as El Niño and La Niña. The El Niño Southern Oscillation (ENSO) is a phenomenon of the atmosphere-ocean interaction, associated to the changes in normal patterns of the Sea Surface Temperature (SST) and trade winds in the region of Equatorial Pacific between the coast of Peru and the west Pacific near Australia (Berlato & Fontana, 2003). ENSO is composed of an oceanic component (El Niño), which is characterized by the warming of surface waters in the tropical eastern Pacific Ocean. The Southern Oscillation is the atmospheric component of ENSO and is characterized by changes in surface pressure in the tropical western Pacific. ENSO is popularly called just El Niño, which is a Spanish word and refers to the Christ child due to this annual warming in the Pacific that usually occurs around Christmas. Nowadays, the term is applied only to anomalous events.

The cold phase of ENSO is known as La Niña when the cooling of surface water in the eastern Pacific intensifies and the trade winds strengthen. Both oceanic and atmospheric components are coupled because the surface pressures in the western Pacific are high during the warm phase (El Niño) and these pressures are low when the cold phase is in effect (La Niña).

ENSO is associated with floods, droughts and other weather extreme events in different regions of the World. In El Niño years, there is an increase in the number of occurrences of heavy rains between October and February with a break in January in the southeast region of Brazil. However, the signal is less pronounced about the total of monthly rainfall occurring in some regions and not in others. In years of La Niña the opposite occurs, reducing the number of extreme events and their intensity.

Figure 2.4 presents SST variation in the region of Niño 3.4 (one of the four regions where temperature in Pacific ocean is measured) for the period 1970 to 2010, where it can be seen the three strongest warm events (ENSO) of the twentieth century (1973, 1982/1983 and 1997/1998). Moreover, it can also be seen a concentration of events ENSO between 1980 and 2000 with six El Niños and five La Niñas.

Recent studies have indicated a disturbing situation regarding the temperature and

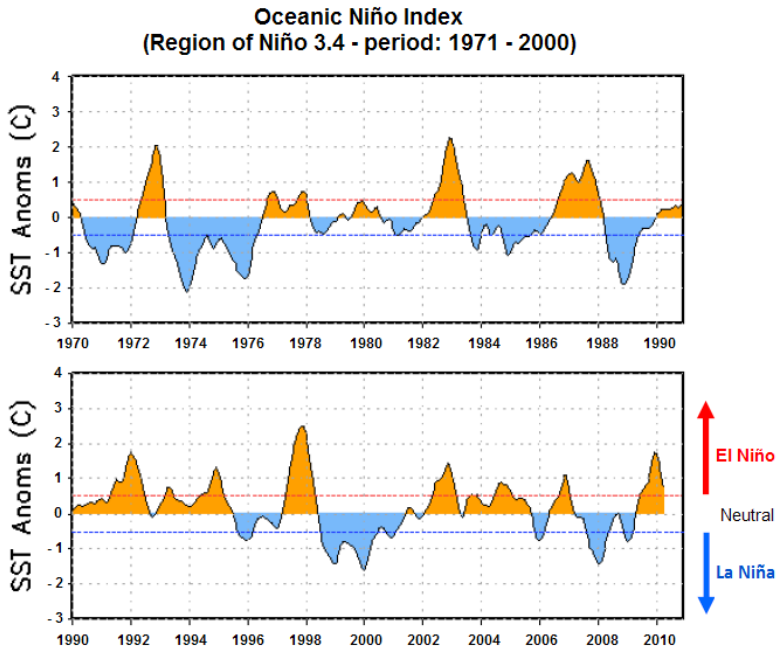



Figure 2.4: Anomalies of Sea Surface Temperature (SST) in the region of Niño 3.4 (1970 - 1990 and 1990 - 2010) (adapted from NOAA).

precipitation in the Planet. Specifically, the results of several analyses have shown that some extreme weather events have changed in frequency, duration and intensity over the last years (Meehl & Tebaldi, 2004; Vincent et al., 2005; Groisman et al., 2005; Goswami et al., 2006; Alexander et al., 2006; Ganguly & Steinhaeuser, 2008). Consequently, increased temperatures and regional changes in rainfall patterns can have adverse effects on natural and human systems.

Since the global warming affects the whole Planet, the Intergovernmental Panel on Climate Change (IPCC) was created, aimed at evaluating and analyzing the data concerning such changes and proposing ways to deal with the problems derived from the climate changes. In February 2007, IPCC published its Fourth Assessment Report on Climate Change, or IPCC-AR4 for short (IPCC, 2007). It indicates increases in the global average temperature between 1.8°C and 4.0°C by 2100. Regarding Brazil, in the same period, the average temperature shall increase around 0.75°C (Marengo et al., 2007).

In order to assess the real impact of such increases, as well as on how to deal with it, initiatives of collaborative work involving meteorologists, mathematicians, statisticians and computer scientists have emerged in several countries with promising results. One of them is the definition of a suite of climate change indexes derived from daily temperature and rainfall data in order to organize and allow comparisons among works around the World.

To understand trends of extreme events it is very important that governments and

communities learn and are prepared to mitigate the problems, and more importantly, to make decisions in a timely manner. Additionally, analyses of temperature time series indicate that it is crucial to define methods to reduce the emission of greenhouse gases and to adapt agricultural crops to the new conditions of increasing temperatures.

An alternative to reduce emissions of greenhouse gases is to replace fossil fuels with renewable sources. In Brazil, sugar cane is the main agricultural crop used to produce ethanol. Although sugar cane benefits from the increase of temperature, other agriculture crops will suffer with the global warming. Several studies simulating spatial distribution for agriculture crops, and considering the climate changes perspectives pointed by IPCC were performed by Brazilian experts. One of the most recent studies claims that temperature increases can lead to harvest losses in grain crops in the order of R$ 7.4 billions in 2020, and that the loss can achieve R$ 14 billions in 2070, what would deeply affect the geography of the agriculture production in Brazil (Pinto & Assad, 2008; Assad et al., 2007).

The work described in (Pinto & Assad, 2008) shows that the temperature increase predicted by IPCC (2007) would lead to a new distribution of agriculture crops in Brazil by the end of the 21st century. Among the most damaged crops will be coffee. Therefore, it would be advantageous for the country to plan ahead how to profit of such scenarios without damaging other crops. Uncontrolled sugar cane expansion can impair other food crops, thus negatively impairing the country economy.

The studies simulating the impact of temperature increases on agriculture crops consider the calculus of agrometeorological variables that include temperature. Based on these new values of agrometeorological indexes, experts are able to assess and to propose alternatives in order to mitigate the effects of global warming on Agriculture.

### 2.2.2 Water Balance and WRSI

Interaction of the soil-plant-atmosphere system refers to a certain amount of water that enters and leaves of each one of these components, what implies in a constant variation in the water stored in soil. This variation represents the balance of input and output of the system, whereas the intensity depends on the environment conditions. Water balance is an accounting soil water by applying the principle of mass conservation for water in a volume of vegetated soil (Pereira et al., 2002).

Basically, there are six possible inputs (rainfall, dew, superficial runoff, side drain, rising damp and irrigation) and four outputs (evapotranspiration, superficial runoff, lateral drainage and deep drainage). Rainfall and dew depend on the climate of the region, while the other entries depend on soil type and topography of the region. Thus, the driving

force of the system is climate (Pereira et al., 2002).

One of the most widely used methods to calculate the water balance was proposed by Thornthwaite & Mather (1955). This method requires basically evapotranspiration data, crop coefficient over the period of crop growing, rainfall, temperature and available water capacity for the considered period.

**Definition 2.3** ***Evapotranspiration*** *is defined as the process of transferring water to atmosphere by evaporation of water from the soil and transpiration from plants (Thornthwaite & Mather, 1955; Pereira et al., 2002).*

Other variables have been used to improve the water balance calculation, such as potential evapotranspiration, real evapotranspiration and maximum evapotrasnpiration.

**Definition 2.4** ***Potential evapotranspiration*** *occurs in an area vegetated with grass and availability of water in the soil, in active growth phase, covering the soil surface. This evapotranspiration is indicative of the evapotranspiration demand of atmosphere in a given place and period (Ometto, 1988; Pereira et al., 2002).*

**Definition 2.5** ***Real evapotranspiration:*** *is considered the real amount of water used by a large surface that is vegetated with grass in an active growth with or without water restriction. The real evapotranspiration is equal to potential evapotranspiration when water restriction does not occur (Pereira et al., 2002).*

**Definition 2.6** ***Maximum evapotranspiration*** *or* ***agricultural crop evapotranspiration*** *is the evapotranspiration when plants are able to maximize production and reach the maximum yield. That is, it is the amount of water used by an agriculture crop at any stage of its development (planting to harvest) without water restriction (Pereira et al., 2002).*

Maximum evapotranspiration can be obtained from the potential evapotranspiration, according to Equation 2.1, proposed by Jensen (1968).

$$ETM = Kc * ETP \tag{2.1}$$

where ETM is the maximum evapotranspiration, Kc is the crop coefficient and ETP is the potential evapotranspiration.

The *crop coefficient* (Kc) is the adjustment factor between the maximum evapotranspiration and the potential evapotranspiration that varies according to the maturity stage

of the plant, species and cultivars. This coefficient is calculated considering the leaf area index (LAI).

According to Doorenbos & Kassam (1994), adequate moisture during the period of sugar cane growth is important to obtain maximum yields, since vegetative growth is directly proportional to transpired water.

One way to assess the climate impact on the sugar cane production can be through the use of indexes, which cover the main atmospheric parameters simultaneously rather than using each one individually. One example is the Water Requirements Satisfaction Index (WRSI), which is calculated through simulations of water balance during the agricultural crop cycle. This index is obtained by Equation 2.2.

$$WRSI = \frac{ETR}{ETM} \tag{2.2}$$

where ETR is the real evapotranspiration and ETM is the maximum evapotranspiration.

WRSI expresses the relationship between the quantity of water consumed by the plant and desirable one in order to ensure maximum productivity of the plant(Assad & Sano, 1998). This index ranges from zero to one, being the highest when the amount of stored water in soil is the maximum one. Thus, WRSI is related to the volume of rainfall and the water storage in soil.

## 2.3 Sugar Cane Crops

Sugar cane (*Saccharum officinarum L.*) is originated from Asia, probably in New Guinea, mainly produced between latitudes 35° north and 35° south (Doorenbos & Kassam, 1979). It is a semi-perennial grass that can have vegetative cycle of 12 months when planted early in the rainy season (from September to November). When it is planted in the middle of the rainy season, its growing cycle enlarges from 14 to 21 months.

According to Alfonsi et al. (1987), Brazil is the only country with two seasons of annual harvests, one in the north-northeast, which starts in September and extends until April and another in the central-south from June to December.

Before the first cut, sugar cane is called *plant cane*. After the cut, clogs or knuckles sprout after 20 to 30 days, resulting in a *ratoon cane* that has its completed cycle in approximately one year (Camara, 1993). Figure 2.5 shows the growth cycle of both plant cane and ratoon cane.

Sugar cane is essentially a tropical plant and its best growing conditions occurs under high temperature and humidity. The optimum temperature for sprouting on the stalk is 32°C to 38°C. The optimum growth is achieved by daily average temperatures between

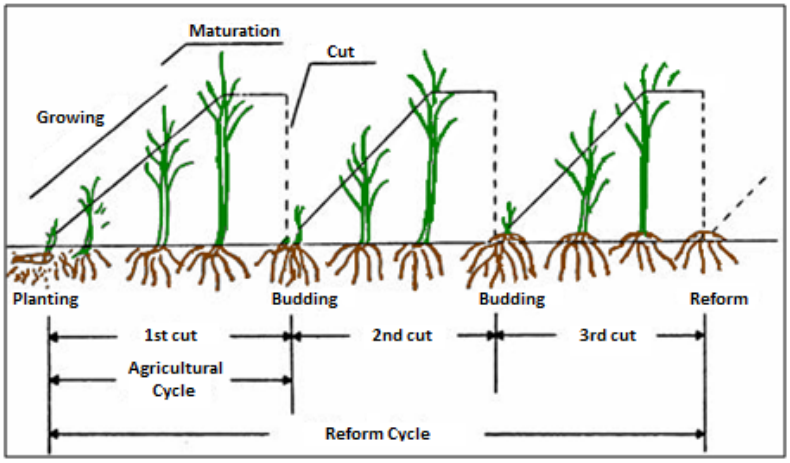



Figure 2.5: Cycle of sugar cane with three cuts (adapted from (Gonçalves, 2008)).

22°C to 30°C. Minimum temperature for strong growth is approximately 20°C. However, for sugar cane maturation are convenient temperatures around 20°C to 10°C (Doorenbos & Kassam, 1979).

Sugar cane grows rapidly in ideal weather conditions, producing wide and relatively thick internodes. The sugar cane growth becomes slower and the internodes become shorter and thinner whether one of the weather factors fail (Fauconier & Bassereau, 1975). Sugar cane matures later when the heat and rain in the Summer extend to the Fall. Consequently, the amount of sugar stored in stems is reduced. On the other hand, the best yields of sugar cane rich in sucrose are obtained when Summer is hot and humid, and Fall is cool and dry. It grows less during periods of drought in the Summer and in the overcast days (Godoy & Toledo, 1972).

Generally, sugar cane requires six to eight months with high temperatures, intense solar radiation and regular rainfalls to allow full vegetative growth. After that, it needs four to six months with dry season and/or low temperatures, which are adverse conditions to growth, but extremely beneficial and stimulating to the accumulation of sucrose (Casagrande, 1991).

Sugar cane provides the raw material for the production of alcohol, which is mixed with diesel generating biodiesel. Researchers in Brazil have investigated others oleaginous species as potential sources for biodiesel production (Nass et al., 2007). However, ethanol deriving from sugar cane is the main alternative to replace fossil fuels and researchers have been studying the issues related to greenhouse gases emissions in the production and use of ethanol (Goldemberg, 2007; Macedo et al., 2008; Goldemberg et al., 2008). Economic factors contribute to the use of ethanol as fuel due to the inherent instability in the supply of oil as well as its high price. Also, ethanol is considered renewable/energy sources.

Environmental and social issues are linked to the expansion of sugar cane for ethanol

production in Brazil, such as atmospheric pollution due to burning, degradation of soils and water resources as well as exploitation of cane cutters, among others (Martinelli & Filoso, 2008). In their work, Martinelli & Filoso (2008) provide some recommendations to aid at establishing a code for ethanol production by policy makers and Brazilian government. The production of sugar cane in Brazil, in general, has been growing in recent years. The production of São Paulo state represents approximately 60% of the national production of sugar cane.

In this scenario, a massive data volume has been generated by industry, research institutes and universities due to the increasing importance of the sugar cane production. Traditional statistical methods and new approaches based on data mining methods have been employed to assess data related to sugar cane production and to predict sugar cane yields (Garcia & Vieira, 2008; Everinghama et al., 2009; Ferraro et al., 2009). In addition, remote sensing images have been used in studies about sugar cane production as detailed in the next section.

## 2.4 Concepts of Remote Sensing

*Remote sensing* is defined as the use of modern sensors, equipment for processing and transmitting data, aircraft and spacecraft, aimed at studying the terrestrial environment by recording and analyzing the interactions between electromagnetic radiation and substances, which compose the Earth planet regarding their various manifestations (Novo, 1992).

Remote sensors are devices designed to measure electromagnetic energy (in certain intervals of the electromagnetic spectrum) from an object. Remote sensors measure the energy from targets in the Earth's surface, as shown in Figure 2.6.

Remote sensors transform electromagnetic energy into an electrical signal that can be stored or transmitted in real time. Thus, the signal is converted into information describing the features of objects, which compose the Earth's surface. The multispectral characteristic of remote sensing allows recording the electromagnetic energy in the wavelength ranges called bands or channels. Knowledge about spectral behavior of land surfaces in different wavelength bands of the electromagnetic spectrum is crucial to be able to use satellites to monitor agriculture. Based on this knowledge, several vegetation indexes have been developed in recent years.

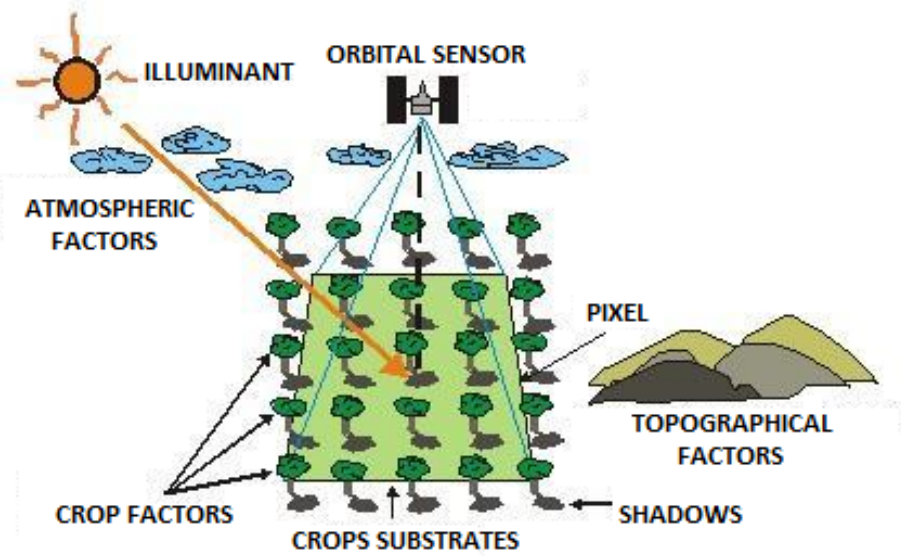



Figure 2.6: Illustration showing the relationship between source-target-sensor (adapted from (Antunes, 2005)).

### 2.4.1 Vegetation indexes

Vegetation indexes are based on the combination of electromagnetic radiation reflected by vegetation in some spectral bands of the electromagnetic spectrum. The spectral measurement represents the relationship between the amount and condition of vegetation in the region, where this measurement was made (Moreira & Shimabukuro, 2004). Spectral vegetation index is a quantity obtained by sum, ratio, difference, or other processing of spectral data, to represent the characteristics of vegetation cover (Ponzoni, 2001).

Vegetation indexes have been studied to characterize biophysical parameters of vegetation. They indicate the presence of some characteristics and/or some condition regarding the vegetation. For example, biomass is related to the solar energy absorbed. In general, vegetation indexes are associated with leaf area index (Xavier & Vettorazzi, 2004; Wang et al., 2005), green biomass (Anyamba & Tucker, 2005) and vegetation productivity (Holben et al., 1980).

Several vegetation indexes have been proposed due to different possibilities, such as sensor type, and applications. The most used is the Normalized Difference Vegetation Index (NDVI), which was proposed by Rouse et al. (1973). NDVI is based on data from channels 1 (red) and 2 (near-infrared) combined through the Equation 2.3.

$$NDVI = \frac{(\rho_{NIR} - \rho_R)}{(\rho_{NIR} + \rho_R)} \tag{2.3}$$

where $\rho_{NIR}$ is the reflectance in the near-infrared (channel 2) and $\rho_R$ is the reflectance in the red (channel 1).

NDVI values vary in the range (-1, +1). Values close to zero indicate regions without vegetation, and values close to +1 indicate vegetated areas with the highest possible density of green leaves. As NDVI has limitations (rapid saturation for increasing green biomass, sensitivity to soil background effects and less effective for atmospheric correction), other indexes have been proposed such as the Enhanced Vegetation Index (EVI), which is a modified version of NDVI and was developed for Moderate Resolution Imaging Spectroradiometer (MODIS) data (Huete et al., 1997).

### 2.4.2 NOAA satellites and AVHRR sensor

Several satellites can be used to help the monitoring and estimation of agricultural production, mainly the satellites from the National Oceanic and Atmospheric Administration (NOAA), with the Advanced Very High Resolution Radiometer (AVHRR) sensor. This sensor is applied to studies of ecosystems due to availability of long time series of its data. Moreover, other advantages of AVHRR sensor are global coverage and free access to data. NOAA-AVHRR images have been used in land surface studies, such as drought investigation (Bajgiran et al., 2008), estimation of crop area and yield (Liu & Kogan, 2002), vegetation phenology estimation or evaluation (Maignan et al., 2008).

NOAA meteorological satellites are scheduled to accomplish two daily passes (day and night) focusing the same target on Earth (Kampel, 2004). They are polar-orbiting and spend approximately 102 minutes to cross the Equator line again. The passes of each satellite occur in the same solar hour regardless of the latitude. Thus, the frequency of daily image acquisition on the same point is high, particularly when several satellites are working well simultaneously.

Since February 2009 four satellites (NOAA-15, NOAA-17, NOAA-18, and NOAA-19) have been in orbit with their main sensors in full operation. Therefore, it is possible to have at least four images by day and four images by night every day at each point of Earth's surface. Table 2.1 shows the hours of Equatorial crossing time for the latest NOAA-AVHRR satellites.

Table 2.1: Equatorial crossing time of latest NOAA-AVHRR satellites.

| | **Orbit** | |
|---|---|---|
| **Satellites** | **Ascending (LT)** | **Descending (LT)** |
| NOAA-15 | 05:00 p.m. | 05:00 a.m. |
| NOAA-17 | 10:00 p.m. | 10:00 a.m. |
| NOAA-18 | 04:00 a.m. | 04:00 p.m. |
| NOAA-19 | 02:00 a.m. | 02:00 p.m. |

AVHRR sensor has five spectral channels in the visible, near-infrared, mid-infrared and

thermal infrared. The NOAA-AVHRR has a 12 hour temporal resolution and a spatial resolution of 1 km by 1 km at nadir[1] and 2.4 x 6.9 km in extremes (Townshend, 1994). Figure 2.7 presents an RGB-321 (where the channel 3 is in the red (R), channel 2 is in the green (G) and channel 1 is in the blue (B)) image of the São Paulo state acquired by NOAA17-AVHRR on May/27th/2006 at 13:04GMT.

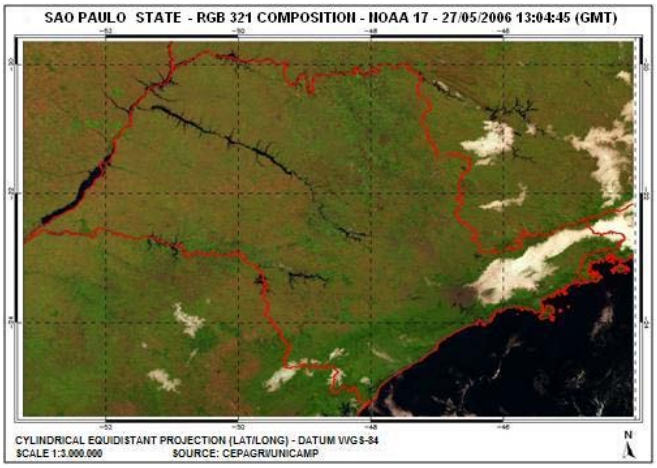



Figure 2.7: An example of RGB-321 image of the São Paulo state acquired by NOAA17-AVHRR on May/27th/2006 at 13:04GMT

AVHRR images are received by an acquisition station, located at the view angle of NOAA satellites to capture and to record data. Transmission mode of AVHRR images is the High Resolution Picture Transmission (HRPT). NOAA-AVHRR images have geometric distortions caused by Earth curvature, rotation, attitude errors and imprecise orbital (Rosborough & Baldwin, 1994). These distortions must be corrected specially for land applications that require a highly accurate geometric matching, following some phases, such as:

- Format conversion from HRPT to standard;
- Radiometric calibration;
- Geometric correction;
- Identification of pixels classified as cloud.

We have used the NAVPRO system developed by Esquerdo et al. (2006) to perform the preprocessing of AVHRR images, specially the geometric correction. NAVPRO is an automatic set of C-shell scripts that call subroutines of the NAV system described

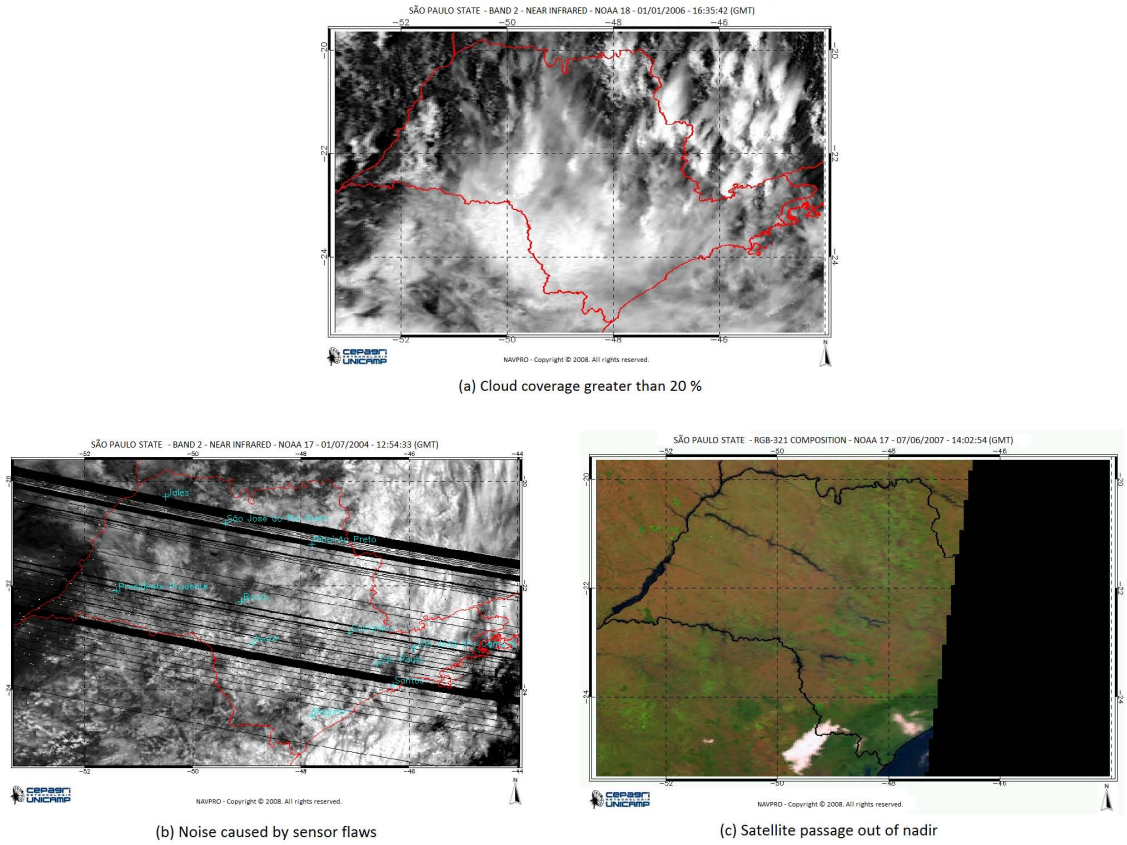
[1]Intersection point between vertical line that is perpendicular to the horizontal plane and the celestial sphere, but in the opposite hemisphere to that one, where observer is located.

by Emery et al. (1989), developed by the Colorado Center for Astrodynamics Research (CCAR), Aerospace Engineering Sciences, with the University of Colorado, Boulder, USA.

NAVPRO converts the images from High Resolution Picture Transmission (HRPT) raw data format to another one named CCAR, similar to Level 1B Local Area Coverage (LAC). The next step is the radiometric calibration, when digital numbers are transformed to reflectance at the top of atmosphere for 1, 2 and 3A AVHRR channels and brightness temperatures for 3B, 4 and 5 channels.

Geometric correction combines indirect navigation and spacecraft attitude error estimation. After that, the Maximum Cross Correlation (MCC) technique is used to detect the geographic displacement between a base image and a target one (Emery et al., 2003). When the MCC technique finishes, output images are calibrated and submitted to an accurate navigation. Thus, the images are stored in a file containing the seven channels. The first five channels are composed of each AVHRR channels.

Some images are eliminated during the process of image correction due to problems, such as cloud coverage greater than 20% or different kinds of noise, as illustrated in Figure 2.8.




(a) Cloud coverage greater than 20 %

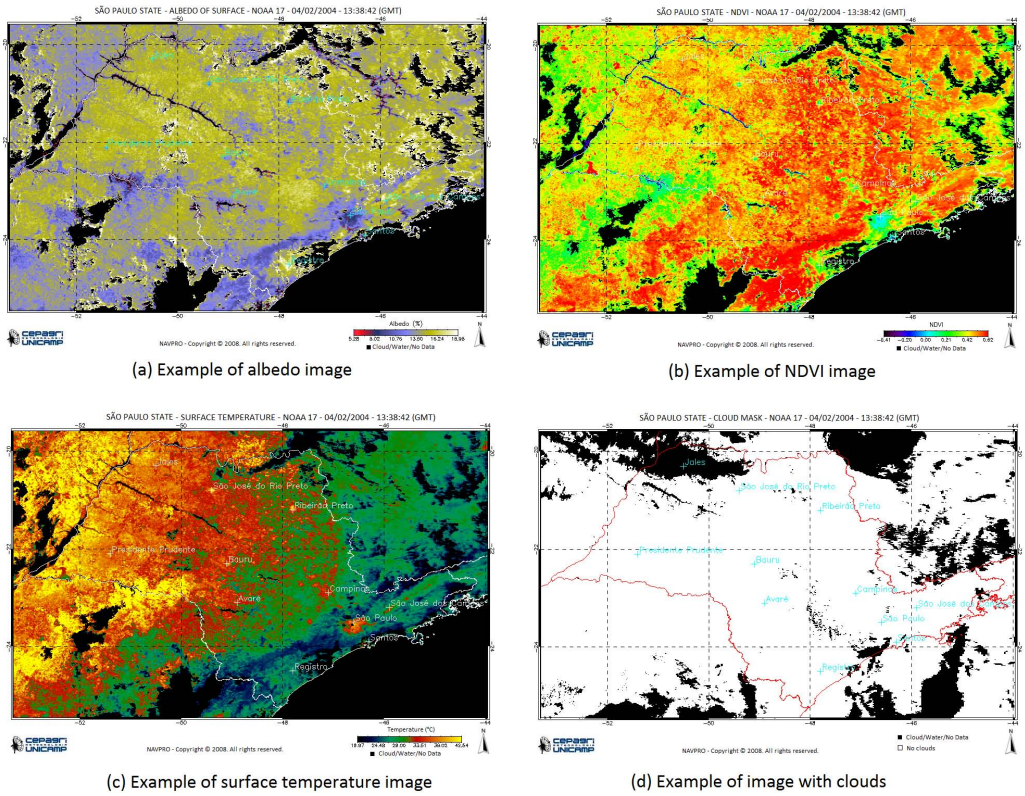
(b) Noise caused by sensor flaws

(c) Satellite passage out of nadir

Figure 2.8: Examples of images that were eliminated due to different problems during geo-processing.

Measurements and indexes, such as cloud mask, surface temperature, albedo and

NDVI images are generated wherever an image satisfies all conditions in the steps of the NAVPRO system. Examples of these images are presented in Figure 2.9. Albedo is the ratio between the amount of radiation reflected by the Earth's surface (including the atmosphere) and the total of incident radiation (from Sun) at a given temperature. Surfaces of sand and snow have high albedo values while forests have low values of albedo (Song & Gao, 1999).




(a) Example of albedo image

(b) Example of NDVI image

(c) Example of surface temperature image

(d) Example of image with clouds

Figure 2.9: Examples of images generated by the NAVPRO system.

The effect of shadows, aerosols and water vapor are minimized by the use of Maximum Value Composite (MVC) for NDVI, as described by Holben (1986), using images from the same satellite. The vegetative evolution of sugar cane can be understood by analyzing MVC images of NDVI in the northeast of São Paulo. The growing season usually begins in June and this aspect is represented by green and blue shades in the NDVI images. These colors represent the low NDVI values, which indicate areas with exposed soil and sparse vegetation. These colors also appear in the NDVI images from July to November. From December, when sugar cane crops have more biomass, these regions acquire yellow, orange and red shades in the images. The maximum NDVI is represented by a stronger red shade when sugar cane crops reach their peak of development from February to May. Dark areas represent regions covered by clouds. This phenomena occurs mostly in December, January and February.

### 2.4.3 Application of remote sensing in Agriculture

NOAA-AVHRR images are a source of spectral information about agricultural regions in Brazil. In the literature, there are several papers analyzing NDVI time series to improve agriculture monitoring. Wang et al. (2003) have obtained a meaningful value of cross-correlation between NDVI and soil humidity, precipitation and temperature for different kinds of vegetation, such as forest, grass and some agricultural crops from 1989 to 1997, in Kansas. This research indicated a strong correlation between pluviometric precipitation and NDVI. According to Lucas & Schuler (2007), NDVI behavior is similar to precipitation trend. This study showed that precipitation of a preceding month is more relevant to calculate NDVI values than rain values of the current month. The relationship between pluviometric precipitation and spectral data from NDVI and EVI MODIS images were studied by Fontana et al. (2005). This research showed that NDVI and EVI indexes are indicators of Winter crops yield in Australia. Both indexes had similar behaviors in temporal scale, being associated to pluviometric precipitation accumulated in the period from April to June.

Sugar cane crops are cultivated on large fields and can be monitored by remote sensing images of medium and low resolution, such as the NOAA-AVHRR. In recent years, several studies have been developed to identify areas for sugar cane expansion, to assess its social and economic impact, to predict its yield, to monitor diseases, among other applications. Reflectance of crop canopies is a combination of reflectance of plants and soil (Guyot, 1990). Specifically, the spectral response of sugar cane depends on canopy architecture, foliar chemistry, agronomic parameters and geometry of data acquisition and atmospheric conditions. Abdel-Rahman & Ahmed (2008) have cited several works that discussed how light interacts with the sugarcane canopies.

Another important contribution of remote sensing to sugar cane monitoring is the identification of sugar cane areas and accurate forecast of the cultivated crop fields, which are needed for crop yield estimation. Xavier et al. (2006) used an unsupervised method to classify sugar cane crop in Brazil through EVI images from MODIS. The results showed that sugar cane can be distinguished from natural and planted forests, peanuts, soybean, water bodies and urban regions. However, it was difficult to distinguish sugar cane from pasture, but the use of images from higher spatial resolution sensors could aid to minimize this spectral mixture.

Nascimento et al. (2009) have used harmonic analysis applied to time series of NOAA-AVHRR to identify sugar cane areas in the São Paulo state. They generated a decision tree to search patterns that could represent sugar cane along the crop season of 2006/2007. The application of both methods (harmonic analysis and decision tree) was able to identify

sugar cane fields with 92% of confidence when compared to ground truth information from the CANASAT project[2].

Mapping of sugar cane varieties is extremely important for crop damage risk assessment and yield prediction. Multispectral remotely sense data have been used to identify sugar cane varieties (Fortes & Demattê, 2006). Remote sensing data have also been used in the prediction of sugar cane yield. In Brazil, Rudorff & Batista (1990) used LandSat data and agrometeorological model to predict yield. According to Nascimento et al. (2009) a phenology-spectral model, such as the proposed by Pellegrino (2001), with a time series of NOAA17-AVHRR images presented satisfactory results in the estimation of sugar cane productivity with relative errors below 5% and anticipation of about 110 to 150 days before the harvest.

Multi-temporal images are a useful source of information for monitoring agricultural crops fields. Nevertheless, data acquired from satellite images have often missing or uncertain radiometric values. Hajj et al. (2009) proposed an approach that addresses this issue by combining time series of satellite images with information from crop growth modeling and the expert's knowledge. They used the fuzzy logic and modeled linguistic terms, which helped them to build expert decision rules.

In recent years, the sugar cane crops have expanded due to several reasons, such as biofuel production, potential benefits to the environment as a possible way of mitigation of greenhouse gases, economic impact and others. Although there are traditional ways to evaluate the sugar cane expansion, remote sensing images have been an important source of information to evaluate the direct land conversion to sugar cane. Rudorff et al. (2009) have used time series of EVI images from MODIS to identify the land use prior to the conversion to sugar cane in Brazil. They observed that pasture land in 2000 were gradually converted to annual crops until 2005 and then to sugar cane. In a recent study, Rudorff et al. (2010) confirmed that remote sensing images have been efficient to aid at evaluating important characteristics of the sugar cane cultivation, providing relevant results to the debate of sustainable ethanol production from sugar cane in Brazil.

## 2.5 Summary

In this chapter, we detailed some important concepts about Climatology, Agrometeorology and Remote Sensing to better understand the analyses and experiments that are presented further in this work. We also discussed aspects concerning sugar cane crops and its economic and social impacts. The importance of sugar cane production for Brazil as well

[2]http://www.dsr.inpe.br/canasat

as the demand for new techniques and methods to analyze sugar cane data is one of the motivations for the achievement of this research project. As it could be seen, few works propose the use of data mining techniques to discover correlations, patterns and extremes in these datasets.

In this doctorate project, we have used remote sensing images of NOAA-AVHRR due to existence of long time series and global coverage. As it can be seem previously in this chapter, several works have applied remote sensing images to monitor sugar cane fields. The majority of these studies have used satellite of medium or high spatial resolution. However, as sugar cane crops are cultivated in large and contiguous fields, satellites of low spatial resolution, such as NOAA can be used with satisfactory results. Several experiments that corroborate to this issue will be presented in the following chapters.

This chapter also points different problems that often occur with remote sensing images that make difficult the use of this kind of image. For instance, it is very complicated to deal with noise and cloud coverage in the images, as well as to extract time series from multi-temporal images. We will show that some of these problems were satisfactorily solved in this work, while others are still open problems. In the next chapter, we present important concepts about the fractal theory, which was the first approach used to provide a solution involving agrometeorological and remote sensing data in this thesis.

# Chapter 3

# Fractal Theory

## 3.1 Introduction

*Fractal* is defined as an object that presents roughly the same characteristics regardless of the scale where it is analyzed, i.e., it is a self-similar object. Therefore, parts of a fractal (a mathematical structure, an object or a dataset) are similar, exact or statistically, to the whole fractal. That is, small scale details are similar to large scale characteristics (Schroeder, 1991; Traina Jr. et al., 2005).

For example, the Sierpinski triangle is a geometrical fractal. It is built by a recursive iterative process, theoretically infinite. Given an equilateral triangle $ABC$, we first remove the central triangle $A'B'C'$; from each of the three remaining triangles whose sides have length equal to half of the side of $ABC$, we withdraw again the central triangle, and so on. Figure 3.1 shows the initial steps of the building process of a Sierpinski triangle. The remaining triangle has "holes" regardless of the scale and each triangle inside the first one is a "miniature" of the whole triangle.

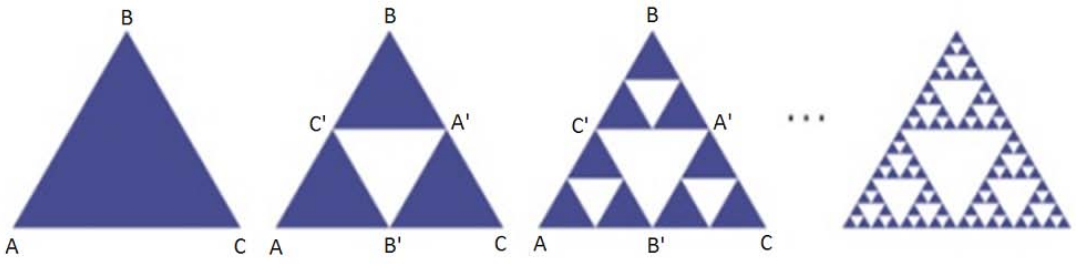



Figure 3.1: Steps of the building process of Sierpinski triangle.

There are many other mathematical structures defined as fractal, such as the Koch curve, the Cantor set, and the Mandelbrot set that are presented in Figure 3.2. There are also examples of fractals in Nature, for example: clouds, mountains, vegetables, trees, the coast of continents, islands and others (Mandelbrot, 1983).

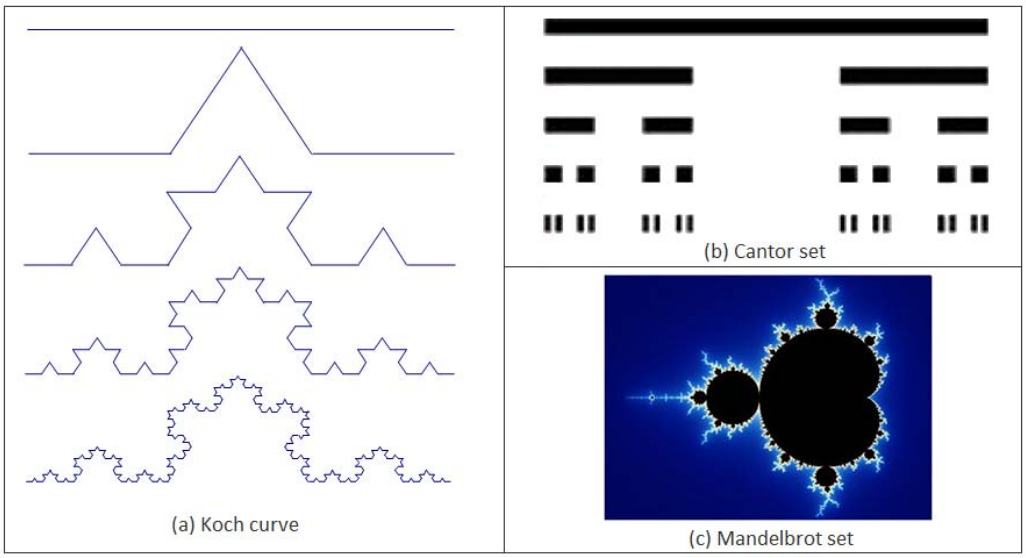



Figure 3.2: Some examples of fractals: (a and b) geometrical fractals and (c) algebraic fractal.

Fractal concepts have been applied to several tasks in data analyses and data mining. One of them is the estimation of the intrinsic dimension ($D$) of the dataset, which is related to the concept of embedding dimension ($E$) (Faloutsos & Kamel, 1994).

**Definition 3.1** ***Embedding dimension E:*** *Given a finite dataset $A$, the embedding dimension $E \in \mathbb{N}$ is the number of attributes that define $A$, i.e., $E$ is the dimension of the space in which the dataset is embedded.*

**Definition 3.2** ***Intrinsic dimension D:*** *Given a finite dataset $A$, its intrinsic dimension $D \in \mathbb{R}^+$, is the dimensionality of the object represented by the data, regardless of the dimension of the space in which it is embedded.*

The intrinsic dimension ($D$) is a measure of the amount of information that the dataset represents. For example, the intrinsic dimension of a set of points distributed along a line is equal to one; if the set is embedded in a higher dimensional space, the intrinsic dimensionality continues equal to one as illustrated in Figure 3.3.

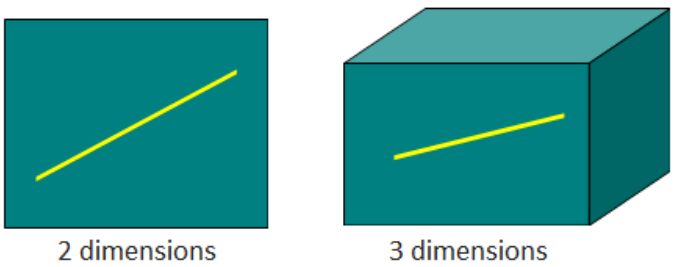



Figure 3.3: A line embedded in two and three dimensions where $D = 1$.

Faloutsos & Kamel (1994) proposed the use of the intrinsic dimension as a tool to measure the non-uniform behavior of real datasets. Moreover, the authors presented empirical studies to demonstrate that real data usually have self-similar behavior, which is the fundamental characteristic of fractal objects. Therefore, the intrinsic dimension $D$ of a real dataset can be estimated by calculating its fractal dimension.

The intrinsic dimension based on the fractal dimension has been employed as a useful tool for clustering analysis (Barbará & Chen, 2003), mining of temporal association rules (Barbará et al., 2004), attribute selection (Traina Jr. et al., 2000, 2010), time series forecasting (Chakrabarti & Faloutsos, 2002) and spatial data mining (Traina et al., 2001).

In this chapter, we present the main concepts related to the fractal theory that are used in some of the methods proposed in this thesis. Section 3.2 shows different ways to calculate the fractal dimension. The correlation computation indicated by the Correlation Fractal Dimension is detailed in Section 3.3. The fractal theory employed to monitor data stream is discussed in Section 3.4.

## 3.2 Fractal Dimension

Fractals usually have unusual characteristics that can be considered paradox. For instance, the Sierpinski triangle has infinite perimeter (proportional to $lim_{i\to\infty}(1 + 1/2)^i$) and null area (proportional to $lim_{i\to\infty}(3/4)^i$) since at each iteration of its building process, which is theoretically infinite, its perimeter increases and its area decreases. Due to these properties, this fractal can neither be considered a one-dimensional Euclidean object (since its perimeter is not finite) nor a bi-dimensional Euclidean object (since its area is null). Thus, it is possible to consider a fractionating dimensionality called fractal dimension (Mandelbrot, 1983). Intuitively, the fractal dimension of Sierpinski triangle is a value between 1 and 2. Mathematically, the precise value is 1.58.

There are several definitions for fractal dimension, which are briefly presented in this section. The basic measurement of the fractal dimension is devoted to fractals denominated *exactly self-similar*. This kind of fractal is composed of $M$ replicas being each one a scaled-down version 1:s of the original fractal.

**Definition 3.3** ***Fractal dimension $\mathfrak{D}$:*** *Let $M$ be the number of replicas and $s$ the scale factor by which each replica is reduced, the fractal dimension $\mathfrak{D}$ of an exactly self-similar fractal defined in an E-dimensional space is:*

$$\mathfrak{D} \equiv \frac{logM}{logs} \tag{3.1}$$

The Sierpinski triangle, for example, is an exactly self-similar fractal, because its rule of construction generates three replicas in 1:2 scale for each iteration. Therefore, the fractal dimension of Sierpinski is $\mathfrak{D} = log3/log2 \approx 1.58$. Similarly, in Figure 3.2 the Cantor set and the Koch curve are exactly self-similar fractals with the fractal dimension $\mathfrak{D} = log2/log3 \approx 0.63$ and $\mathfrak{D} = log4/log3 \approx 1.26$, respectively.

This definition of fractal dimension $\mathfrak{D}$ is suitable for mathematical exactly self-similar fractals with well-defined recursive construction rules. However, for datasets called *statistically* self-similar fractals, which do not have well-defined rules of construction, it is more appropriated to calculate the fractal dimension by using the Box-Counting method (Schroeder, 1991), which defines the Correlation Fractal Dimension $D_2$ as presented in Equation 3.2.

**Definition 3.4** ***Correlation Fractal Dimension*** $D_2$***:*** *Given a dataset self-similar in the range of scales* $[r_1, r_2]$*, its Correlation Fractal Dimension* $D_2 \rightarrow \mathbb{R}^+$ *is measured as*

$$D_2 \equiv \frac{\partial log(\sum_i C_{r,i}^2)}{\partial log(r)} \qquad r \in [r_1, r_2] \tag{3.2}$$

*where* $r$ *is the side of the cells in a (hyper) cubic grid that divides the address space of the dataset, and* $C_{r,i}$ *is the count of points in the* $i$*th cell.*

In a practical way, the derivate value that defines the fractal dimension $D_2$ can be obtained by the construction of the box-count plot, which represents the values of $log(\sum_i C_{r,i}^2)$ and $log(r)$ in a graph. For fractal datasets, the resulting curve is linear in an interval $(r_1, r_2)$ and the fractal dimension $D_2$ is estimated by the slope of the line that best fits the analyzed interval. Figure 3.4 shows a set of 6561 points in the Sierpinski triangle and the plot in log–log scale of the sum of squared occupancy $\sum_i C_{r,i}^2$ versus the grid cell size (radius) $r$.

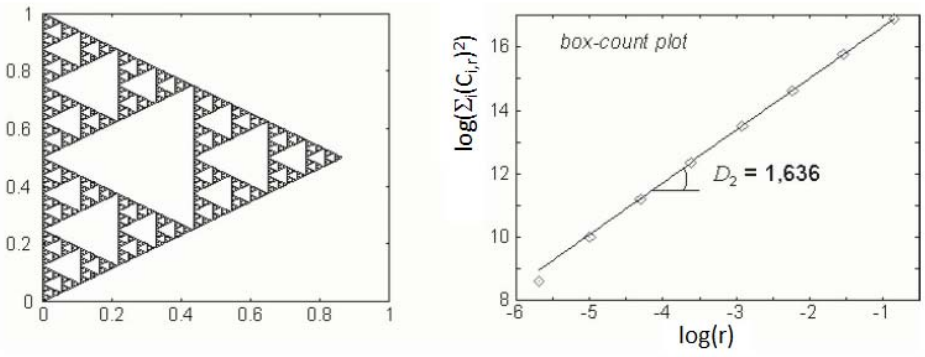



Figure 3.4: The box-counting plot for Sierpinski triangle (adapted from (Sousa, 2006)).

The fractal dimension was calculated by the Liboc() algorithm (linear cost on the number of elements in the dataset) presented in (Traina Jr. et al., 2000, 2010) and

presented as follows. Consider the address space of a point-set in an $E$-dimensional space, and impose an $E$-grid with grid-cells of side size $r$. Focusing on the $i$-th cell, let $C_{r,i}$ be the count ('occupancies') of points in each cell. Then, compute the value $S(r) = \sum_i C_{r,i}^2$. The fractal dimension is the derivative of $log(S(r))$ with respect to the logarithm of the radius. Thus, Liboc algorithm can obtain the fractal dimension $D$ of a dataset plotting $S(r)$ in log-log scales for different values of the radius $r$, and calculating the slope of the resulting line.

It is needed to process $S(r)$ for a quantity $R$ of values of $r$, so the algorithm can achieve a suitable statistical approximation of the line. To avoid reading the dataset again for each value of the radius, Traina Jr. et al. (2010) proposed to create a multi-level grid structure, where each level has a radius the half of the size of the previous level ($r = 1, 1/2, 1/4, 1/8$, etc.). Each level of the structure corresponds to a different radius, so the depth of the structure is equal to the number of points in the resulting graph. The structure is created in main memory, so the number of points in the graph is limited by the amount of main memory available. If this graph is linear for a suitable range of radii, the dataset is a fractal and its fractal dimension $D$ is the slope of the fitting line of this graph.

For each given cell side $r$, only the cells having at least one already processed point are maintained, counting the sum of occupancies $C_{r,i}$ of this cell. In this way, each new point is directly associated to a cell in each level, without the need to be compared with the previously read points. Figure 3.5 shows the structure used in the algorithm for 2- and 3-dimensional datasets.

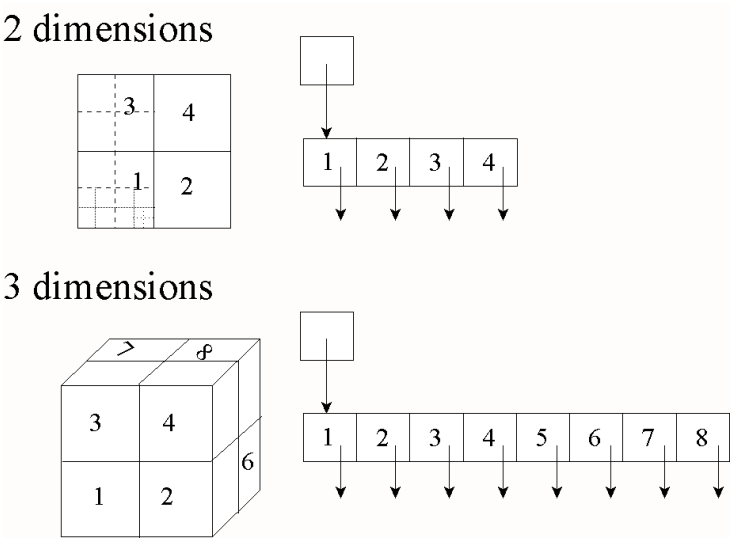



Figure 3.5: Representation of grid-cells in 2- and 3-dimensional spaces (adapted from (Traina Jr. et al., 2010)).

The largest cell side of the space of points generates $2^n$ cells. In the next level, each cell is split into other $2^n$ cells, and so on. Given that the position of each cell in the

space is always known, each cell is represented by: the sum of occupancies $C_{r,i}$ in this cell, and the pointers to the cells in the next level covered by this cell (see Figure 3.5). This structure is a kind of a multidimensional "quad-tree" (oct-tree for a 3D space, or *E-dim*-tree). Figure 3.6 shows an example of this structure for a dataset with five points in three levels in a 2-dimensional space.

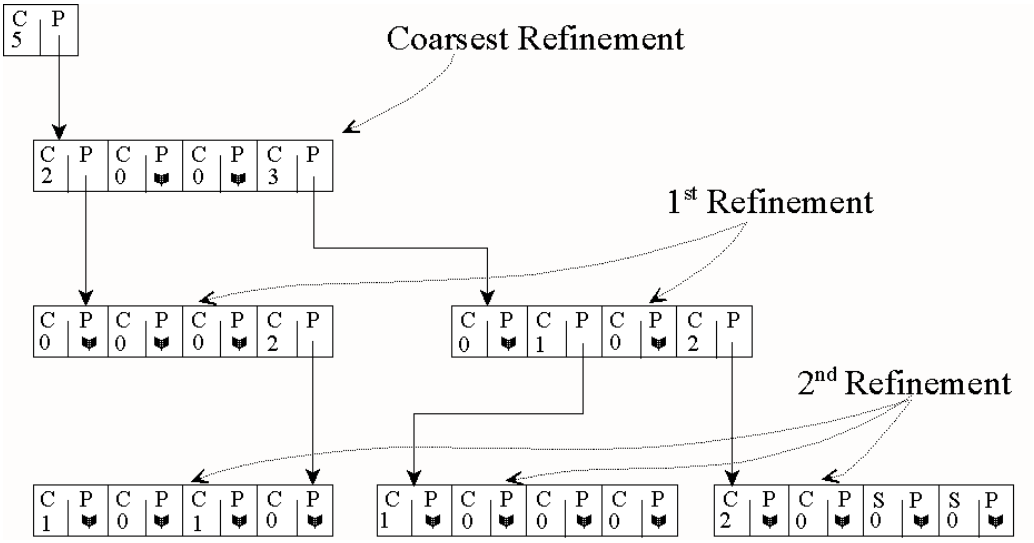



Figure 3.6: Example of the data structure used for calculating the Sum of Occupancies of a dataset with 5 points (with three levels of resolution) (adapted from (Traina Jr. et al., 2010)).

Notice that new cells are added to the structure on demand. Thus, only cells occupied by at least one point are created ($C_{r,i} > 0$). The algorithm processes the points set only once, so it is indeed very fast. Algorithm 1 summarizes this computation process.

**Algorithm 1** Compute the fractal dimension $D$ of a dataset $A$ (box-count approach).

**Input:** Normalized dataset $A$ ($N$ rows, with $E$ dimensions/attributes each)
**Output:** Fractal dimension $D$

1: **for** each desirable grid-size $r = 1/2^j, j = 1, 2, ..., l$ **do**
2: **for** each point of the dataset **do**
3: Decide which grid cell it falls in (say, the $i$-th cell)
4: Increment the count $C_i$ (occupancy')
5: **end for**
6: Compute the sum of occupancies $S(r) = \sum C_i^2$
7: **end for**
8: Print the values of $log(r)$ and $log(S(r))$, generating a plot;
9: Return the slope of the linear part of the plot (linear regression) as the fractal dimension $D$ of the dataset $A$.

In the literature, there are several methods for the fractal dimension calculation (Schroeder, 1991; Faloutsos & Kamel, 1994). However, the most suitable method for our work is the box-counting method since we have only used real datasets. The correlation

fractal dimension $D_2$ stands out for its practical and theoretical relevance. Experimentally, the calculation of $D_2$ for statistically self-similar fractals is relatively simple, and suitable primarily for the fractals formed by isolated points distributed over some regions of space where they are immersed. In theory, $D_2$ is related to the concept of correlation, as described in the next section.

## 3.3 Correlation Detection: FD-ASE algorithm

The number of attributes in a dataset determines its embedded dimension $E$, but if there are correlated attributes, its intrinsic dimension $D$ is smaller than $E$ ($D < E$) (Faloutsos & Kamel, 1994). The intrinsic dimension estimated by the Correlation Fractal Dimension $D_2$ indicates the minimum number of attributes needed to represent a dataset. The ceiling function of the intrinsic dimension ($\lceil D \rceil$) determines a superior threshold for the necessary attributes quantity to represent the fundamental features of the dataset. $D$ can also be used to discover how many and which attributes may be employed to reduce the data dimensionality. With this purpose, Sousa et al. (2007b) proposed the *FD-ASE (Fractal Dimension Attribute Significance Estimator)* algorithm aimed at identifying different types of correlations. This technique applies the forward attribute inclusion approach and uses the intrinsic dimension as a criterion to identify groups of correlated attributes and to select a relevant attribute subgroup to represent the essential data characteristics. The following definitions are needed to better understand the technique.

**Definition 3.5** ***Partial Intrinsic Dimension*** $pD()$***:*** *Given a finite dataset* $A = \{a_1, a_2, \ldots, a_E\}$ *with* $E$ *attributes and a subset of attributes* $C \subset A$*, the Partial Intrinsic Dimension* $pD(C)$ *is the intrinsic dimension projecting the dataset on the subset* $C$*.*

**Definition 3.6** ***Individual Contribution*** $iC()$***:*** *Given a finite dataset* $A = \{a_1, a_2, \ldots, a_E\}$ *with* $E$ *attributes, the Individual Contribution* $iC()$ *of an attribute* $a_k \in A$ *is the maximum potential contribution of* $a_k$ *to the intrinsic dimension of* $A$*, and it is measured as* $iC(a_k) = pD(\{a_k\}) \rightarrow [0, 1]$*.*

Consider a dataset $A$ and a subset of attributes $C \subset A$ with partial intrinsic dimension $pD(C)$. An attribute $a_i \in (A - C)$ increases the partial intrinsic dimension of $C$ by at most its individual contribution $iC(a_i)$, according to the level of correlation between $a_i$ and the attributes of $C$. If $a_i$ is completely uncorrelated to every attribute in $C$, the partial intrinsic dimension will increase by the individual contribution $iC(a_i)$, i.e., $pD(C \cup \{a_i\}) - pD(C) \cong iC(a_i)$. On the other hand, if $a_i$ is strongly correlated to the attributes in $C$, the partial intrinsic dimension will increase by a value of almost zero,

i.e., $pD(C \cup \{a_i\}) - pD(C) \cong 0$. Additionally, if $a_i$ is weakly correlated to the attributes in $C$, the partial intrinsic dimension will increase by an amount between zero and the individual contribution $iC(a_i)$, i.e., $0 \leq pD(C \cup a_i) - pD(C) \leq iC(a_i)$.

Correlations mean that the value of an attribute can be approximated from other attributes. Sousa et al. (2007b) defined the terms "strong correlation" and "weak correlation". The first one is used when the value of one attribute can be closely deduced from a subset of other attributes, as in linear correlations. A weak correlation indicates that an attribute can be only approximated from other attributes, as in fractal correlations. In order to quantify the correlation among attributes, a threshold $\xi$ ranges from zero – meaning complete correlation – up to one, when the attributes are totally independent.

**Definition 3.7** ***ξ*-Correlation:** *Given a dataset defined on $A = \{a_1, a_2, \ldots, a_E\}$, a subset $B \subset A$ is said to be $\xi$-correlated to a subset $C \subset A, B \cap C = \oslash$ if each attribute $a_i$ in $B$ does not contribute more than $\xi * iC(a_i)$ to the partial intrinsic dimension of $C$.*

**Definition 3.8** ***Attribute Set Core* $\xi C$:** *Given a dataset defined on $A = \{a_1, a_2, \ldots, a_E\}$ with intrinsic dimension $D$, an Attribute Set Core $\xi C$ is a subset of attributes in $A$ such that $|pD(\xi C) - D| < \sum \xi * iC(a_i)$, $\forall a_i \in (A - \xi C)$, and there is no attribute $a_k \in \xi C$ such that $|pD(\xi C) - pD(\xi C - a_k)| < \xi * iC(a_k)$.*

**Definition 3.9** ***Correlation base* $\xi B_p$:** *Given a dataset defined on $A = \{a_1, a_2, \ldots, a_E\}$ and an Attribute Set Core $\xi C \subseteq A$, a Correlation Base $\xi B_p$ is a subset of attributes $\xi B_p \subseteq \xi C$ such that either $\exists a_k \in (A - \xi C) | \exists M_k, M_k(\xi B_p) \rightarrow a_k$ or there are no $\xi$-correlated attributes in the dataset and $\xi B_p = \xi C = A$, where $M_k$ is a mapping indicating that $a_k$ is $\xi$-correlated to all the attributes in $\xi B_p$.*

**Definition 3.10** ***Correlation group* $\xi G_p$:** *Given a dataset defined on $A = \{a_1, a_2, \ldots, a_E\}$ and a Correlation Base $\xi B_p \subseteq A$, a Correlation Group $\xi G_p$ is the subset of attributes $\xi G_p \subseteq A$, such that, $\xi G_p = \xi B_p \cup a_k \in (A - \xi C) | \ |pD(\xi G_p) - pD(\xi G_p - a_k)| < \xi * iC(\{a_k\})$ and $\exists M_k(\xi B_p) \rightarrow a_k$, where $M_k$ is a mapping indicating that $a_k$ is $\xi$-correlated to all the attributes in $\xi B_p$.*

A correlation group $\xi G_p$ includes the correlation base $\xi B_p$ and every attribute $\xi$-correlated to all attributes in $\xi B_p$, but excludes the attributes not $\xi$-correlated to the full correlation base $\xi B_p$. In other words, attributes $\xi$-correlated to some attributes, but not to all attributes in $\xi B_p$ are not in $\xi G_p$.

For example, consider a dataset defined by five attributes $A = \{a_1, a_2, a_3, a_4, a_5\}$ as illustrated in Figure 3.7(a). This dataset has the mappings $M_2(\{a_1\}) \rightarrow a_2$ and $M_5(\{a_1\}) \rightarrow a_5$ as shown in Figure 3.7(b) (Definition 3.9). Then $A$ has the correlation

group: $\xi G_1 = \{a_1, a_2, a_5\}$, with Correlation Base $\xi B_1 = \{a_1\}$. Figure 3.7(c) shows the mapping $M_4(\{a_1, a_3\}) \rightarrow a_4$. Then $A$ has another Correlation Group: $\xi G_2 = \{a_1, a_3, a_4\}$, with Correlation Base $\xi B_2 = \{a_1, a_3\}$. The Attribute Set Core $\xi C = \{a_1, a_3\}$ of $A$ is composed of $\xi B_1$ and $\xi B_2$ as it can be seen in Figure 3.7(d).

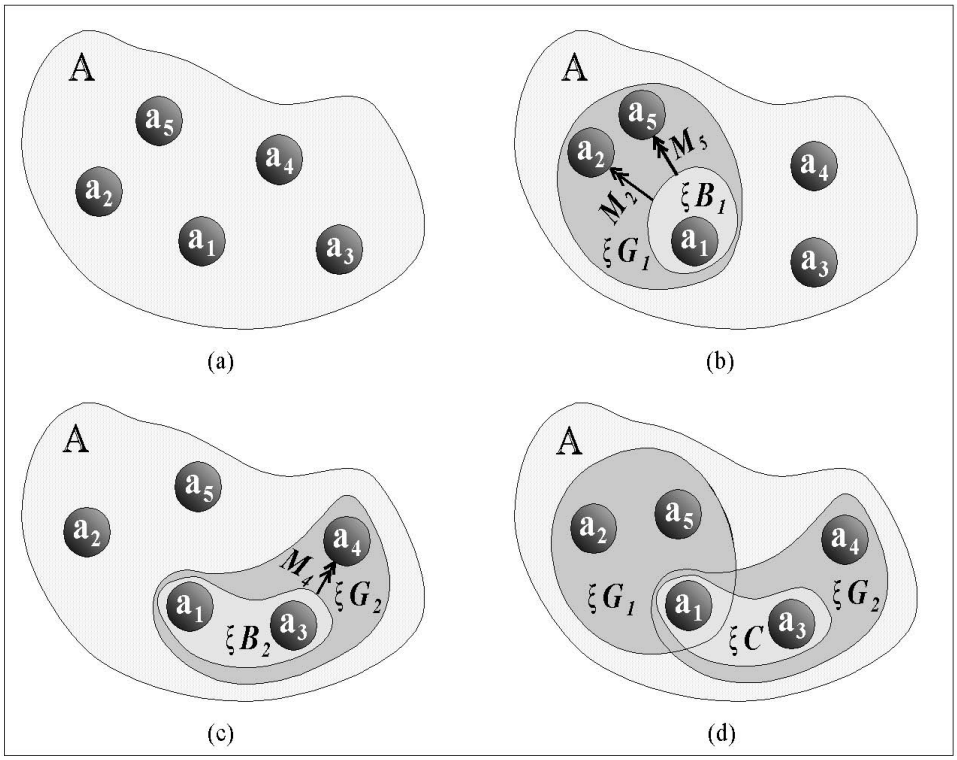



Figure 3.7: Example of correlation groups of a dataset with five attributes (adapted from (Romani et al., 2010b)).

Notice that if a function exists that $f_k(\xi B_p) \rightarrow a_k$, $\xi B_p \subset \xi G_p$, $a_k \in A$, then $a_k \in \xi G_p$ and $pD(\xi G_p) = pD(\xi G_p - a_k)$. However, correlations are not limited to functions. In fact, any mapping $M_k(\xi B_p) \rightarrow a_k$ such that $|pD(\xi B_p \cup a_k) - pD(\xi Bp)| < \xi * iC(a_k)$ defined as a $\xi$-correlation, so for every attribute $a_k \in G_p$, $a_k \notin \xi B_p$ we have:

$$|pD(\xi G_p) - pD(\xi G_p - a_k)| < \xi * iC(a_k) \tag{3.3}$$

Therefore, an attribute $a_i$ in a Correlation Group but not in the Correlation Base does not increase the partial intrinsic dimension of the group by more than $\xi * iC(a_i)$, that is $|pD(\xi G_p) - pD(\xi B_p)| < \sum \xi * iC(a_i)$, $\forall a_i \in (\xi G_p - \xi B_p)$.

Intuitively, FD-ASE first calculates the intrinsic dimension of a dataset through the LiBOC algorithm. Afterwards, FD-ASE measures the partial intrinsic dimension considering incremental sequences of variables. For instance, FD-ASE first calculates $D$ considering only the first variable in the dataset. After, it calculates $D$ considering the first and the second variable in the dataset and so on. The FD-ASE algorithm uses the variation

in the partial intrinsic dimension as a criterion to identify groups of correlated variable. In addition, FD-ASE uses the partial intrinsic dimension to select a relevant attribute subgroup to represent the essential data characteristics. For example, let be a dataset of three attributes $a_1$, $a_2$ and $a_3$ that represents a line. The embedding dimension of this dataset is three ($E = 3$) and its intrinsic dimension is equal to one ($D = 1$). Executing the FD-ASE algorithm, the Correlation Group $\xi G_1 = \{a_1, a_2, a_3\}$ and the Correlation Base $\xi B_1 = \{a_1\}$ are found. It means that $a_1$ is correlated to $a_2$ and $a_1$ is also correlated to $a_3$. FD-ASE also generates an Attribute Set Core $\xi C = \{a_1\}$, indicating that only the attribute $a_1$ is enough to represent the dataset.

This approach allows spotting the attributes that define others, as well as how strong their correlations are. Therefore, the analyst of the database can drop the attributes that are not meaningful and save memory space as well as time processing when managing and querying the data. The detection of correlation groups is a powerful tool to understand the relation among variables from meteorological stations and those extracted from remote sensing images. In Chapter 5, we will demonstrate the applicability of the FD-ASE method in the agrometeorologic domain.

## 3.4 Data Stream Monitoring through SID-meter

Intuitively, a data stream is a flow of data items ordered (explicitly by a time stamp or implicitly by order of arrival to the system that handles it), potentially unbounded and usually generated in real time.

**Definition 3.11** ***Data stream*** *is an ordered sequence of events (or items)* $\{ev_1, ev_2, \ldots, ev_n, \ldots\}$ *in which an event* $ev_j$ *is defined by a set of* $E$ *measured attributes* $a_i$*, such that each* $ev_j = (a_1, \ldots, a_E)$ *(Sousa et al., 2007a).*

In general, data streams are characterized by large amounts of data generated in synchronous or asynchronous processes potentially infinite. Another characteristic quite common in applications involving data streams is the evolution, i.e., trends of the data undergo significant changes over time.

These changes may mean temporary events (for example, a week of freezing temperatures caused by an isolated wheather event) or relevant changes in the process of generating the stream that result in variations in the distribution of data (for instance, climate change caused by factors, such as global warming). The identification of these tendency variations in evolving data streams is crucial in some types of applications, such as monitoring of climate variations, monitoring of industrial processes, systems of fraud detection in credit card, among others.

The intrinsic dimension estimated by fractal dimension can be used in algorithms for analyzing and processing data streams. The variation of the intrinsic dimension over time into a data stream reflects changes in trends, such as changes in the distribution of the data or in the correlations among attributes that define the stream. For example, in a data stream defined by quotations of currencies based on U.S. dollars, changes in local or international financial market and exchange policies may alter the correlations between currencies, as well as increase and decrease the number of correlated currencies.

In the literature, several authors work on analysis of behavior changes of evolving data (Aggarwal, 2003; Kifer et al., 2004; Papadimitriou et al., 2004), burst detection (Kleinberg, 2003; Zhu & Shasha, 2003), classification (Aggarwal et al., 2004b; Gama et al., 2005; Ferrer-Troyano et al., 2006; Aggarwal & Yu, 2008), clustering (Guha et al., 2003; Aggarwal et al., 2004a; Rodrigues et al., 2008), frequent items identification, data streams maintenance and processing (Jin et al., 2003; Manjhi et al., 2005; Sakurai et al., 2007).

The general idea of using the intrinsic dimension as a tool to monitor evolving data streams is to continuously measure $D$ over time in order to detect significant variations of successive values of $D$ and, consequently, identify meaningful behavior changes. An approach to measure the intrinsic dimension of data stream was proposed by Sousa et al. (2007a). The authors presented the algorithm SID-meter (*data Stream Intrinsic Dimension meter*), which considers a data stream as a sequence of events $\{ev_1, ev_2, ..., ev_n\}$, each of which represented by an array of $E$ measurements. The events occurring within a time interval are considered as a dimensional dataset of dimension $E$. The fractal dimension $D_2$ is thus used to estimate the intrinsic dimension $D$ of successive sub-sequences of events.

SID-meter applies an event-based sliding window divided into $n_c$ sequential periods, named *counting periods*. Each period can process events arriving during a given time or a predefined number of incoming events, i.e., $n_i$ events are processed in each counting period. When a counting period is completed, the events of the oldest one are discarded. Therefore, $n_c$ and $n_i$ respectively specify the length and the movement step of the measuring window. Figures 3.8(a) and 3.8(b) illustrate successive sliding windows, divided into four counting periods, through a data stream composed of the attributes $a_1, a_2$ and $a_3$.

When a counting period finishes, a new value of $D$ is computed considering the events in the current window. The value of $D$ is based on the count of events inside the whole window, following Equation 3.2. Thus, SID-meter continually measures $D$ for successive windows, tracking the stream behavior over time.

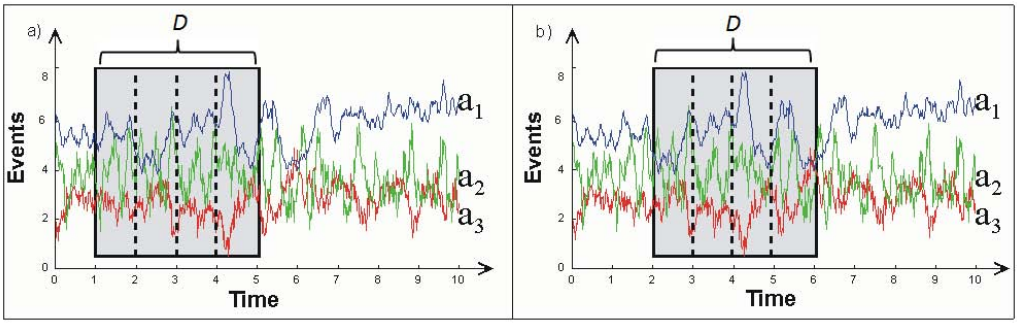



Figure 3.8: Counting periods of a sliding window (adapted from (Romani et al., 2009a)).

## 3.5 Summary

In this chapter, we presented concepts from the fractal theory that have been explored with promising results to analyze the behavior of real datasets. Several studies have shown that, in general, real data have the property of self-similarity and thus can be modeled as a fractal. In particular, the correlation fractal dimension has been successfully used as a tool for analysis of data distribution in the space of attributes, especially in data mining.

We also described two methods that use the intrinsic dimension estimated by fractal dimension: FD-ASE in Section 3.3 and SID-meter in Section 3.4. FD-ASE identifies correlation groups that define which attributes are correlated and which of them best represent each correlation found (correlation base). SID-meter allows the monitoring process of data streams through measurement of intrinsic dimension in different windows indicating changes in the data flow.

The atmospheric conditions represented in the meteorological measurements that comprise the historical series used in this work are complex dynamic systems whose analyses can benefit from fractal theory. Thus, the methods briefly discussed in this chapter (FD-ASE and SID-meter) were used as part of some proposed methods for analysis of the climate and remote sensing series. These new methods and the experimental results are presented in Chapter 5 of this thesis.

# Chapter 4

# Data and Time Series Mining

## 4.1 Introduction

In recent years, improvements in the data acquisition technology have decreased the time interval of data gathering, bursting the quantity of the produced data. In addition, storage capacity of databases has also increased generating huge amounts of data that exceed our human ability for comprehension without the support of data analysis tools. Therefore, there is imminent and constant need for developing of new algorithms, techniques and methods to aid specialists from different fields to analyze the enormous volume of data in order to discover useful information and knowledge.

Consequently, several methods have been developed in order to mine large amounts of data to discover data patterns contributing to knowledge bases, business strategies, and scientific research. Data Mining is an interdisciplinary field that combines a set of disciplines including database systems, statistics, machine learning, visualization, information retrieval, pattern recognition, image analysis, and others. Complementary, information extracted through data mining techniques may be used in a variety of applications, such as business analysis, production control, biomedical systems, climate change, and agriculture. One topic of Data Mining field that focuses on discovering patterns, specifically in historical series of data, is called Time Series Mining.

In this doctorate thesis, data mining techniques were employed to analyze climate data and remote sensing images. First of all, data mining techniques were used to optimize similarity searching in time series in an efficient and effective way. In a second step, time series mining was employed to transform time series in a symbolic representation to discover association patterns among heterogeneous series. Therefore, data mining is one of the pillars of this thesis highlighting the association rule mining as the main task of data mining that was explored in this work. Additionally, other tasks of data mining were also

used, such as feature selection, discretization, data transformation, and clustering. An overview of discretization and data transformation concepts is presented in this chapter.

This chapter is organized as follows. Section 4.2 presents the concept of Knowledge Discovery in Databases (KDD) and the main Data Mining (DM) tasks. Section 4.3 shows the steps of time series mining process. Some important methods of preprocessing phase are detailed in Section 4.4. The association task is defined in Section 4.5. Finally, the distance functions applied to similarity search in time series are discussed in Section 4.6.

## 4.2 The KDD process

Knowledge Discovery in Databases or simply KDD is an interactive sequence of steps with the purpose of discovering useful information and knowledge from data. One of the most used definitions for the KDD concept was proposed by Fayyad et al. (1996) "*KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*".

The KDD process refers to steps in which knowledge is extracted from data, as it can be seen in Figure 4.1. In the process, data mining corresponds to an essential step where methods are applied in order to find data patterns. Due to the importance of this step in the KDD process, the term Data Mining is actually used to summarize the whole process (Han & Kamber, 2001). In general, the knowledge discovery process means finding patterns in data, in an interactive and iterative way, through execution of algorithms followed by analysis of their results.

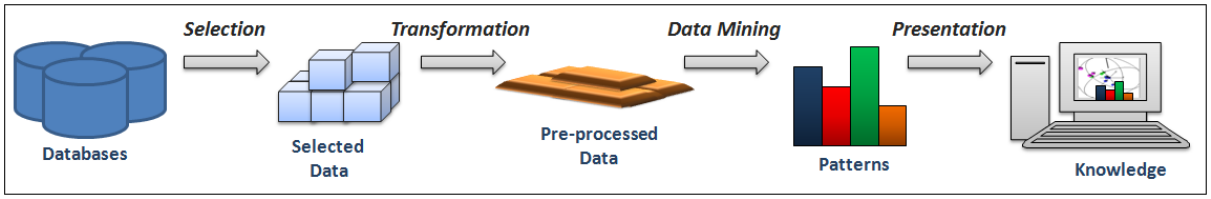



Figure 4.1: An overview of the steps in the KDD process (adapted from (Han & Kamber, 2001)).

According to Fayyad et al. (1996), the major steps of KDD process are:

1. *Data Cleaning*: Real-world databases often have incomplete, noisy and inconsistent data that can damage the data analysis hindering the patterns detection. Data cleaning routines work to prepare the data to the next steps in a KDD process by filling in missing values, smoothing noisy data, identifying and removing outliers, and resolving inconsistencies.

2. *Data Integration*: In some circumstances, data from multiple sources must be merged and transformed into appropriate forms to be included in the same analysis. Thus, this step involves techniques to correctly integrate multiple databases, data cubes, or files in a data warehousing.

3. *Data Selection*: This step corresponds to the identification of relevant data to the analysis task and their retrieving from the databases.

4. *Data Transformation*: In this step, data are transformed or consolidated into suitable forms for mining using operations, such as summarization, aggregation, generalization, or normalization.

5. *Data Mining*: Core step in the KDD process, where computational techniques are applied to extract unknown and useful patterns from the data.

6. *Pattern Evaluation*: In this step, interestingness measures are used to identify the truly interesting patterns representing knowledge.

7. *Knowledge Presentation*: Visualization and knowledge representation techniques are employed to present the mined knowledge to the users.

Data mining functionalities define the type of patterns to be discovered in data mining tasks. These tasks consist of a set of techniques, procedures, and algorithms used to extract patterns from the data. Usually, data mining tasks are subdivided in two categories: *predictive* and *descriptive*. Predictive mining tasks construct models on current data in order to make predictions. Descriptive mining tasks reveal patterns and properties over analyzed data. The major data mining tasks are:

- *Association*: Algorithms in this class are designed to find relationships among items in a database.

- *Classification*: This task is composed of techniques to predict the class of a new object.

- *Clustering*: Algorithms to group similar objects, following a given criterion.

- *Summarization*: In general, techniques based on statistics or aggregation are used to summarize the data.

- *Outliers detection*: Algorithms that search for the objects that do not follow a standard data/trend behavior.

This thesis proposes to employ and to develop data mining techniques to discover relevant patterns in climate and remote sensing data. As this kind of database is composed of long historical series of data, methods and algorithms to mine time series were proposed. In the next section, the time series mining process is detailed.

## 4.3 Time Series Mining

A time series is any set of observations ordered in a period of time. In other words, a time series consists of a sequence of values changing with time. These measures are gathered at equal time intervals. Time series have been used in several fields for instance agriculture, economy, geo-physics, meteorology, etc. There are different reasons to study time series, such as to investigate mechanisms responsible for generating this kind of data, to forecast values in time series for a short and long time, to describe trends in series, and to search for important periodicities in the data (Morettin & Toloi, 2006; Wei, 2006).

Time series mining brings the complexity of dealing with time series to the data mining field. Important tasks of time series mining include trend analysis, similarity search, as well as mining sequential and periodic patterns. The complexity can increase whenever the original data must be pre-processed to build the times series. For example, time series can be built from measurements taken from periodical satellite images. Moreover, different tasks in data mining require the data to be in the frequency domain. Thus algorithms of data transformation are essential to transform time series to a suitable format in order to ease the use of data mining algorithms.

In general, a time series mining process involving climate and remote sensing data is composed of a set of five steps, as illustrated in Figure 4.2 and described in Table 4.1.

Table 4.1: The steps of time series mining process.

| **Step** | **Objective** |
|---|---|
| Extraction | extract time series from images and other kind of complex data |
| Pre-processing | remove errors and noise, and transform the data |
| Integration | integrate different time series from several sources |
| Mining | apply algorithms to discover patterns and knowledge |
| Presentation | present the mined knowledge |

The most important phase in the time series mining process is the preprocessing, since time series need to be converted into discrete intervals or symbolic sequences for some mining tasks. In general, the same steps of KDD process can be defined in the time series mining. However, the data mining tasks have some particularities in their definitions, which are described as follows.

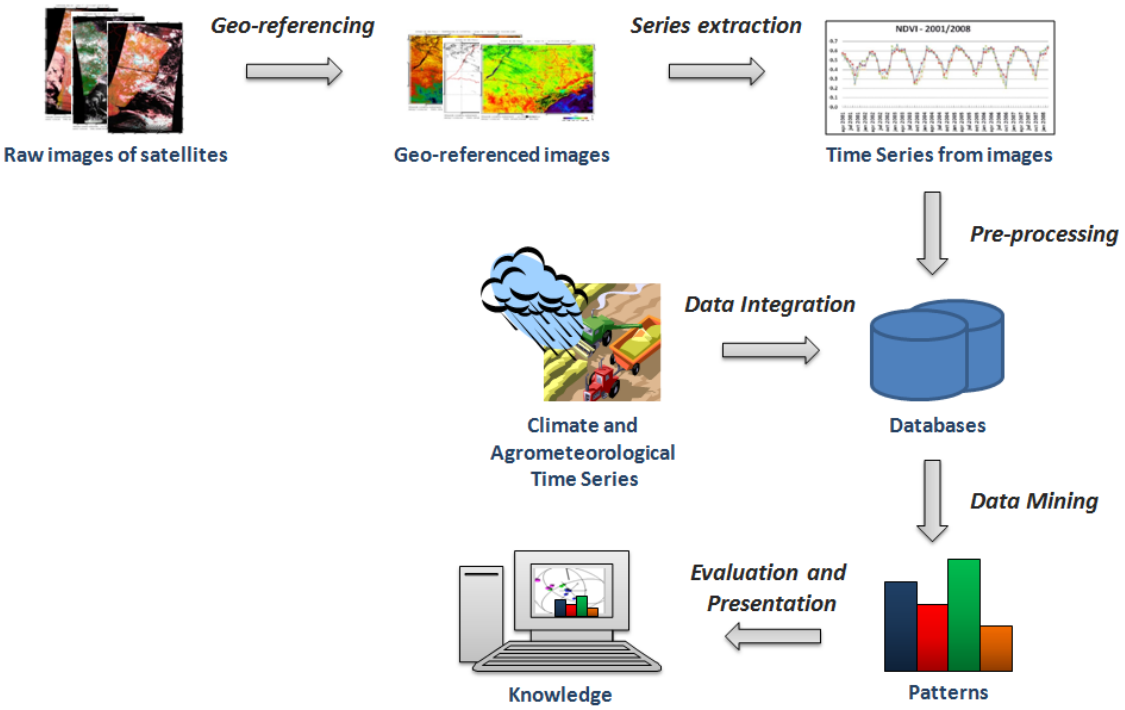



Figure 4.2: An overview of the steps of time series mining process adapted to the context of this doctorate thesis.

- *Similarity Search*: Algorithms to find the most similar time series considering a query time series and some similarity measures.
- *Rule discovery*: Task related to the problem of finding rules relating patterns in a time series or in different time series.
- *Clustering*: Task aimed at finding groups of time series according to some similarity measure or criterion.
- *Classification*: Task that assigns an unlabeled time series to one or more predefined classes.
- *Summarization*: Task that creates an approximation of a given time series that retains its essential features.
- *Anomaly Detection*: Techniques to find parts of time series that contain anomalies or unexpected/novel behavior.

## 4.4 Data Preprocessing

According to Zhang et al. (2005), 80% of the work in Knowledge Discovery process is concentrated in the preprocessing phase, since real-world databases are susceptible to

noisy, missing and inconsistent data. Incomplete data can occur for different reasons, such as: attributes of interest are not always available, data may not be included because they were not considered important when the database was created, relevant data may not be registered due to misunderstanding or because of equipment malfunctions, and other reasons (Han & Kamber, 2001).

In general, noisy data occurs mainly due to failures in data collection instruments, human or computer errors during data entry, technology limitations, such as limited buffer size for synchronized data transfer and consumption. Moreover, incorrect data may result from inconsistencies in naming conventions and data codes used (Han & Kamber, 2001).

As a solution, data preprocessing techniques may be used. For instance, *data cleaning* can be applied to correct inconsistencies and remove noise from the data. *Data integration* is used to merge data from different sources into a consistent database. *Data transformations*, such as normalization methods, may improve the accuracy of mining algorithms. Aggregating, eliminating redundant attributes or clustering to reduce the data size are examples of *data reduction* techniques.

The main purpose of the preprocessing phase is improving the overall quality of the mined patterns. In this doctorate work, different preprocessing techniques were used to prepare data to be submitted to the proposed and extended mining methods. Some techniques for data pre-processing, which were employed in this work are described in details in the next sections.

### 4.4.1 Discretization techniques

The most common types of attributes used in data mining are nominal (categorical), continuous, and discrete. The nominal attributes only assume a limited number of values without a relationship of order among the values. An example of categorical attribute is weather condition, such as: sunny, cloudy and rainy. On the other side, continuous attributes are composed of an infinite number of values with a relation of order among them. The value of maximum temperature is an example of a continuous attribute. Discrete attributes have a reduced number of values when they are compared to continuous attributes. These attributes also preserve the relation of order between values. The mapping process of continuous attributes into discrete attributes is called *discretization*.

Discretization techniques are used to reduce the number of values for continuous attributes by dividing the range of attributes into intervals that can be used to replace data values (Han & Kamber, 2001). The objective of the discretization algorithms is to determine the best set of cut points to be used to convert continuous into discrete data. A cut point is a threshold in a real values interval. Occasionally, the discretization process

is a mandatory step of a mining algorithm that only works with nominal or discrete data. Thus, a discretization algorithm is necessary whether the real application needs to deal with continuous data.

One disadvantage of using discretization algorithms is the loss of precision of values and loss of information that occur during the discretization process that can lead the data mining algorithms to have tortuous results. Despite the loss of information that is inherent in the discretization process, several works report a meaningful increasing in accuracy and execution time of the mining algorithms when a proper discretization technique is used in the preprocessing phase (Kurgan & Cios, 2004).

The discretization methods can be supervised, i.e. using the class information of the instances (examples) to promote discretization, and not-supervised when they do not use this information. The simplest discretization methods are *equal-width* (ranges of values with the same length) and *equal-frequency* (intervals with the same instances number).

The method 1R is an improvement from the equal-width method, where thresholds of intervals (cut points) are adjusted according to information about instances class (Holte, 1993). Kerber (1992) proposed the *ChiMerge* algorithm that uses the statistical test $\chi^2$ to determine when consecutives intervals must be clustered. Algorithms that accomplish both discretization and feature selection tasks have been proposed in the last years. One example is the *Chi2* algorithm (Liu & Setiono, 1995) that is an upgrading of the ChiMerge algorithm.

*Omega* is a supervised algorithm for feature selection and discretization (Ribeiro et al., 2008). First, Omega sorts the continuous values and defines the initial cut points, where each cut point is a limit of an interval (bin) of real values. When a change in the class label of the instances occurs, a cut point is created. Figure 4.3 shows a simple example of Omega execution with eight real values.

Omega fixes the minimum frequency for a bin, avoiding a great number of cut points in the second step. The algorithm eliminates the right cut points of the intervals that do not satisfy the minimum frequency restriction given by an input parameter $H_{min}$ as it can be seen in Figure 4.3. Thus, if the value of $H_{min}$ is high, the bins obtained in this step are few.

In the third step, consecutive intervals with the same majority class and with an inconsistency rate smaller than the maximum inconsistency threshold ($\zeta_{max}$) are fused. The majority class is the most frequent class in an interval. An inconsistence is an occurrence of a class different from the majority class in an interval. Let $M_{T_i}$ be the majority class of an interval $T_i$. Equation 4.1 gives the inconsistence rate $\zeta_{T_i}$ of an interval $T_i$.

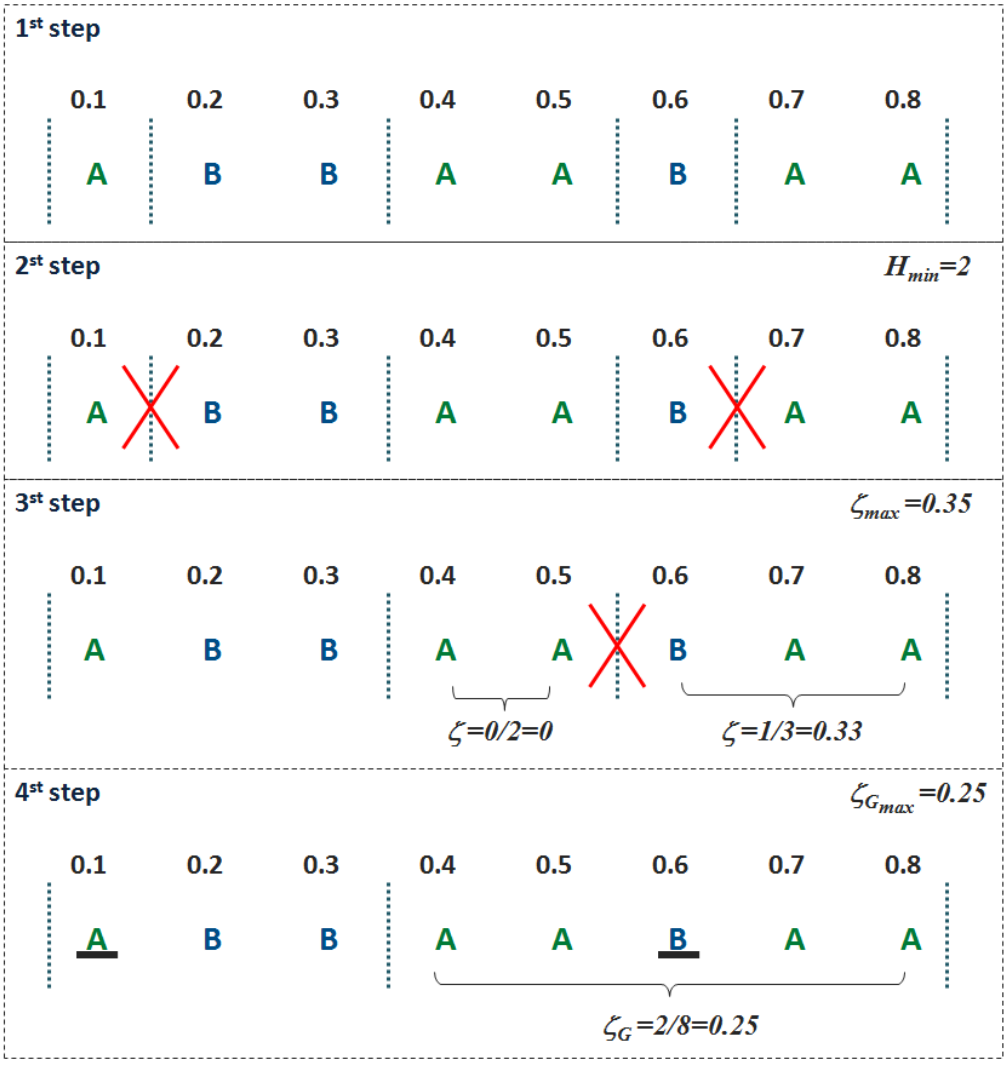



Figure 4.3: Example of the Omega execution. The letters $A$ and $B$ are the class information provided.

$$\zeta_{T_i} = \frac{|T_i| - |M_{T_i}|}{|T_i|} \tag{4.1}$$

where $|T_i|$ is the number of instances in the interval $T_i$ and $|M_{T_i}|$ is the number of instances of the majority class in $T_i$.

Figure 4.3 shows an example of the cut point that is removed in the third step using $\zeta = 0.35$. The inconsistence rates $\zeta_{T_i}$ correspondent to the second and third interval are $\zeta_{T_2} = 0/2$ and $\zeta_{T_3} = 1/3 = 0.33$, respectively. As $T_2$ and $T_3$ have the same majority class, i.e. $M_{T_2} = M_{T_3} =$ "A" and $\zeta_{T_2} \leq \zeta_{max}$ and $\zeta_{T_3} \leq \zeta_{max}$, the second and third intervals are fused.

In the last step, Omega performs the feature selection. Let $T$ be the set of intervals in which a feature $f$ is discretized. For each feature, the algorithm computes the global inconsistence $\zeta_G$, according to Equation 4.2.

$$\zeta_G = \frac{\sum T_i \in T(|T_i| - |M_{T_i}|)}{\sum T_i \in T(|T_i|)} \tag{4.2}$$

Every attribute whose global inconsistence value is greater than an input threshold $\zeta_{G_{max}}$ is removed from the set of attributes. Figure 4.3 shows the end cut points determined by the Omega algorithm to the feature $f$. From eight instances, only two have classes different from the majority class in their intervals. Thus, the global inconsistency of the feature $f$ is $\zeta_G = 2/8 = 0.25$. If $\zeta_G \leq \zeta_{G_{max}}$, the feature $f$ is selected, otherwise it is eliminated from the feature vector.

The Omega algorithm is used as a module in a new method proposed in this work to mine association rules from time series, because this algorithm reached better results when it was compared to the 1R, ChiMerge and Chi2 algorithms (Ribeiro et al., 2008). This new method is presented in details in Chapter 5.

### 4.4.2 Time Series Representation

One reason to represent time series by symbolic data is to provide a more concise way to spot the major characteristics of the data. In addition, other important aspects must be considered, such as data compression, processing speed up, and noise removal. In general, several techniques for time series analysis require data in a frequency domain. Distance-preserving orthonormal transformations are often used to transform data from the time domain to the frequency domain. However, the appropriate choice of representation is very important since this process affects the efficiency of time series mining (Han & Kamber, 2001). Accordingly, a large number of time series representations have been proposed as illustrated in Figure 4.4, that shows a hierarchy inspired by Lin et al. (2003). We can classify these techniques in two groups: data adaptive and non data adaptive.

Some examples of non data adaptive techniques are Discrete Wavelet Transform (Chan & Fu, 1999), Discrete Fourier Transform (Faloutsos et al., 1994), Discrete Cosine Transform (Korn et al., 1997), Chebyshev polynomials (Cai & Ng, 2004), and Piecewise Aggregate Approximation (Keogh et al., 2001a). Other techniques are data adaptive, such as Piecewise Linear Approximation (Chen et al., 2007a), Adaptive Piecewise Constant Approximation (Keogh et al., 2001b), Single Value Decomposition (Faloutsos et al., 1994), Symbolic Aggregate approXimation (Lin et al., 2007), and Multi-resolution Vector Quantized (Wang et al., 2010). An overview of some techniques for data transformation, where $n$ is the original dimensionality of the data and $N$ is the reduced dimensionality of the data, is presented as follows.

The *Discrete Fourier Transform (DFT)* represents time series as a linear combination of sine and cosine functions keeping only the first $n/2$ coefficients, because each sine

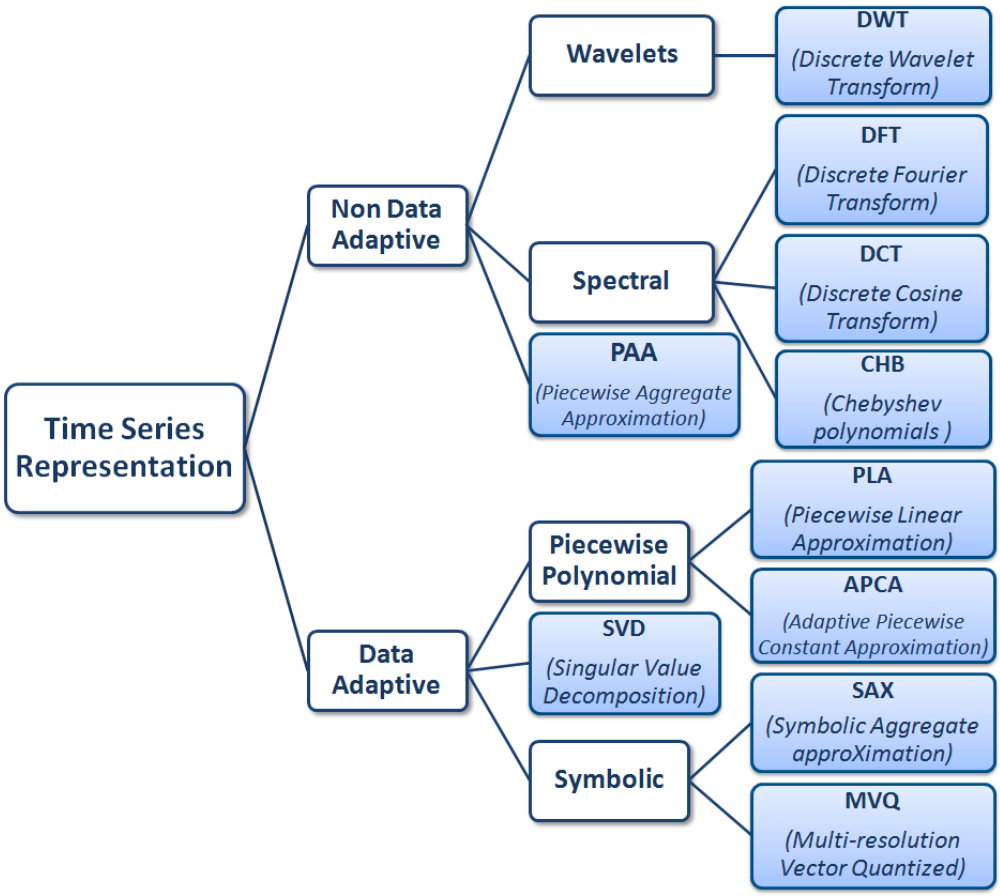



Figure 4.4: A hierarchy of various time series representation, where the techniques are highlighted in blue.

function requires two numbers for the phase ($w$) and amplitude ($A, B$) (Agrawal et al., 1993a; Faloutsos et al., 1994). Figure 4.5(a) presents an example of the application of DFT to transform a time series. DFT has an ability to compress most natural signals, which is an advantage. Moreover, there is a fast algorithm, called Fast Fourier Transform (FFT) (Winograd, 1976; Gonzalez & Woods, 1992), that calculates the DFT coefficients in $O(nlogn)$ time. However, this technique has some problems, such as its difficulty in dealing with sequences of different lengths.

The *Discrete Wavelet Transform (DWT)* method represents time series as a linear combination of Wavelet basis functions keeping the first $N$ coefficients. There are many types of wavelets (Chan & Fu, 1999; Popivanov & Miller, 2002). However, researchers in indexing and time series mining usually use the *Haar* wavelets. In addition, Haar wavelets is also simple to implement. An example of applying this technique can be seen in Figure 4.5(b). DWT has a good ability to compress stationary signals. Algorithms for DWT can be executed in linear time that is also positive.

Although there are many different types of wavelets, researchers in time series min-

ing/indexing generally use Haar wavelets.

Similarly to DFT and DWT, *Singular Value Decomposition (SVD)* represents time series as linear combination of *eigenwaves* keeping the first $N$ coefficients (Korn et al., 1997; Keogh et al., 2001b). However, SVD differs from the other methods because the *eigenwaves* are data dependent. An SVD example is shown in Figure 4.5(c). This is an optimal technique for dimensionality reduction although computationally expensive.

Another method, *Piecewise Linear Approximation (PLA)* represents time series as a sequence of straight lines that could be connected, and it is allowed $N/2$ line segments. When lines are disconnected, it is allowed only $N/3$ lines (Shatkay & Zdonik, 1996; Morinaka et al., 2001). This technique is usually employed by either performing interpolation or regression. Figure 4.5(d) presents an example of PLA transformation. There are fast linear time algorithms for PLA, which have an ability to compress natural signals (Morinaka et al., 2001).

The *Piecewise Aggregate Approximation (PAA)* method represents a numeric time series as a sequence of box basis functions, where each box has the same length. Given the reduced dimensionality representation, it is possible to calculate the approximate Euclidean distance (Yi & Faloutsos, 2000; Keogh et al., 2001a). An example of applying this method can be seen in Figure 4.5(e). This technique is fast to calculate.

The *Adaptive Piecewise Constant Approximation (APCA)* is a generalization of PAA that allows the piecewise constant segments to have arbitrary lengths (Geurts, 2001; Keogh et al., 2001b). As time series have little details in some parts and high details in others, APCA can fit itself to the data getting a better approximation. An APCA example is presented in Figure 4.5(f). There is a fast algorithm for APCA with linear complexity (Keogh et al., 2001b). However, the implementation of this technique is complex.

The *Symbolic Aggregate approXimation (SAX)* is a symbolic representation of time series that allows time series of length $n$ to be reduced to a string of length $w$ where $w < n$ (Lin et al., 2003). The time series is transformed using the PAA representation and after that the PAA representation allows to build a symbolic discrete string. The major feature of this technique is the lower bounding approximation to the Euclidean distance, which is useful for indexing.

The *Multi-resolution Vector Quantized (MVQ)* approximation is a representation for time series similar to DWT, but it keeps both local and global information about the data (Wang et al., 2010). The main characteristics of this method are: a "vocabulary" of subsequences is discovered using time-tested "vector quantization" methods; it considers multiple resolutions that improves the accuracy, and it uses text-based techniques from information retrieval to weigh down uninteresting matches in order to provide a new distance metric. The MVQ method maintains high-level feature information instead of

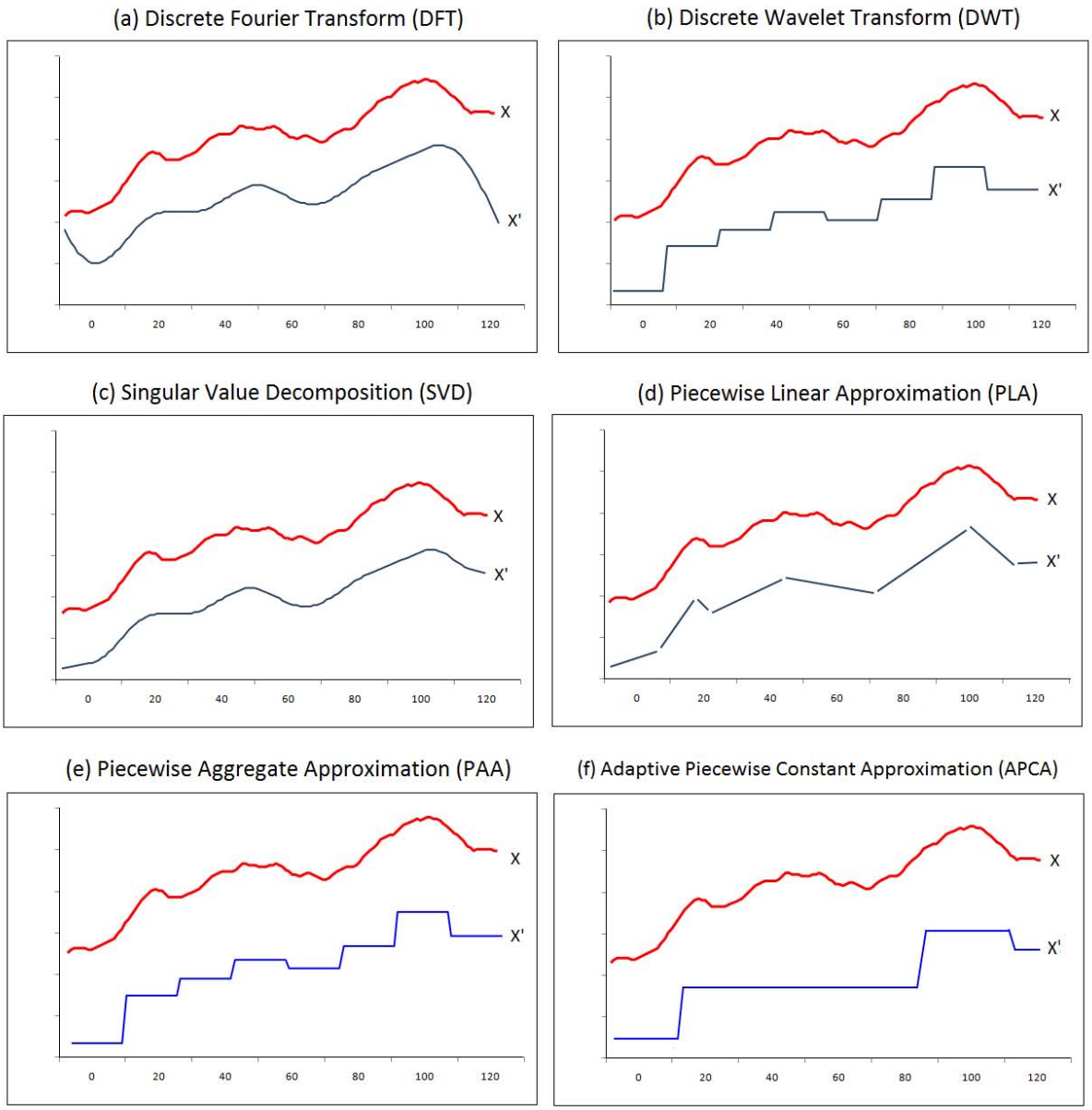



Figure 4.5: Examples of the most used representations for time series mining where lines in red represent the original signal and lines in blue correspond to the transformed signal (adapted from (Keogh, 2001)).

keeping low-level time series values. This aspect makes it easy to introduce more significant similarity measures. MVQ uses a multi-resolution distance function and scales linearly with the database size and dimensionality.

## 4.5 Association Rules

Association rule mining aims at finding association or correlation relationships among a set of data items. The association task is largely used due to its applicability and easy comprehension of patterns generated by this task. Association rules were first proposed

by Agrawal et al. (1993b) to solve the problem of discovering which items occur together in a transaction.

Let $I = \{i_1, \ldots, i_n\}$ be a set of literals called *items*. A set $X \in I$ is called *itemset*. An itemset $X$ with $k$ elements is defined as *itemset-k*. Let $R$ be a table with transactions $t$ involving elements that are subsets of $I$. The transaction $t$ supports an itemset $X$, if $X \in t$. An association rule is an expression of the form $X \rightarrow Y$, where $X$ and $Y$ are itemsets. $X$ is called body or antecedent of the rule, and $Y$ is called head or consequent of the rule. An association rule $X \rightarrow Y$ can be translated as "if $X$ then $Y$" indicating that when $X$ occurs $Y$ also occurs. Support *sup* and confidence *conf* measures, described respectively in Equations 4.3 and 4.4, are used to determine the rules returned by the mining process, where |R| is the number of transactions in relation R.

$$sup(X \rightarrow Y) = \frac{|X \cup Y|}{|R|} \tag{4.3}$$

$$conf(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \tag{4.4}$$

The problem of mining association rules, as it was first stated, involves finding rules that satisfy the restrictions of *minimum support* and *minimum confidence* specified by the user.

The support of an itemset $X$ is the ratio between the number of transactions in $R$ that support $X$ and the total number of transactions of $R$. *Support* is used as a restriction over itemsets frequency to mine the rules. An itemset $X$ is called *frequent itemset* if the support of $X$ is greater or equal to the minimum support specified by the user. *Confidence* of a rule $X \rightarrow Y$ is the ratio between the number of transactions that contains $X$ and $Y$, and the number of transactions that contains $X$.

A well-known example of association rule involving data from a market basket is "70% of shopping that contains diaper also contains beer and 4% of all shopping contains both items". In this example, 70% is the rule confidence and 4% is the rule support. Typically, association rules are considered strong and interesting when they satisfy both a minimum support threshold and a minimum confidence threshold.

The support measure has the monotone property, which means that all nonempty subsets of a frequent itemset must also be frequent. This property, also called Apriori property, is used to reduce the search space improving the efficiency of the level-wise generation of frequent itemsets.

When a database is composed of time series, the problem of association rules mining may be divided into two categories: mining of single time series and multiple time series. Multiple time series may be mined through traditional association rules algorithms. As-

sociation rule mining for single time series is considered a problem of mining sequential patterns or sequence mining. Algorithms for association rules mining as well as sequential patterns mining are presented in the next sections.

### 4.5.1 Algorithms for association rules mining

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called *strong association*. Association rule mining is a process divided in two steps (Agrawal & Srikant, 1994):

- Find all frequent itemsets,
- Generate strong association rules from the frequent itemsets.

The critical phase of association rules mining is the determination of frequent itemsets. The next phase of rules generation from the frequent itemsets is common for most of the algorithms. The first algorithms to determine frequent itemsets were AIS (Agrawal et al., 1993a) and SETM (Houtsma & Swami, 1993). The most used algorithm for association rules mining is Apriori, which was proposed by Agrawal & Srikant (1994).

#### The Apriori algorithm

Apriori uses the monotone property to prune rules. The pseudo-code of Apriori is described in Algorithm 2, where $L_k$ is a set of frequent itemsets with length $k$ ($k$-*itemsets* that satisfy the minimum support threshold *minsup*). $C_k$ is a set of candidates itemsets of length $k$ ($k$-*itemsets* potentially frequent).

**Algorithm 2** Apriori Algorithm

**Data**: Table with transactions $t$, minimum support *minsup*
**Result**: Set of frequent *itemsets*

```
1 L_1 = {frequent itens}
2 for (k = 1; L_k ≠ ∅;k++) do
3 |   C_{k+1} = new candidates generated from L_k
4 |   for each transaction t in database do
5 |   |   Increase the count for all candidates in C_{k+1} that are included in t.
6 |   end
7 |   L_{k+1} = candidates in C_{k+1} that satisfy minsup
8 end
9 return ∪_k L_k
```

In the first interaction (line 1) the algorithm scans all transactions in order to count the number of occurrences of each item and determines $L_1$ (set of frequent 1-itemsets). In

lines 2 to 8 is determined $L_k$ (set of frequent *k-itemsets*). The set $L_k$ is used to generate $C_{k+1}$ (set of candidates *k-itemsets*). Thus, the itemset candidates $C_{k+1}$ are generated through a join of $L_k$ with itself according to the condition that $k-1$ items from the join data are the same data as it can be seen in line 4. Next, Apriori verifies for each generated itemset if it has a subset of non-frequent itemsets. If it is true, the itemset is excluded from the set of candidate itemsets, on contrary it is added to $C_{k+1}$. On lines 4-6 is computed the count of each candidate *k-itemset*, where the counter is incremented by one for each transaction in which it appears. Finally, only the candidates *k-itemset* that have support greater or equal than the minimum support threshold are added to $L_k$ and returned.

Figure 4.6 illustrates the Apriori algorithm for finding frequent itemsets in R. First, each item is a member of the set of candidate 1-itemset, $C_1$. The algorithm scans all of the transactions to count the number of occurrences of each item. Suppose that the minimum support required is 40% (2). Then, the set of frequent 1-itemset, $L_1$, can be determined, eliminating all candidate that not satisfy support $\geq$ min_sup.

The algorithm generates a candidate of 2-itemsets, $C_2$, applying the joint operation in $L_1$ ($L_1 \bowtie L_1$). Next, transactions in R are scanned and the algorithm accumulates the support count of each candidate itemset in $C_2$, as it can be seen in the second row of Figure 4.6. The algorithm then generates the set of candidate 3-itemsets, $C_3$, as detailed in Figure 4.6. Based on the Apriori property that all subsets of a frequent itemset must also be frequent, some candidates are eliminated a priori and they do not appear in table of $L_3$. As $C_4 = \oslash$ due to the application of the Apriori property, the algorithm ends, having found all of the frequent itemsets. Once the frequent itemsets from transactions in the database $R$ have been found, the association rules from them are generated since they satisfy the minimum confidence.

The phase in data mining, which requires more processing is the determination of frequent itemsets. New algorithms were developed to make this phase more efficient, among them stand out the algorithms: *Partition* (Savarese et al., 1995), *FP-Growth* (Han et al., 2000) and *Eclat* (Zaki et al., 1997).

A problem of association rules mining is the large number of rules generated. Yamamoto et al. (2008) presented techniques of itemsets visualization, allowing visual analysis of the itemsets, granting the user to select the more interesting itemsets that appear in the rules.

Another very important issue in association rules mining refers to the *measure of interest* to be used. A *measure of interest* is the importance degree of a rule. It defines, which rules will be returned by the mining algorithm. In addition to the most used measures of interest - support and confidence - other measures of interest: all-confidence

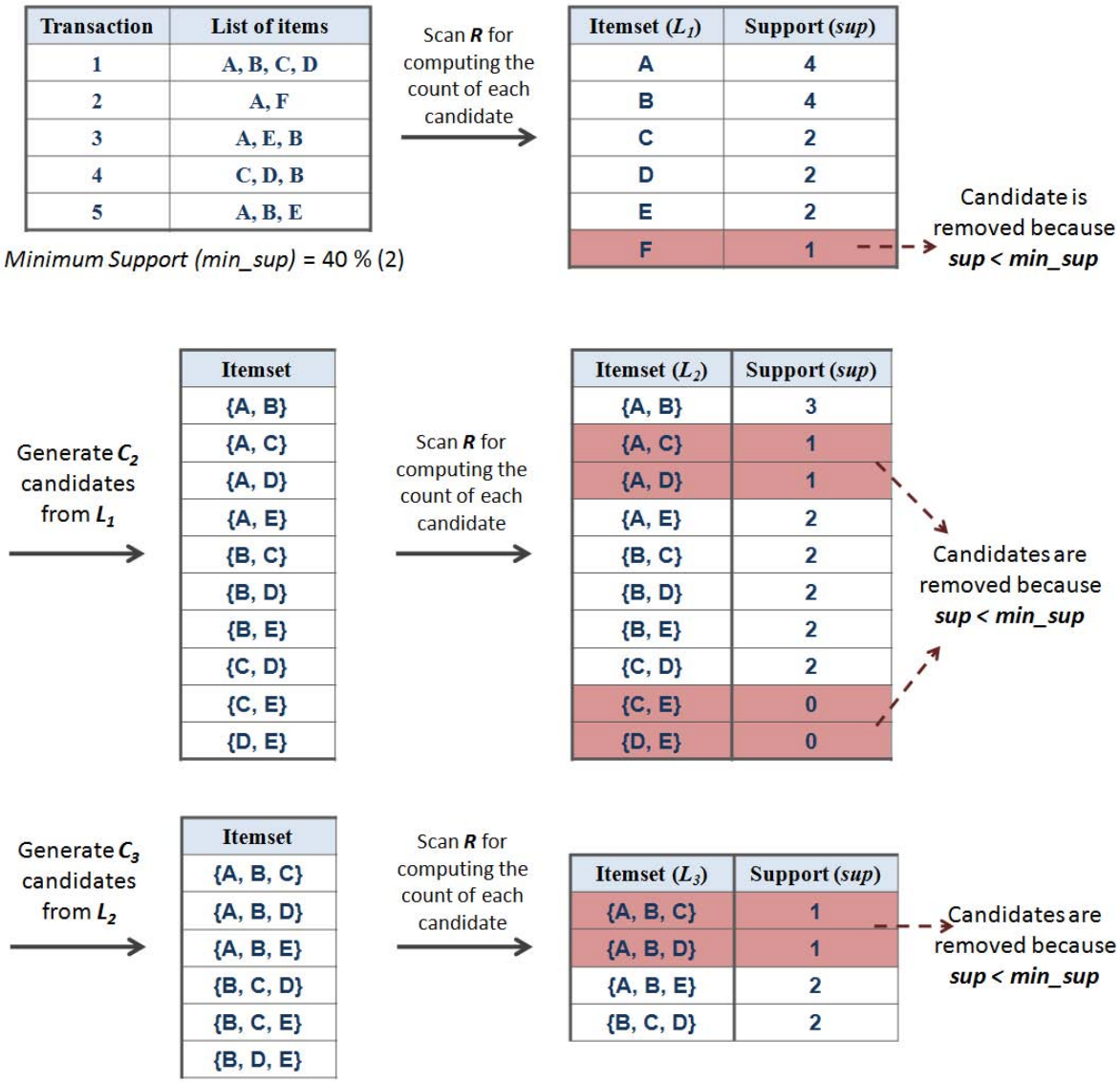



Figure 4.6: Generation of candidate itemsets by the Apriori algorithm, where the minimum support count is 40%.

(Omiecinski, 2003), conviction (Brin et al., 1997), and lift (McNicholas et al., 2008) are presented in Table 4.2.

### 4.5.2 Mining sequential patterns

A new category of data mining techniques called *sequence mining* or *sequential patterns mining* was created to deal with the sequential nature of the data considering the time or the relation of time of the events occurrence. The importance of this research topic is justified by the number of potential applications areas where sequential patterns can be mined, such as telecommunication, financial market, weather forecast, medicine, among others.

Table 4.2: Measures of interest used in the association rules mining.

| Measure | Calculus | Meaning |
|---|---|---|
| *all-confidence* | $allconf(X) = \frac{sup(X)}{max(sup(x \in X))}$ | All rules generated from $X$ have confidence at least $allconf(X)$. |
| *conviction* | $convi(X \rightarrow Y) = \frac{1-sup(Y)}{1-conf(X \rightarrow Y)}$ | Compares the probability of $X$ occurs without $Y$ with its frequency of occurrence. |
| *lift* | $lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{sup(X \rightarrow Y)}$ | *Lift* measures how much more frequent $X$ and $Y$ occur together than expected if they are statistically independents. |

In this context, *sequential patterns* can be described in the form: "*when A occurs, B also occurs within a certain time*". Basically, the difference to traditional association rules is that in the sequential patterns the time information is included in the rule itself and also in the mining process as time constraints (Ahola, 2001).

In general, the sequence data is defined by three columns: *object*, *timestamp* and *events*. Events can be different types of alarm in telecommunications, low or high rainfall, etc. Thus, each transaction in a database of sequence data corresponds to occurrences of events on an object at a specific time (Ahola, 2001). The main task associated with this type of data is to find sequential patterns in the data, which can be useful for forecasting future events, for example.

Over the years, many different algorithms have been proposed with the purpose of mining sequential patterns. However, these algorithms were designed to solve problems for specific applications, which implies in different forms to represent sequence, patterns, and rules discovered. According to Ahola (2001), *"the process of discovering sequential patterns involves two main issues: the structure of the patterns in terms of their representation and constraints and the method by which a pattern's strength is computed".*

The universal formulation of the sequential patterns defines the output of the mining process as a set of *frequent sequences* or *sequential patterns*. In addition, a sequence $s$ denoted by $< s_1, s_2, s_3, \ldots, s_n >$ consists of elements $s_i$, which are *events* or sets of events. The length of a sequence $s$ is given by $|s| = k$ where $k$ is the number of events in the sequence. For instance, considering the sequence $< (A), (C, B), (D), (G, F, E) >$, its length is seven and $s_1 = A$, $s_2 = (C, B)$, $s_3 = D$ and $s_4 = (G, F, E)$.

Algorithms for sequences mining allow users to define mechanisms that restrict the sequential patterns of interest, besides the restriction imposed for the support measure. Constraints are conditions imposed by the user. The sequential patterns to be mined must satisfy the constraints, which can be classified into two categories: restrictions of "generation" and restrictions of "validation". The first type of restriction is imposed on the

generation stage of mining algorithms to reduce the search space of patterns. The latter are constraints that can only be verified in the validation phase of mining algorithms.

The validation constraint Min-Max is a pair of integers $(m, M)$ where $m$ is the minimum threshold and $M$ is the maximum threshold (Srikant & Agrawal, 1996). Thus, a client $c$ supports the sequence pattern $c = s < s_1, \ldots, s_n >$ with the restriction $((m, M)$ if there are instants $t_1, t_2, \ldots, t_n$ such that $(c, s_1, t_1), \ldots, (c, s_n, t_n)$ constitutes the transactions database and for all $i = 1, \ldots, n-1$, $m \leq |t_{i+1} - t_i| \leq M$.

A restriction of Time-Window is a number $W \geq 0$ (Srikant & Agrawal, 1996). A customer supports the sequential pattern $c = s < s_1, \ldots, s_n >$ with the constraint Time-Window W if there are instants $t_1, \ldots, t_n$ such that for every item $i \in s_j$ exist $t_0^j \in [t_j - W; t_j + W]$ such that $(c, i, t_0^j)$ composes the database of transactions.

Constraints of sets are restrictions imposed to the patterns in the generation phase, such as: only generates patterns $s =< s_1, \ldots, s_n >$ where the itemsets $s_i$ satisfy a certain condition involving operations between sets. Restrictions of Regular Expression were introduced by Garofalakis et al. (1999, 2002) to mine sequential patterns $< s_1, \ldots, s_n >$ that satisfy a given regular expression.

Figure 4.7 illustrates some methods for sequential patterns mining developed along the last years, such as *WINEPI* and *MINEPI* (Mannila et al., 1997), *AprioriAll* (Agrawal & Srikant, 1995), GSP (Srikant & Agrawal, 1996), *SPADE* and *cSPADE* (Zaki, 2001), SPIRIT (Garofalakis et al., 2002), *MSDD* (Oates et al., 1997), *TAG* (Bettini et al., 1998), and *PrefixSpan* (Pei et al., 2001). Some of these algorithms are briefly presented in this section.

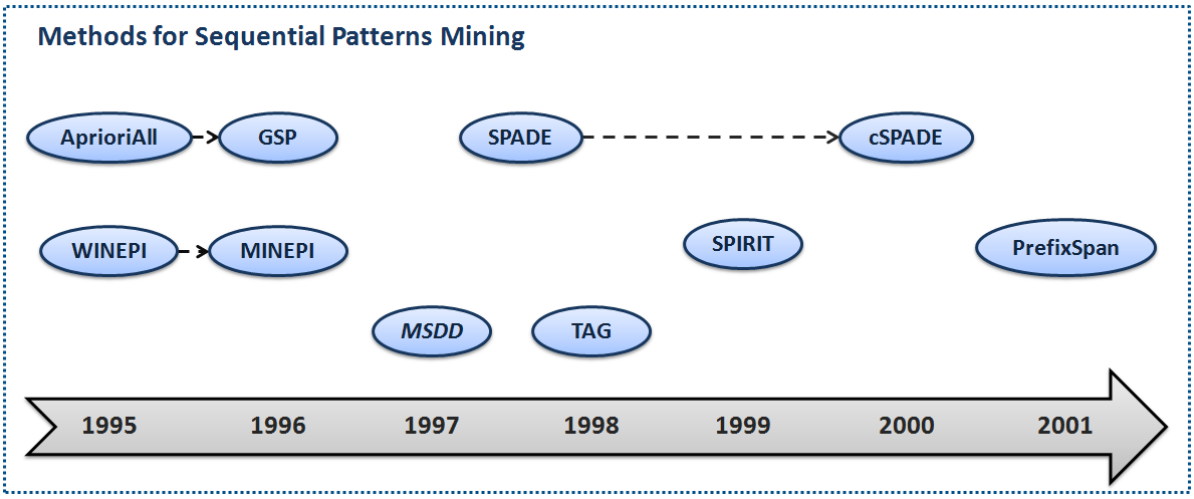



Figure 4.7: Timeline diagram highlighting the most important sequential patterns methods.

*WINEPI* is a set of algorithms proposed for discovering frequent sequences from alarm logs of telecommunication network (Mannila et al., 1997). This log consists of a single

and long sequence of events (alarms). In WINEPI, there is the definition of episodes that is a partially ordered collection of events occurring together. Although designed to telecommunication network alarms, this algorithm may be applied to any single sequence for different applications.

Sequences can be serial or parallel. The first one requires temporal order of events while the latter does not have requisition of relative order. WINEPI uses a time window to guarantee that events of sequences are close to each other. The time window slides over the data and only the occurrences within the window are considered. The support measure is calculated by counting the number of windows in which the sequence occurs. Thus, episode $\alpha$ is frequent if $fr(\alpha, s, win) \geq minfr$, i.e, "if the frequency of $\alpha$ is greater than the minimum frequency threshold in data sequence $s$ and with window of width $win$" (Mannila et al., 1997). The confidence measure is the conditional probability taken after the occurrence of a whole episode in a window. A given window can hold a number of episodes.

WINEPI rules are similar to association rules, but with a supplementary time aspect. In order to search for minimal occurrences, Mannila et al. (1997) proposed an extension of WINEPI, called MINEPI. It is similar to WINEPI giving the conditional probability that a certain combination of events happens within some time bound, given that another combination of events has occurred within a time bound.

The AprioriAll algorithm was proposed by Agrawal & Srikant (1995) to deal with the problem of mining sequential patterns over databases of customer transactions. The algorithm is executed in four phases. First, the database is sorted. Implicitly, the original transaction database is converted into a database of customer sequences.

In the second phase, the algorithm finds the set of all *litemsets*, which are itemsets with minimum support. Then, the *litemsets* is mapped to a set of contiguous integers in order to compare two *litemsets* for equality in constant time, reducing the time required to check if a sequence is contained in a customer sequence. In the next step, sequence phase, AprioriAll uses the set of *litemsets* to find the desired sequences. Finally, in the maximal phase, the algorithm finds the maximal sequences among the set of large sequences.

The *GSP (Generalized Sequential Patterns)* algorithm was proposed by Srikant & Agrawal (1996) for transaction data, where each sequence is a list of transactions ordered by time. Each transaction is defined as a set of items. GSP follows the same idea of Apriori to generate the frequent *k-sequences* (sequence with k items). Each iteration consists of the generation, pruning, and validation phases.

Initially, every item in the database is a candidate of length one. For each sequence of length $k$, GSP scans the database to collect support count for each candidate sequence and generates the candidate length of $(k+1)$ sequences from length $k$-frequent sequences

using Apriori. These steps are repeated until no frequent sequence or no candidate can be found.

GSP presents better performance than the AprioriAll algorithm. GSP prunes more candidates on the stage of pruning, and thus leads to the validation phase with much less elements to be tested. Especially on real data, typically the minimum level of support is very small, which leads to many candidates in later stages. Thus, eliminating the most possible candidates in the pruning phase that are not potentially frequent, the mining process is optimized. This is precisely the strategy used by the GSP algorithm.

*SPADE (Sequential Pattern Discovery using Equivalence classes)* was proposed by Zaki (2001) to discover all frequent sequences in large databases. SPADE is a vertical format sequential pattern mining method. Overall, the main features of the SPADE algorithm are:

1. it uses a vertical database format, that is, the database is re-organized where the rows of database are object-timestamp pairs associated with an event. Thus all frequent sequences can be enumerated via simple temporal joins on id-lists.

2. a lattice-theoretic approach is used for decomposing the original search space into smaller pieces that can be processed independently in main memory. Usually three database scans are required, or only a single scan when some preprocessed information is provided.

3. the problem decomposition is decoupled from the pattern search. Depth-first search is the strategy adopted for enumerating the frequent sequences within each sublattice.

An extension of the SPADE algorithm was proposed to mine constrained frequent sequences. This new algorithm is called cSPADE (*constrained Sequential Pattern Discovery using Equivalence classes*). The algorithm involves constraints, such as length, width, and time limitations on the sequences. Moreover, cSPADE considers minimum or maximum gap constraints on consecutive sequence elements, applying a time window on allowable sequences, incorporating item constraints, and finding sequences predictive of one or more classes (Zaki, 2000).

*SPIRIT (Sequential Pattern Mining Regular Expression Constraints)* is a family of algorithms for mining frequent sequential patterns that also satisfy user-specified regular expression constraints (Garofalakis et al., 1999, 2002). The algorithm uses a less restrictive version of the constraint to push the constraining inside the mining process. The constraints are imposed to prune the search space of patterns during computation. Basically, the algorithm is executed in several steps. First, it starts from the set of frequent

events and then next steps result in the discovery of longer patterns. In the $k^{th}$ step, a set of candidates sequences of length $k$ is generated from a set of frequent sequences of length $k-1$ and pruned. Thus, the data is scanned and the support of the candidates is counted, generating a set of frequent sequences of length $k$ (Ahola, 2001).

## 4.6 Distance Functions

When comparing data elements, distance function or dissimilarity function are important elements. A distance function must comply to the four following axioms to be considered a metric:

1. $d(s_1, s_2) >= 0$ (positiveness)
2. $d(s_1, s_2) = 0$ when $s_1 = s_2$, otherwise $d(s_1, s_2) > 0$ (reflexiveness)
3. $d(s_1, s_2) = d(s_2, s_1)$ (symmetric)
4. $d(s_1, s_3) = d(s_1, s_2) + d(s_2, s_3)$ (triangle inequality) for any $s_i$ pertaining to the data domain.

There are distance functions that do not satisfy the four axioms. These distance functions are called, in specific situations, pseudo-metrics or semi-metrics.

Figure 4.8 shows a classification for some distance functions commonly applied to time series. Given two time series, a similarity function calculates the distance between them. The diagram in Figure 4.8 refers to distance functions that compare the i-th point of one time series to the i-th point of another time series as lock-step measures, such as *$L_p$ norms* (Yi & Faloutsos, 2000) and *DISSIM* (Frentzos et al., 2007). Figure 4.8 shows Elastic Measures that allow comparison of one-to-many points, such as *Dynamic Time Warping (DTW)* (Berndt & Clifford, 1994) and *Derivative Dynamic Time Warping (DDTW)* (Keogh & Pazzani, 2001). Moreover, Figure 4.8 also presents Edit Distance Measures, such as *Longest Common Subsequence Model (LCSS)* (André-Jönsson & Badal, 1997), *Edit Distance on Real sequence (EDR)* (Chen et al., 2005), *Edit Distance with Real Penalty (ERP)* (Chen & Ng, 2004), and *Sequence Weighted Alignment model (Swale)* (Morse & Patel, 2007). Distance measures based on thresholds, for instance *Threshold Queries* (TQuEST) (Aßfalg et al., 2006), and patterns, such as *Spatial Assemble Distance* (SpADe) Chen et al. (2007b) are discussed as well.

The simplest similarity measure for time series is the *Euclidean Distance* (Faloutsos et al., 1994) and its variants, also known as the Minkowski family or $L_p$ norms (Yi & Faloutsos, 2000). The Euclidean distance corresponds to $L_2$, which is also commonly

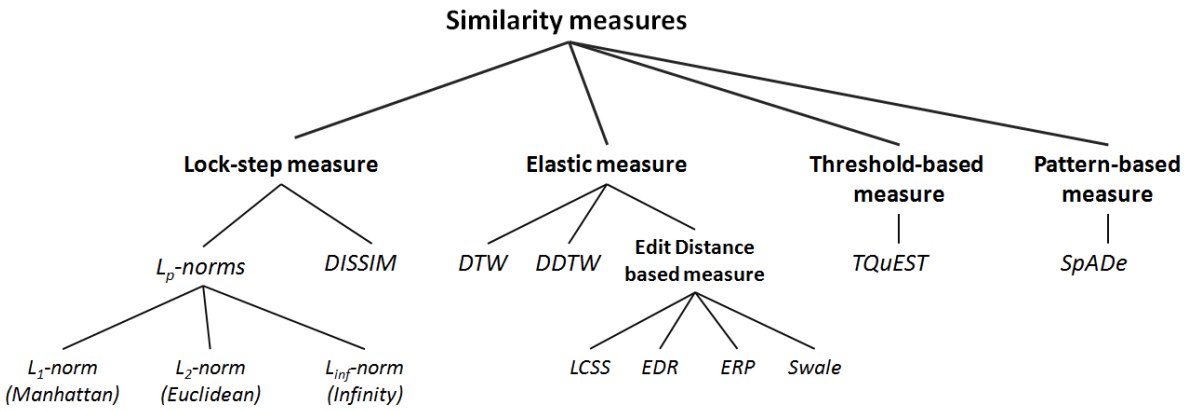



Figure 4.8: A hierarchy for distance functions (adapted from (Ding et al., 2008)).

used to calculate the distance between multi-dimensional arrays and vectors. Given two time series $Q$ and $C$, of length $n$, where:

$$Q = q_1, q_2, \ldots, q_n$$
$$C = c_1, c_2, \ldots, c_n$$

Equation 4.5 shows how to calculate the Euclidean Distance.

$$d(q_i, c_i) = \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2} \tag{4.5}$$

The Euclidean distance and its variants are intuitive, linear and easy to implement and to index with metric access methods. According to Ding et al. (2008), the Euclidean distance is competitive with other more complex approaches, particularly whether the database size is relatively large. However, these distance functions are very sensitive to noise and misalignment.

With the purpose of handling time warping in similarity computation, Berndt & Clifford (1994) proposed the *Dynamic Time Warping (DTW)*, which "stretches" or "compresses" one time series to provide a better match with another time series. Its main objective is to keep close time series that have similar behavior, but are delayed or distorted along the time axis. Thus, this technique has a good sensibility to warping because the comparisons between corresponding points become more flexible. In this case, points of a series can be compared to adjacent ones in other series, as illustrated in Figure 4.9.

To align two sequences using DTW, an $n, m$ matrix is built where the ($i^{th}$, $j^{th}$) element of the matrix contains the Euclidean distance $d(q_i, c_j)$ between the two points $q_i$ and $c_j$. Each element of the matrix corresponds to the distance between the points that it

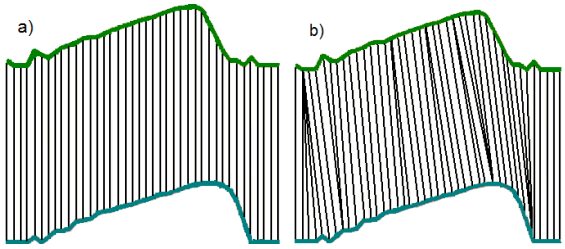



Figure 4.9: Comparisons between time series: a) conventional method; b) using DTW

represents. A warping path $W = (w_1, w_2, \ldots, w_k)$ is a contiguous set of matrix elements that defines a mapping between $Q$ and $C$. The adjustment route is defined by the following rules:

- it starts at $w_1 = (1, 1)$ and finishes at $w_k = (m, n)$
- the sequence of route must be to adjacent elements of the matrix (including diagonally adjacent cells)
- the points in $W$ must be monotonically spaced in time, that is, the sequence must not go back in the route.

There are many warping paths, but DTW is a sum of $w_k$ elements in the path that minimizes the warping cost. DTW is calculated by Equation 4.6.

$$DTW(Q, C) = min \left\{ \sqrt{\frac{\sum_{k=1}^{K} w_k}{K}} \right\} \tag{4.6}$$

where $w_k$ is the $k^{th}$ element of the adjustment route and $K$ is the number of elements of the adjustment route.

In Equation 4.6, $K$ in the denominator is used to compensate the size of the deviation between the two time series, because the warping paths may have different lengths. Dynamic programming is an efficient way to find the path, which is employed in Equation 4.7.

$$\gamma(i, j) = d(q_i, c_j) + min \{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \tag{4.7}$$

Improvements have been proposed to speed up similarity search using DTW introducing several lower bounding measures (Yi et al., 1998; Kim et al., 2001; Keogh & Ratanamahatana, 2005). As a result, the cost for computing DTW on large datasets was ameliorated becoming linear (Keogh & Ratanamahatana, 2005).

In this thesis, we have performed experiments with DTW to assess whether DTW is a suitable similarity measure to be applied in analyses of time series extracted from

satellite images. In this example, NDVI time series from ten sugar cane producing regions (Araraquara, Araras, Jaboticabal, Jardinópolis, Jaú, Luis Antônio, Pitangueiras, Pontal, Ribeirão Preto and Sertãozinho) of São Paulo state were used.

Similarity searches were computed for time series of the same region as well as distinct ones. Queries using a specific harvest as center were employed to identify similar crop seasons in the dataset. Figure 4.10 shows the graphics of the most similar time series (2004/2005) that indicates the crop season 2004/2005 had a similar behavior to crop season 2005/2006, considering Jaboticabal. The crop season 2003/2004 also had a similar trend to 2005/2006, which indicates a pattern of spectral response by NDVI from 2003 to 2006, which coincides with the sugar cane production values obtained from IBGE.

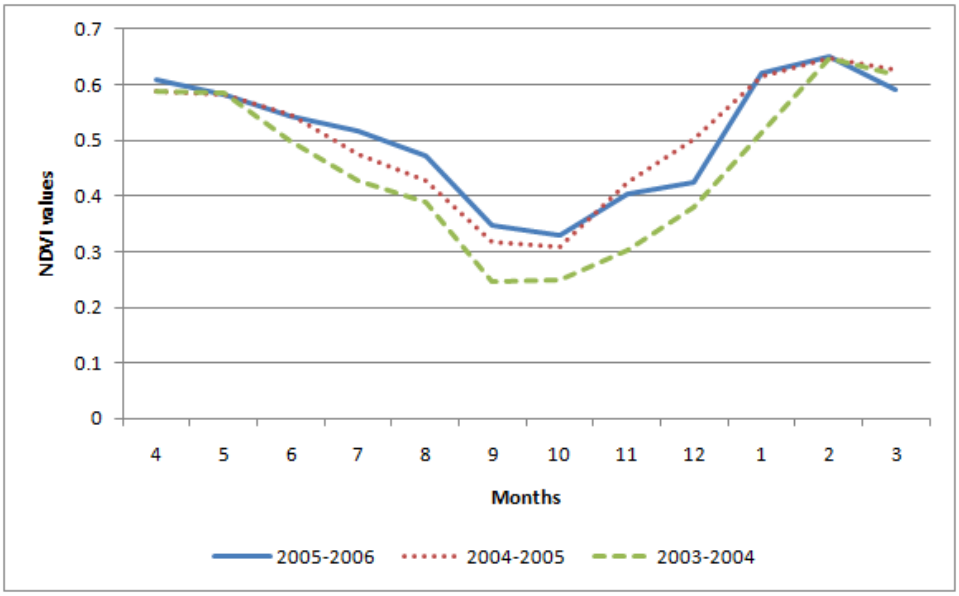



Figure 4.10: Graphs showing the result of similarity search for NDVI time series of Jaboticabal

Another example shows similarity search using time series for seven-crop-season years of all regions and had the purpose of analyzing which cities had a similar pattern of NDVI along the series. Figure 4.11 graphically shows these results.

Accordingly, a DTW-based method is appropriate to perform similarity search in NDVI time series from NOAA-AVHRR imagery. This approach makes the analysis of time series easier, because it finds similar series to a specific pattern presented by an automatic system. This technique has been used in time series mining in many areas, and can also be successfully employed to multi-temporal satellite images.

The *Longest Common Subsequence Model (LCSS)* is a distance function for time series based on the concept of the edit distance for strings (André-Jönsson & Badal, 1997; Vlachos et al., 2002). A threshold parameter $\epsilon$ was introduced indicating that two points from two time series are equivalent if the distance between them is less than $\epsilon$. Vlachos

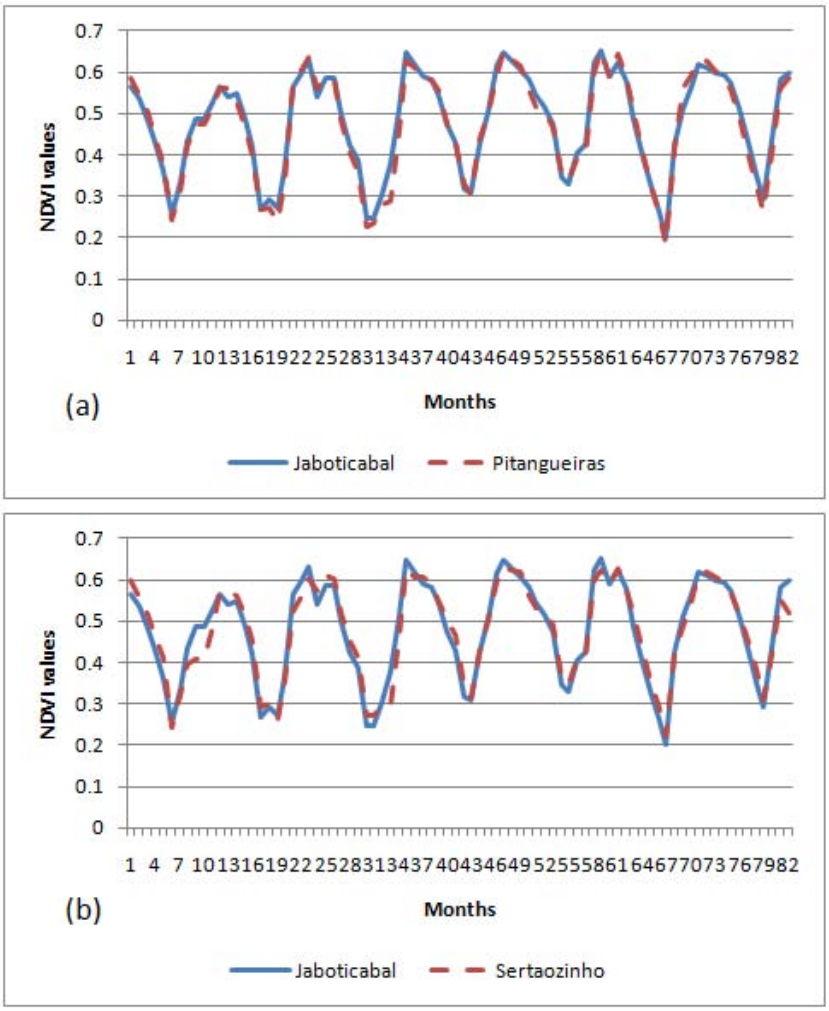



Figure 4.11: Graphs showing results of similarity search for complete NDVI time series using DTW: (a) Pintangueiras and (b) Sertãozinho.

et al. (2002) used a warping threshold to restrict the matching of points along the temporal dimension.

The *Edit Distance on Real sequence (EDR)* is a similarity measure of the group of edit distances (Chen et al., 2005). EDR also uses a threshold $\epsilon$ just as LCSS, though its function is to quantify the distance between a pair of points to 0 or 1. Contrarily to LCSS, EDR penalizes the gaps between two correspondent segments regarding to the lengths of the gaps.

The *Edit Distance with Real Penalty (ERP)* combines advantages of DTW and EDR through a constant reference point for calculating the distance between gaps of two time series (Chen & Ng, 2004). Basically, ERP uses the distance value between one of those points and the reference point if the distance between two points is too large.

A pattern-based distance function for time series, called *Spatial Assemble Distance (SpADe)* was proposed by Chen et al. (2007b). The SpADe algorithm discovers corre-

spondent segments (patterns) within the entire series, by allowing scaling and shifting in amplitude and temporal dimensions. The algorithm deals with the problem of finding the most similar set of matching patterns instead of computing the similarity value between time series. One difficulty related to the algorithm is the need to adjust various parameters, such as a temporal scale factor, amplitude scale factor, pattern length, etc.

## 4.7 Summary

In this chapter we presented the Knowledge Discovery on Databases (KDD) concept focusing on analysis of time series. Recently, given the growing importance of research involving climate change and agriculture data, where the amount of data increases every day, there is an opportunity for specialists in the data mining to develop new methods capable of handling large volumes of data in linear time.

We also described some preprocessing techniques since this task is very important to prepare time series to improve the performance of mining algorithms. Association rules is one of the most used tasks of data mining. We discussed the major algorithm for association rules (Apriori), which can be used to mine multiple time series that have been discretized. Due to its importance, we also detailed algorithms for sequence mining, that are more used to mine patterns in time series. The task of association is explored in this doctorate work as a way to detect association patterns in climate and remote sensing time series to contribute for the improvement of the monitoring process in agriculture crops, as we will present in Chapter 8.

Finally, we briefly described several distance functions used to similarity search in time series. DTW, also described in this chapter, is employed in this work to allow the similarity search of multidimensional objects that will be detailed in Chapter 6.

# Part II

# Contributions

# Chapter 5

# Employing Fractal Dimension in Time Series

## 5.1 Introduction

The proposition of new approaches and techniques to assist in the monitoring of agricultural crops is one of the goals of this thesis. Although there are traditional statistical methods that are widely used in the analysis of agrometeorological data, it is still necessary to develop new methods for specific problems, such as correlation detection involving several variables, joint analysis of multiple variables, and extreme events analysis. Furthermore, computational techniques can increase the data processing capacity, since the volume of data has increased in recent years due to improvements in sensor technology. Certainly, the analysis process and knowledge discovery in large amounts of data is a research challenge in different areas.

Consider, for example, datasets integrating climate data and remote sensing images from some sugar cane fields. A feature selection algorithm can identify the most relevant attributes of both datasets, which represent the majority of the information related to the agricultural yield and the correlations among attributes. Moreover, it is interesting to know which attributes can better approximate the values of others. In fact, detection of correlated attributes, their importance and precedence can improve the agrometeorological models for sugar cane monitoring and forecasting.

Additionally, in real climate applications, an impressive amount of time series is available, both generated by meteorological stations and interpolated over distributed grid points. As the data distribution in this application domain usually changes over time, climate time series can be seamless considered as evolving data streams. Therefore, tracking the behavior of evolving climate data can be very useful to agricultural monitoring, for

example, to monitor precipitation, air temperature and soil water content.

In this chapter, we present three different approaches to discover and to analyze patterns on time series of climatic data and remote sensing images. The described methods are based on fractal theory and data mining. The first technique in Section 5.2 is the FD-ASE algorithm, applied to identify sets of correlated attributes and to select relevant attributes to represent the meaningful features in the data.

The second one combines three algorithms in order to find association rules in datasets composed of climate and remote sensor data. We used the FD-ASE algorithm to select the meaningful attributes in datasets and extended its presentation format, before submit them to the Omega algorithm that transforms continuous data into discrete ones. The last step consists in applying the Apriori algorithm to extract association rules from the set of discrete data. This method called Apriori-FD is described in Section 5.3.

We also explore the fractal dimension as a tool to support a framework for data stream monitoring in agrometeorological applications (Section 5.4). The suitability of the fractal-based approach to monitor data streams is obtained by employing a statistical approach to compare the data in consecutive time periods, pointing out the attributes that are responsible for the trend changes and how they influence them.

## 5.2 Correlation Detection

In this section, we describe one case study that exemplifies the applicability of the approach we present in this chapter. We apply the FD-ASE algorithm (see Chapter 3 for details) to analyze the *SugarCaneRegion* dataset, which is composed of rain, maximum and minimum temperature, NDVI and WRSI values taken from the eight sugar cane productive areas from the São Paulo state from 04/01/2001 to 03/31/2008. Figure 5.1 presents the geographical distribution of these regions in the São Paulo state map.

We divided the *SugarCaneRegion* dataset into eight subsets, one for each evaluated region. The dataset attributes are presented in Table 5.1. Figure 5.2 shows a mapping in a 3-dimensional space of Araras and Pitangueiras dataset used in the experiments.

Table 5.1: Attributes description

| **Attribute** | **Meaning** |
|---|---|
| $a_1$ | Rainfall |
| $a_2$ | Maximum Temperature |
| $a_3$ | Minimum Temperature |
| $a_4$ | NDVI |
| $a_5$ | WRSI |

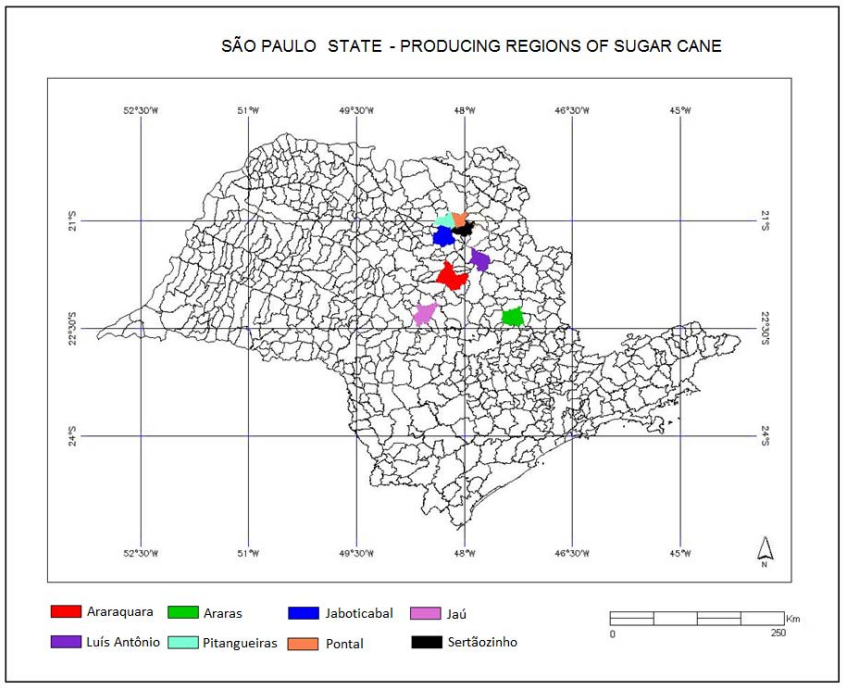



Figure 5.1: São Paulo state, located at southeastern Brazil (54° 00' to 43° 30' W and 25° 30' to 19° 30' S), where major sugar cane producers are found.

The intrinsic dimensions of the Araras and Pitangueiras datasets, measured through LiBOC (Traina Jr. et al., 2000, 2010), are illustrated in Figures 5.3(a) and 5.3(c). We applied the FD-ASE algorithm to the dataset and evaluated the threshold $\xi$ values above 0.7 for weak correlations. The intrinsic dimension calculation presented $\lceil D \rceil = 3$ or $\lceil D \rceil = 4$, depending on the dataset, as it can be seen in Table 5.2. The partial intrinsic dimension ($pD()$) of the reduced dataset (see Figures 5.3(b) and 5.3(d)) is very close to the intrinsic dimension of the full dataset. This value is lower than the embedded dimension $E = 5$. This means that there are 3 to 4 relevant attributes in the datasets that represent each of the eight cities studied, indicating that at least one of the attributes is correlated to the others.

By analyzing the correlations found, we can observe some interesting relationships between regions. It can be noted that the groups of correlated attributes (Correlation Group), the relevant attributes in each group (Correlation Base) and the set of relevant attributes considering the whole dataset (Attribute Set Core) are similar for different regions, for instance Araraquara and Luis Antonio.

Table 5.2 presents the Fractal Dimension, the Attribute Set Core ($\xi C$), Correlation Groups ($\xi G$) and their Correlation Bases ($\xi B$) generated for each region. For instance, FD-ASE found a Correlation Group $\xi G_1 = \{a_4, a_1\}$ and $\xi B_1 = \{a_4\}$ for Jaboticabal city. Thus, as $\xi B_1$ base contains NDVI ($a_4$), we can affirm that rainfall ($a_1$) is correlated to NDVI. On the other hand, the algorithm generated $\xi G_1 = \{a_3, a_2, a_1\}$ and Correlation

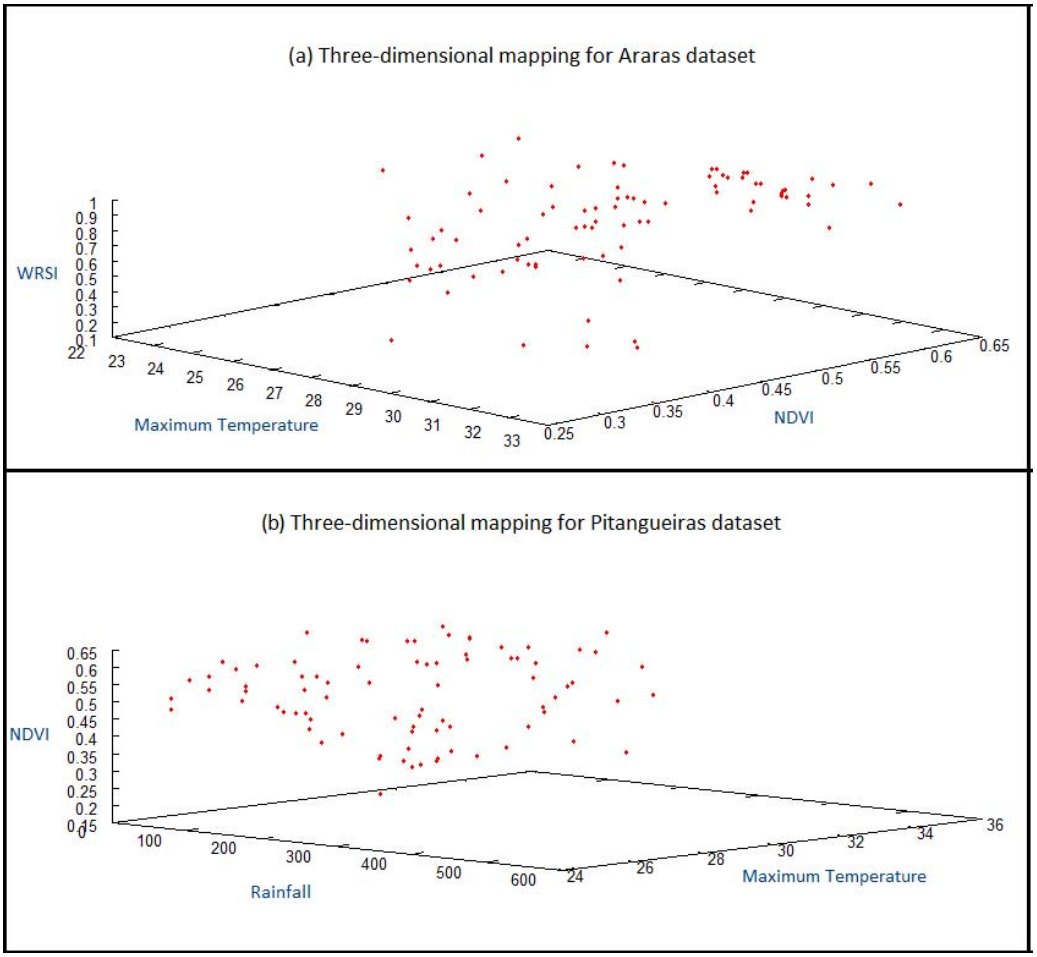



Figure 5.2: Representation of data in a 3-dimensional space. (a) Araras dataset; (b) Pitangueiras dataset

Base $\xi B1 = \{a_2, a_3\}$ for Araras city, i.e. as the base $\xi B_1$ contains maximum temperature ($a_2$) and minimum temperature ($a_3$), we can affirm that the rainfall ($a_1$) is correlated to the maximum and minimum temperatures for the $\xi$ threshold employed.

Table 5.2: Results of FD-ASE execution for a threshold $\xi$ indicating weak correlation

| **City** | **Fractal Dimension** | $\xi \geq 0.7$ | |
|---|---|---|---|
| Araraquara | 3.1705 | $\xi G_1 = \{a_3, a_5\}$ and $\xi B_1 = \{a_3\}$ | $\xi C = \{a_4, a_3, a_1, a_2\}$ |
| Araras | 3.6605 | $\xi G_1 = \{a_3, a_2, a_1\}$ and $\xi B_1 = \{a_2, a_3\}$ | $\xi C = \{a_2, a_3, a_4, a_5\}$ |
| Jaboticabal | 3.3286 | $\xi G_1 = \{a_4, a_1\}$ and $\xi B_1 = \{a_4\}$ | $\xi C = \{a_3, a_4, a_2, a_5\}$ |
| Jau | 3.8601 | $\xi G_1 = \{a_1, a_5\}$ and $\xi B_1 = \{a_1\}$ | $\xi C = \{a_3, a_2, a_1, a_4\}$ |
| Luis Antonio | 3.2222 | $\xi G_1 = \{a_3, a_5\}$ and $\xi B_1 = \{a_3\}$ | $\xi C = \{a_4, a_3, a_1, a_2\}$ |
| Pitangueiras | 3.6131 | $\xi G_1 = \{a_4, a_5\}$ and $\xi B_1 = \{a_4\}$ | $\xi C = \{a_2, a_3, a_1, a_4\}$ |
| Pontal | 3.1265 | $\xi G_1 = \{a_4, a_1, a_5\}$ and $\xi B_1 = \{a_4\}$ | $\xi C = \{a_3, a_4, a_2\}$ |
| Sertaozinho | 2.5071 | $\xi G_1 = \{a_2, a_5\}$ and $\xi B_1 = \{a_2\}$ | $\xi C = \{a_4, a_1, a_3, a_2\}$ |

All regions keep NDVI and the maximum temperature in the Attribute Set Core ($\xi C$), evidencing the importance of these variables in the datasets. Thereafter, by using the method of fractal correlation, we discovered the existence of correlations between

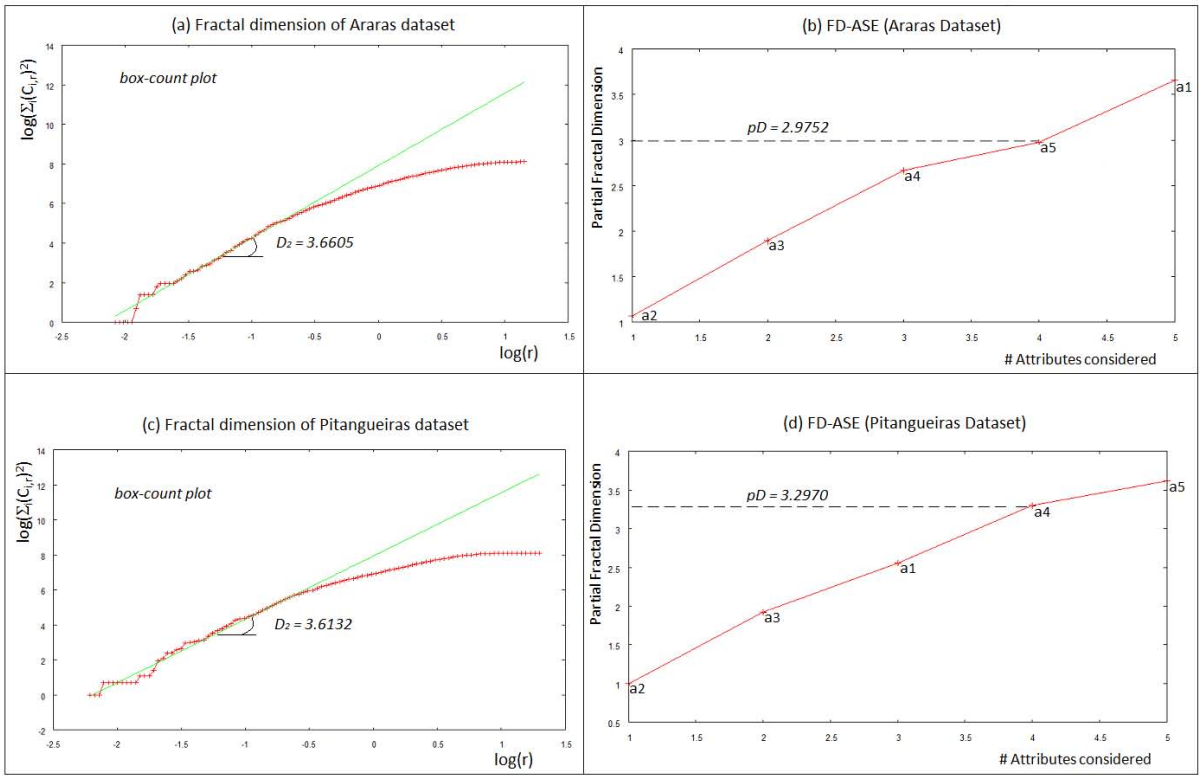



Figure 5.3: Araras and Pitangueiras dataset: (a and c) intrinsic dimension; (b and d) attributes for $\xi C \geq 0.7$

NDVI and precipitation, which is not identified when employing the Pearson's correlation technique (Pearson, 1896). The coefficient of the Pearson's correlation is calculated by Equation 5.1.

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.1}$$

where $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are the measured values for both variables, and $\bar{x}$ and $\bar{y}$ are the arithmetical average for both variables.

It is worth to mention that the Pearson correlation is the technique usually employed by agrometheorologists to find correlations among data. As the correlation found between NDVI and precipitation is not linear, it cannot be detected by Pearson. The FD-ASE method can also find correlation among more than two attributes, which is an advantage when compared to the aforementioned well-known Pearson's correlation.

Several meetings were conducted with agronomists (doctoral students) to assess the potential use of the FD-ASE algorithm in their research. During these sessions, three experts executed FD-ASE to find groups of correlated attributes in datasets involving data from orbital sensors, agrometeorological indexes, measurements taken in the field and meteorological data. Although we explained the method and followed the experi-

ments helping whenever asked, specialists showed difficulty to interpret outputs of FD-ASE. Thus, we proposed changes to FD-ASE in order to incorporate more semantics on the method. One of the improvements is to divide groups generated by the FD-ASE in subgroups. For example, if the result presented by FD-ASE is a Correlation Group $\xi G_1 = \{a_1, a_2, a_4, a_5\}$ with Correlation Base $\xi B_1 = \{a_1, a_2\}$, specialists have a tendency to interpret the Correlation Group $\xi G_1$ as the only information about the correlation between attributes, disregarding the Correlation Base $\xi B_1$. In this example, the Correlation Base $\xi B_1$ is composed of $a_1$ and $a_2$ what means that $a_1$ and $a_2$ are correlated with $a_4$ as well as $a_1$ and $a_2$ are correlated with $a_5$. However, we cannot affirm that $a_1$ and $a_2$ are correlated with $a_4$ and $a_5$ together. Thus, we proposed to present the subgroups of correlated attributes as subgroups $SubG_1 = \{a_1, a_2, a_4\}$ and $SubG_2 = \{a_1, a_2, a_5\}$.

Knowing how the attributes extracted from the raw data are correlated helps the specialists during the analysis of the data gathered. Furthermore, since the amount of data acquired and provided by satellites and meteorological stations is very large and grows in a very fast pace, a tool that highlights where the specialists should pay more attention is a valuable asset. According to these results, we proposed a method to mine association rules combining FD-ASE algorithm with association rule mining, which is detailed in the next section.

## 5.3 The Apriori-FD Method

According to results presented in the previous section, we showed that it is possible to take advantage of the fractal dimension to select the most relevant attributes from datasets composed of agroclimatic and remote sensing data. As the volume of this kind of data has increased fast and continuously in recent decades, specialists have to spend much more effort and resources to analyze them. This fact becomes a motivating opportunity to the development of new data mining algorithms and methods, as they provide important tools to identify relationships, patterns and correlations that are not previously known by the users.

In this context, we propose the *Apriori-FD* method to mine rules from heterogeneous time series, considering only the most relevant attributes, which are submitted to a pre-processing stage before the association rules mining process. As it can be seen in Figure 5.4, Apriori-FD method is applied in three steps.

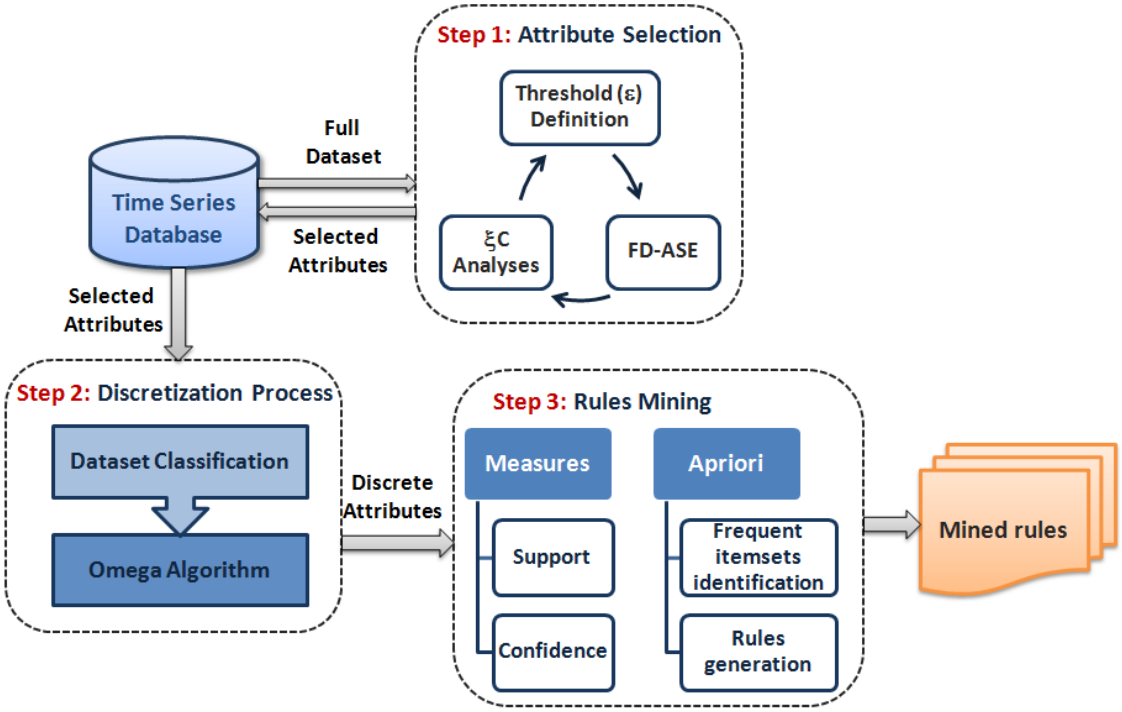



Figure 5.4: Steps to mine rules from relevant time series selected by FD-ASE through Apriori-FD.

### 5.3.1 Step 1: Attribute Selection

In this step, the most relevant attributes are selected by the FD-ASE algorithm. By applying FD-ASE in the beginning of the process, we reduce the number of potential rules to be generated by the association rules algorithm. Thus, this first stage of our method works as a filter that selects the main time series to be considered, making the analyses process of the generated rules easier. Additionally, since only relevant attributes are used as input to the next steps of the process, the processing time to mine rules diminishes considerably.

First of all, the user provides a threshold value ($\xi$) to FD-ASE that is executed generating an Attribute Set Core with the most meaningful attributes. This output can be assessed by the user before the second step starts.

### 5.3.2 Step 2: Discretization Process

The dataset of selected attributes is submitted to a supervised algorithm called Omega (described in Chapter 4) to accomplish the discretization process transforming all climatic, agrometeorological and remote sensing time series of continuous data into discrete ones. This transformation is mandatory since the association rule algorithm (Apriori) only accepts categorical or discrete data as input.

Omega first defines the initial cut points producing intervals that minimize the in-

consistencies gotten with the discretization process. After, the algorithm restricts the minimum frequency that an interval must present, avoiding a huge number of cut points. Thus, Omega joins consecutive intervals controlling inconsistency rate and removes the most inconsistent features. This last stage is not required since FD-ASE already selects the relevant attributes without needing classification. The discrete data generated by Omega are used as input to Apriori algorithm to mine association rules.

### 5.3.3 Step 3: Rules Mining

Apriori (Agrawal & Srikant, 1994) is a well-known algorithm for association rules mining that was chosen to mine rules from discrete datasets generated by Omega. Support and confidence measures must be defined to allow the rules generation in an appropriate way.

The algorithm (detailed in Chapter 4) first finds the frequent item-sets considering the property of frequent item-set to restrict the search space. As a result, a report file is generated with the rules that satisfy the minimum support and minimum confidence measures established.

### 5.3.4 Experimental Results

Experiments were conducted with the *SugarCaneRegion* dataset described in Section 5.2. It is composed of three time series of meteorological data (rain = $a_1$, maximum temperature = $a_2$ and minimum temperature = $a_3$), one time series of remote sensing data (NDVI = $a_4$) and one time series of agrometeorological data (WRSI = $a_5$).

According to the proposed method, eight datasets were submitted to FD-ASE algorithm that initially calculated their intrinsic dimension using the LiBOC() algorithm (details in Chapter 3). The value found for seven datasets is $\lceil D \rceil = 4$ and one of them is $\lceil D \rceil = 3$. It indicates that at least 3 or 4 attributes are required to maintain the intrinsic properties of the dataset. In the majority of the datasets, attribute $a_5$ was discarded.

The datasets Araras and Jaboticabal presented the same Attribute Set Core composed of attributes: $a_2$, $a_3$, $a_4$ and $a_5$. The others obtained attributes $a_1$, $a_2$, $a_3$ and $a_4$ in the $\xi C$. According to these results, datasets were submitted to Omega algorithm to the discretization process considering only attributes in $\xi C$.

The datasets with discrete data were used as input to Apriori algorithm using minimum support = 5% and minimum confidence = 100%. Apriori generated around 50 to 100 rules for each dataset. For convention, we defined three different ranges for NDVI values: minimum NDVI corresponds to (0.20 - 0.35), the range (0.36 - 0.56) is considered average NDVI and extreme values are in the range (0.56 - 0.66).

For Araras and Jaboticabal, rules have shown a direct relationship between the values of WRSI and NDVI as observed in the identification of attributes correlated.

Meaningful rules were found regarding to minimum or maximum temperature for all regions analyzed:

- minimum temperatures between 12°C and 13°C lead to average values of NDVI,
- higher minimum temperatures between 19°C and 20°C are associated with extreme values of NDVI.

Rules that associate the maximum temperature ($a_2$) with NDVI ($a_4$) show a variation in the range depending on the region. For example, for Jaboticabal, Jaú and Pitangueiras, Apriori found rules such as $a_2$[29-30] $\Rightarrow$ $a_4$[0.56-0.63], which means that when attribute $a_2$ is between [29-30] attribute $a_4$ is between [0.56-0.63]. For Araraquara and Luís Antônio, Apriori detected rules with a lower value for the maximum temperature, such as $a_2$[24-25] $\Rightarrow$ $a_4$[0.56-0.66]. Similar behavior was observed in rules involving minimum temperature ($a_3$) and NDVI ($a_4$). For instance, a rule such as $a_3$[20.09 - 20.98] $\Rightarrow$ $a_4$[0.56 - 0.63] was generated to Jaboticabal, Jaú and Pitangueiras, whereas $a_3$[19 - 19.34] $\Rightarrow$ $a_4$[0.56 - 0.66] was found for Araraquara and Luís Antônio.

According to these rules, maximum temperature between 20°C and 30°C are associated with maximum values of NDVI. Moreover, minimum temperature around 20°C is also linked with extreme values of NDVI. These patterns highlighted in the mined rules coincide with earlier studies carried out by agrometeorologists who confirm that temperatures between 22°C to 30°C are ideal for the optimum growth of sugar cane and minimum temperature for strong growth is approximately 20°C as aforementioned in Chapter 2 - Section 2.3. Therefore, this proposed method Apriori-FD allowed the mining of interesting patterns for agrometeorology, an indication of value ranges for temperatures, both maximum and minimum leading to a higher NDVI value, which is an indicator of productivity.

However, rules containing the rainfall attribute ($a_1$) showed lack of coherent association since different ranges of rainfall lead to the same values of NDVI ($a_4$). For example, all intervals composed of discrete values of rainfall (mm) were associated to average values of NDVI, such as $a_1$[0 - 1] $\Rightarrow$ $a_4$[0.36 - 0.56] and $a_1$[100 - 150] $\Rightarrow$ $a_4$[0.36 - 0.56]. That is, for both ranges of rain between 0 and 1, as well as from 100 to 150 the same NDVI is indicated. As the NDVI is an index that indicates the green biomass, the effect of rainfall on the plant growth could be captured by NDVI after a period of time. Thus, rules that consider time lag must be evaluated in order to discover more interesting patterns.

The experimental results of Apriori-FD method pointed to the necessity of new techniques to identify patterns that consider the time in order to predict more accurately the

phenomena. Furthermore, replacing Omega with another unsupervised algorithm could improve the proposed method avoiding the previous role of classification.

## 5.4 Data Stream Monitoring

Improvements in the data acquisition technology have decreased the time interval of data gathering, bursting the quantity of meteorological data. Moreover, since the behavior of this kind of data may change over time, monitoring activities have become more important. As applications in agrometeorology have generated continuous sequences of data over long periods of time, these data can be seamlessly considered *data streams*, as previously defined in Chapter 3.

In this section, we present a framework to monitor evolving climate data by employing a fast and low-cost process based on the fractal dimension extracted from the collected data. Significant changes in data trends are captured by the fractal-based monitoring process. The changes are evaluated by employing a statistical test to compare the data in consecutive time periods, revealing which data attributes are responsible for the trend changes and how they influence them.

The proposed method combines the SID-meter method (discussed in Chapter 3) with a *Data Analysis* module, as illustrated in Figure 5.5. When meaningful changes occur, the fractal-based process (SID-meter) triggers the Data Analysis module in order to analyze the data and to validate the variation of the fractal dimension, also identifying the changes that have occurred in the distribution of attributes. The fractal dimension variation is described in terms of mean and standard deviation of attribute values.

### 5.4.1 Step 1: SID-meter method

As new events (items) of the data stream are received, the SID-meter is executed to update the measures of the intrinsic dimension $D$ for recent events. For explanation purposes, $D_p$ represents the intrinsic dimension calculated over the current sliding window ($p$) and $D_{p-1}$ denotes the intrinsic dimension computed over the preceding window ($p-1$).

The values of $D_{p-1}$ and $D_p$ are continually monitored and compared until a significant difference between them is detected. We quantify the significance of the measurement variations based on a user-defined parameter $\epsilon$. Thus, $|\lceil D_{p-1} \rceil - \lceil D_p \rceil| > \epsilon$ is considered a meaningful difference. The smaller the value of $\epsilon$, the more sensitive the monitoring process. When $|\lceil D_{p-1} \rceil - \lceil D_p \rceil| > \epsilon$, the *Data Analysis* module is triggered, as it can be seen in Figure 5.6.

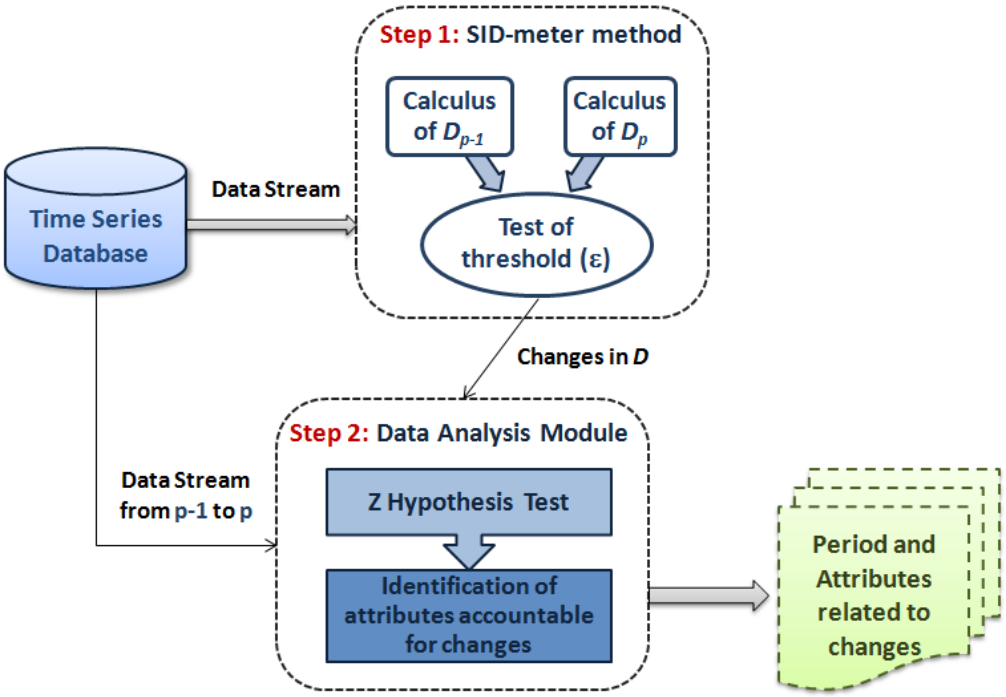



Figure 5.5: Integration of SID-meter and Data Analysis module.

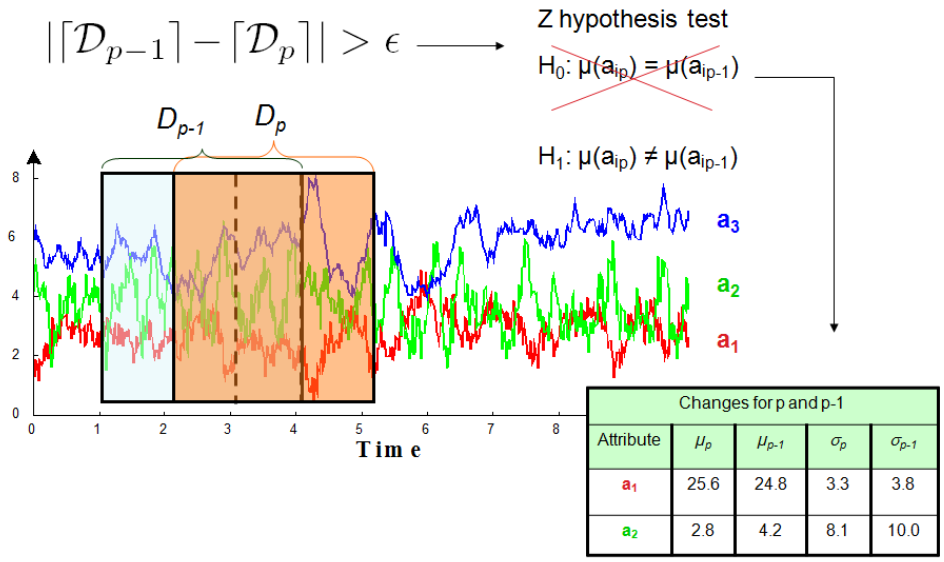



Figure 5.6: Example of execution: SID-meter and Data Analysis module (adapted from (Romani et al., 2009a))

### 5.4.2 Step 2: Data Analysis module

This module is employed to verify the attribute variations and to reveal which significant differences have occurred between the preceding $p-1$ and the current window $p$. Each event $ev_j$ of a data stream $\{ev_1, ev_2, ..., ev_n\}$ is defined by a set of $E$ measured attributes $a_i$, such that each $ev_j = (a_1, ..., a_E)$.

The set of $a_i$ attribute values in the current window $p$ of $n_i * n_c$ events is given by $a_{i_p} = \{a_{ip_1}, a_{i_{p_2}}, ..., a_{i_{p_n}}\}$. Similarly, the set of $a_i$ attribute values in the preceding window $p-1$ is given by $a_{i_{p-1}} = \{a_{i_{p-1_1}}, a_{i_{p-1_2}}, ..., a_{1_{p-1_n}}\}$.

The Data Analysis module runs Z hypothesis test that statistically analyzes data in the current window. For each attribute $a_i$, the Z hypothesis test is employed to compare $a_i$ mean between the preceding window $p-1$ and the current window $p$. Considers that the mean of $a_i$ values in the current and preceding windows are given, respectively, by $\mu(a_{i_p})$ and $\mu(a_{i_{p-1}})$.

The hypothesis $H_0$ should be rejected with a significance $\alpha_{min}$ (the test's probability of incorrectly rejecting the null hypothesis), in favor of the hypothesis that the means $\mu(a_{i_p})$ and $\mu(a_{i_{p-1}})$ are statistically different. Thus, $H_0 : \mu(a_{i_p}) = \mu(a_{i_{p-1}})$ and $H_1 : \mu(a_{i_p}) \neq \mu(a_{i_{p-1}})$. $\alpha_{min}$ is an input parameter that indicates the minimum confidence to reject the hypothesis $H_0$. Rejecting $H_0$ with the confidence $\alpha_{min}$ implies that the averages $\mu(a_{i_p})$ and $\mu(a_{i_{p-1}})$ are statistically different.

If the hypothesis $H_0$ is rejected, the mean ($\mu$) and standard deviations ($\sigma$) of the $a_i$ values are collected for the current($\mu(a_{i_p})$, $\sigma(a_{i_p})$) and preceding windows ($\mu(a_{i_{p-1}})$, $\sigma(a_{i_{p-1}})$) to describe the attribute changes in the data stream windows. Figure 5.6 presents a table with $\mu$ and $\delta$ examples for the attribute $a$ in windows $p$ and $p-1$.

The proposed method shows that it is possible to gather in a single graph the behavior of several variables and parameters, which otherwise would have to be analyzed separately, imposing to the specialist much more effort and time.

### 5.4.3 Experimental Results

Experiments on real and synthetic data streams were performed to evaluate and to validate the proposed method. Table 5.3 presents a summary of the datasets used, giving the number of attributes ($E$) and the number of events ($N$) in each one.

Table 5.3: Datasets definition

| Dataset | Description | Source | E | N |
|---|---|---|---|---|
| *Synt* | Attributes with similar distribution to climate data | | 5 | 18,000 |
| *ClimateCps* | Real data composed of measures of daily rain, maximum and minimum temperature from Campinas collected from 01/01/1890 to 01/19/2009 | IAC | 3 | 41,658 |
| *ClimatePira* | Real data composed by measures of daily rain, maximum and minimum temperature from Piracicaba collected from 01/01/1991 to 01/18/2009 | ESALQ/USP | 3 | 6,593 |

For all experiments, the significance $\alpha_{min}$ of the statistical test employed in the *Data Analysis* module was 0.01.

### Experiment 1: The *Synt* dataset

The synthetic data (*Synt* dataset) is employed to demonstrate the results of the proposed method. The *Synt* dataset simulates the trend of real agroclimatic datasets. It is composed of five float attributes (a, b, c, d, e) and 18,000 events generated according to Table 5.4.

Table 5.4: Definition of synthetic dataset (*Synt*)

| Attribute | Meaning | Range of Values |
|---|---|---|
| a | maximum temperature | 10..45 |
| b | minimum temperature | -5..31 |
| c | daily rainfall | 0..130 |
| d | vegetation index | 0..1 |
| e | agrometeorological index | 0..1 |

In this experiment we set the parameter of the monitoring process as $\epsilon = 0.5$ and defined a window with $n_c = 2$ counting periods of $n_i = 365$ events each, i.e., we created a two-year sliding window with a movement step of one year, considering daily events. The graph in Figure 5.7 shows the values of the intrinsic dimension measured over time for the *Synt* dataset.

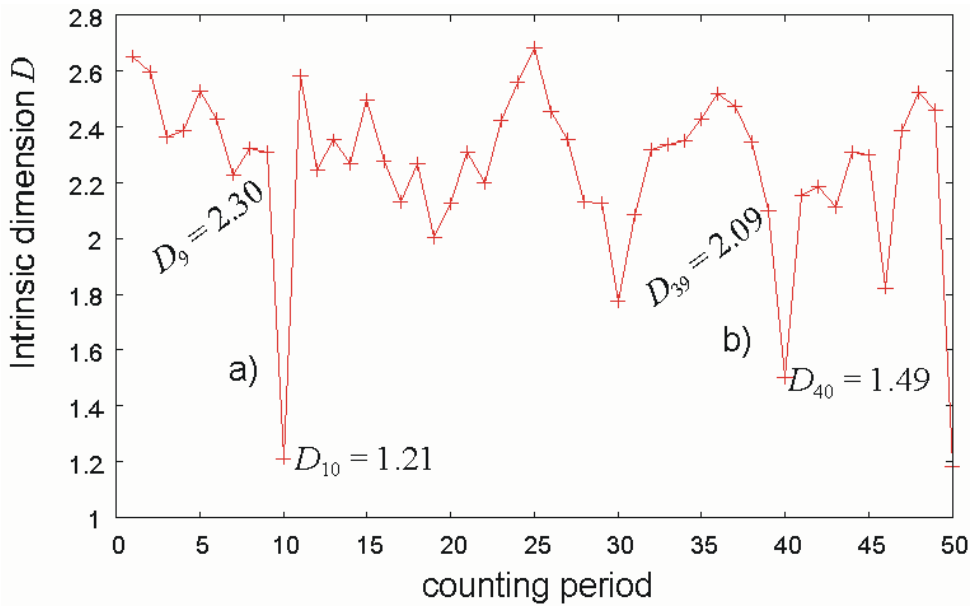



Figure 5.7: Monitoring process - *Synt* dataset, with dimension $D$ highlighted for meaningful periods.

As shown in Figure 5.7, the highest variations in the intrinsic dimension are in the counting periods $p = 10$ and $p = 40$. The first period of trend change is identified when the significant difference between $D_9$ and $D_{10}$ is pointed out. Thus, at period $p = 10$, the monitoring process triggers the *Data Analysis* module to analyze the attributes

considering the events covered by two consecutive sliding windows, namely the current window in period $p = 10$ and the preceding window in period $p - 1 = 9$. The hypothesis test indicated that attributes $a$, $b$, $d$ and $e$ had significant changes in their values in these consecutive periods. These changes are described in terms of mean and standard deviation of the attributes values in the current ($p$) and in the precedent window ($p - 1$) presented in Table 5.5.

Table 5.5: Attributes of *Synt* dataset revealed by Data Analysis module as presenting significant changes in their values of mean and standard deviation for the windows $p = 10$ and $p - 1 = 9$.

| **Attribute** | $\mu_p$ | $\mu_{p-1}$ | $\sigma_p$ | $\sigma_{p-1}$ |
|---|---|---|---|---|
| a | 35.51 | 19.25 | 0.28 | 7.95 |
| b | 11.09 | 16.49 | 8.74 | 9.21 |
| d | 0.49 | 0.46 | 0.17 | 0.17 |
| e | 0.41 | 0.30 | 0.15 | 0.16 |

As shown in Table 5.5, attribute $a$ had the major significant variation in the windows $p = 10$ and $p - 1 = 9$. The mean value increased from 19.25 in window $p = 10$ with a standard deviation of 0.28 to a mean of 35.51 with a standard deviation of 7.95. In climate data, this alteration may correspond to a variation between a very dry season and a wet season with days without raining. As also shown in Table 5.5, despite being significant, the $d$ attribute had the least significant variation in the windows $p = 10$ and $p - 1 = 9$.

The monitoring process goes on until period 40, when a new significant change is detected and the Data Analysis module is triggered to analyze the attributes considering the events covered by two consecutive sliding windows $p = 40$ and $p - 1 = 39$. The hypothesis test indicated that attributes $a$, $b$, and $e$ had significant changes in their values in these consecutive periods. These changes are described in terms of mean and standard deviation of the attribute values in the current ($p$) and in the precedent windows ($p - 1$) presented in Table 5.6.

Table 5.6: Attributes of *Synt* dataset revealed by Data Analysis module as presenting significant changes in their values of mean and standard deviation for the windows $p = 40$ and $p - 1 = 39$

| **Attribute** | $\mu_p$ | $\mu_{p-1}$ | $\sigma_p$ | $\sigma_{p-1}$ |
|---|---|---|---|---|
| a | 35.5 | 18.13 | 0.3 | 7.3 |
| b | 10.8 | 15.0 | 8.5 | 8.3 |
| e | 0.42 | 0.3 | 0.18 | 0.17 |

As in real climate data, after some seasons the climate tendency tends to repeat. Thus the behavior change triggered at window $p = 40$ is similar to the behavior triggered at $p = 10$. For instance, attribute $a$ had the major significant variation in both cases. However, attribute $d$, that had the least significant variation at window $p = 10$, now, at

window $p = 40$, was not returned by *Data Analysis* module as an attribute that had a significant trend variation.

Just for test purposes, we also executed the *Data Analysis* module for the windows with the smallest variation in the intrinsic dimension, which are the windows starting at periods $p = 32$ and $p - 1 = 31$. In this case, as the theory points out, no attribute was revealed by the *Data Analysis* module as having a significant value change. This is what a direct analysis from the raw data also shows.

### Experiment 2: The *ClimateCps* dataset

The *ClimateCps* dataset has three attributes: the daily minimum ($t_{min}$) and maximum ($t_{max}$) temperatures (° Celsius), and the amount of rainfall ($rain$) (mm) measured for a period of 114 years in Campinas, at São Paulo state. To calculate the intrinsic dimension of *ClimateCps* over time we used three counting periods ($n_c = 3$) and 365 events per period ($n_i = 365$), that is, $D$ is updated every 12 months in a three-year sliding window. The parameter to trigger the Data Analysis module was empirically set as $\epsilon = 0.1$, considering that we were interested in tracking small variations.

The same process described for the *Synt* dataset was carried on. The graph of Figure 5.8 shows the values of the intrinsic dimension over time for the climate data from Campinas. First, the *Data Analysis* module is triggered when the current window is $p = 17$. The statistical tests employed in the *Data Analysis* module indicate that the attributes $t_{max}$ and $rain$ had significant changes in their values in these consecutive windows. These changes are described in terms of mean and standard deviation of the attributes values in the current ($p = 17$) and in the precedent window ($p - 1 = 16$) presented in Table 5.7.

Table 5.7: Attributes of *ClimateCps* dataset revealed by Data Analysis module as presenting significant changes in their values of mean and standard deviation for the windows $p = 17$ and $p - 1 = 16$

| **Attribute** | $\mu_p$ | $\mu_{p-1}$ | $\sigma_p$ | $\sigma_{p-1}$ |
|---|---|---|---|---|
| $t_{max}$ (° Celsius) | 25.6 | 24.8 | 3.3 | 3.8 |
| $rain$ (mm) | 2.8 | 4.2 | 8.1 | 10.0 |

As shown in Table 5.7 the attribute with major significant variation was $rain$. The mean value decreased 3.4 from period $p-1 = 16$ to period $p = 17$. This fact corroborates the domain specialist's knowledge that the highest climate variation in the region was caused by rain variations along the years.

As it can be seen in Figure 5.8, the window $p = 63$ is the period with the highest variation in the intrinsic dimension $D$. The Data Analysis module indicates that attributes $t_{min}$ and $t_{max}$ had significant changes in their values in the preceding window starting at

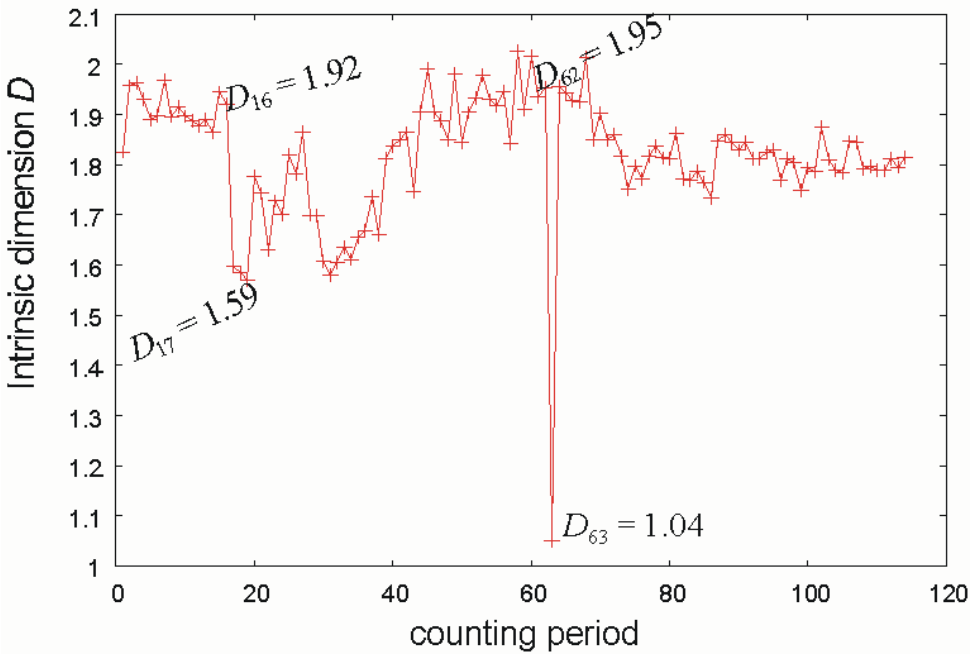



Figure 5.8: Monitoring process - *ClimateCps* dataset

period $p - 1 = 62$ and the current window starting at period $p = 63$. The *rain* attribute was kept without significant variation in these windows.

The large difference between $D_{63}$ and $D_{62}$ associated to the *rain* attribute stable behavior indicates a meaningful change in climate conditions. This fact can be confirmed by observing the graph in Figure 5.9 that illustrates the year 1951 with extreme rainfall (above 100mm) while 1952 presented rainfall bellow average and months without rain in the region of Campinas.

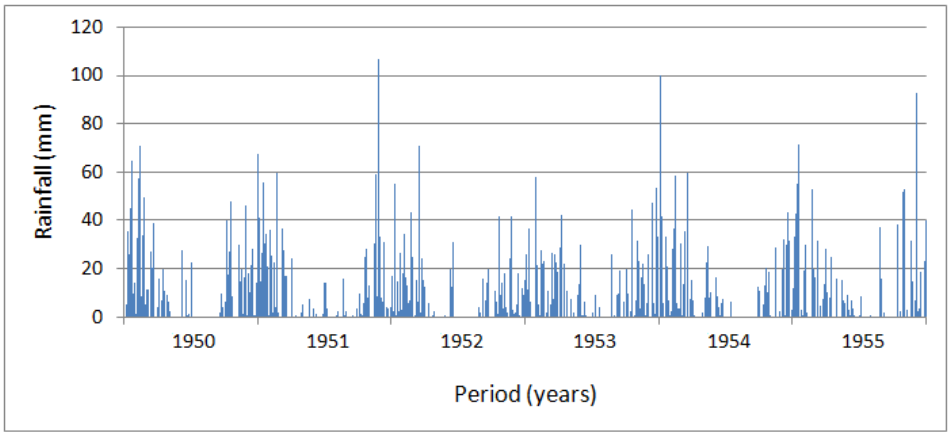



Figure 5.9: Daily Rain from 1950 to 1955

Moreover, Figure 5.10 shows the moving average of annual rain values where the lowest values coincide with the monitoring points indicated by SID-meter.

As it can be seen in Figure 5.8, the most invariant interval of intrinsic dimension

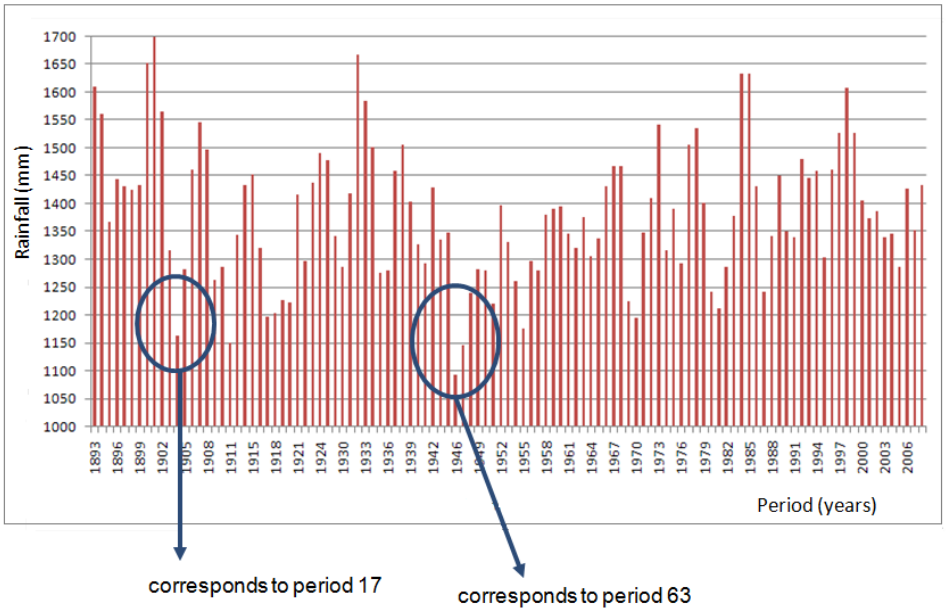



Figure 5.10: Moving Average for Rainfall (*rain*)

occurred between periods $p = 90$ and $p = 93$. Just for test purposes, we also executed the Data Analysis module for the window $p = 92$ and, as expected, no attribute was revealed by the Data Analysis module as having a significant value change.

### Experiment 3: The *ClimatePira* dataset

The *ClimatePira* dataset has three attributes, being each one the value of the daily rainfall (*rain*), maximum ($t_{max}$) and minimum ($t_{min}$) temperature measured for a period of 18 years at Piracicaba region, which is an important sugar cane producing region in Brazil. To calculate the intrinsic dimension of the *ClimatePira* dataset we used 2 counting periods ($n_c = 2$) and 182 events per period ($n_i = 182$), that is, $D$ is updated every 6 months for the climate measures in a one-year sliding window. The parameter to trigger the monitoring process was empirically set as $\epsilon = 0.15$, considering that we were interested in tracking small variations. The same process described for the *Synt* and *ClimateCps* datasets was employed. The graph of Figure 5.11 shows the intrinsic dimension along the time periods of the climate data from Piracicaba.

According to Figure 5.11, the highest variations in intrinsic dimension occurred in $p = 9$ and $p = 22$. The monitoring process activates the Data Analysis module for the counting period 9 (first semester of year 1995), and pointed out that the attributes *rain*, $t_{min}$ and $t_{max}$ had significant behavior changes. These changes are described in terms of mean and standard deviation of the attributes values in the current ($p = 9$) and in the precedent window ($p - 1 = 8$) presented in Table 5.8.

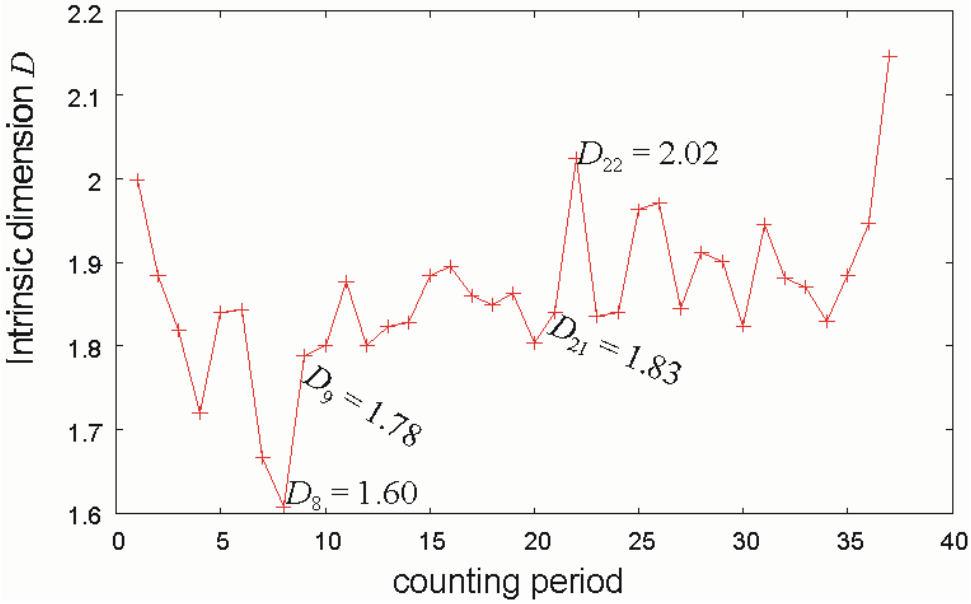



Figure 5.11: Monitoring process - *ClimatePira* dataset

Table 5.8: Attributes of *ClimatePira* dataset revealed by Data Analysis module as presenting significant changes in their values of mean and standard deviation for the windows $p = 9$ and $p - 1 = 8$

| **Attribute** | $\mu_p$ | $\mu_{p-1}$ | $\sigma_p$ | $\sigma_{p-1}$ |
|---|---|---|---|---|
| *rain* (mm) | 2.7 | 6.6 | 7.9 | 12.9 |
| $t_{min}$ (° Celsius) | 29.3 | 28.4 | 3.5 | 3.3 |
| $t_{max}$ (° Celsius) | 14.2 | 14.5 | 4.2 | 3.8 |

As shown in Table 5.8, the attribute rain had the major significant variation in the windows $p = 9$ and $p-1 = 8$. The mean *rain* has dropped 3.9 mm in the counting periods $p - 1 = 8$ and $p = 9$. The analysis of climate data in this period indicates an alteration in the rain distribution during 1994 and 1995, with a drought period (three months from June to August) more severe than normal conditions.

For the period 22 (second semester of year 2001), the Data Analysis module pointed that attributes *rain*, $t_{min}$ and $t_{max}$ had significant behavior change. This period indicates an alteration in the data behavior during 2001 and 2002, with fewer days of rain during the rainy season (three months from November to March).

When the intrinsic dimension does not significantly vary, the attributes also tend to keep their tendency, as the experiments showed. Important climate variations were discovered in the experiments, especially the ones involving large periods of abnormal rain. Therefore, the intrinsic dimension monitoring process proposed in this section is an important and well-suited tool to monitor streams of climate data. That is, instead of spending hours analyzing many different graphs and charts, the specialists now have a method that spots the regions of interest, where they should pay more attention during the

decision making process. Moreover, the proposed method allows to gather the information from all the attributes at once, instead of analyzing them separately.

## 5.5 Summary

In this chapter, we presented different methods that apply the fractal dimension to detect correlations among attributes in a dataset, to select relevant attributes to represent the meaningful features in the data and to support data stream monitoring. The experiments performed to validate the proposed methods involved climatic, agrometorological and remote sensing data.

The first approach applied to identify sets of correlated attributes indicated both linear and non-linear correlations on the contrary of Pearson's correlation method. In addition, the FD-ASE algorithm allowed detecting sets with more than two correlated attributes. Consequently, results showed that NDVI is correlated with rainfall and WRSI concomitantly, what it is not shown by other classic correlation methods, as Pearson's correlation for example. Moreover, the Attribute Set Core generated for the majority of the evaluated regions indicated that NDVI is a relevant attribute and must not be discarded.

Results of the Apriori-FD method presented relevant rules involving maximum and minimum temperatures. These association rules indicated ranges of temperatures that lead to maximum values of NDVI corresponding to agrometeorologists' expectations. The assessment of generated rules also allowed detecting similarity among datasets of different regions. However, outcomes relating rainfall pointed out to the need of new techniques that consider time lag in the generation of rules. In order to address these problems, we proposed two new algorithms that are detailed in Chapters 7 and 8.

The last method presented in this chapter allows monitoring data stream of climate and remote sensing data efficiently. In addition, the statistical test applied to consecutive windows revealed which attributes were responsible for tendency changes and how they impact these changes. During the experiments, we observed that defining the window size is not trivial for the agrometeorologists. There are several possible combinations for the $n_i$ and $n_c$ parameters, making the assessment accomplished by experts quite difficult. This method should be improved in order to become more independent from user definitions. One alternative could be the automatic identification of windows size grouping the generated graphs of fractal dimension according to similar patterns. Preliminary work to solve this problem is presented in Appendix B.

In the next chapter we present new contributions of this work to similarity search

involving climate and remote sensing time series.

# Chapter 6

# Distance Functions for Multiple Time Series

## 6.1 Introduction

As aforementioned, sugar cane has an important role in replacing fossil fuels, contributing to reduce production of greenhouse gases. Moreover, this agricultural commodity is important to the country's economy, becoming fundamental to improve models that assist the crops monitoring process. With the advent of remote sensing imagery, it becomes possible to monitor sugar cane fields on a regional scale by using vegetation indexes such as NDVI.

Consider, for example, multi-temporal NDVI images regarding several crop seasons. When we use time series of NDVI values to represent sugar cane fields, it is possible to find similar regions comparing the time series that symbolize them. However, time series are considered complex objects, which generally do not define a relation of total order ($=, \neq, >, \geq, <, \leq$). Thus, finding similar patterns to classify or to analyze different regions represented by time series is a non-trivial task.

As a solution, we can establish a relationship of similarity in complex objects, using distance functions to find the most similar objects. Therefore, specialists can use an automatic method to analyze a huge volume of time series finding similarities and clustering among them. Detection of similar regions aims at understanding the distribution of agricultural crops in a certain region. Moreover, this functionality can improve the monitoring process in a regional scale.

In this chapter, we present two new methods for finding similar regions represented by time series of climatological data and indexes obtained from satellite images. First of all, we propose a method called CV-DTW (*Correlation and Variance weighting Dynamic Time*

*Warping Distance*) to similarity search considering two-dimensional objects, i.e. objects represented by two different series, for instance NDVI and WRSI. This method takes advantage of the well-known DTW distance extending it when the values are weighted by the correlation between series and the variance of each one. This approach allows specialists to compare regions considering distinct series that represent them, as well as combining attributes of different types of sensors.

Finally, Section 6.3 details another proposed method, called FD-DTW (*Fractal Dimension weighting Dynamic Time Warping Distance*) that uses the fractal dimension to weight DTW. This method allows comparison between multidimensional objects i.e. objects represented by $n$ time series. The correlation fractal dimension measures the intrinsic dimension of the object independently of the space in which the object is embedded.

## 6.2 The CV-DTW method

We propose a method to measure the similarity between two-dimensional objects (datasets), i.e. objects represented by two different time series. This method weights the DTW distance function with the correlation between series and the variance of each one. The CV-DTW method is executed in six steps as presented by Algorithm 3.

**Algorithm 3** The steps of the CV-DTW Method

**Input:** Two objects (datasets) $A$ and $B$ composed of two time series 1 and 2 each one.
**Output:** The CV-DTW distance value.

1: Execute the DTW distance between correspondent series for datasets $A$ and $B$
2: Calculate the variance for time series 1 and 2 of datasets $A$ and $B$
3: Compute the variance factors $f_{V_1}(A, B)$ and $f_{V_2}(A, B)$
4: Execute the Pearson's correlation between two time series of datasets $A$ and $B$
5: Generate the correlation factor $f_C(A, B)$
6: Weight $DTW_1(A, B)$ and $DTW_2(A, B)$ by variance factor and correlation factor

Let $A$ be an object represented by two time series 1 and 2. Let $B$ be another object also represented by two time series 1 and 2. The CV-DTW method that calculates the similarity between objects $A$ and $B$ is given by Equation 6.1, which was empirically defined.

$$\begin{aligned} CV\_DTW(A,B) &= ((DTW_1(A,B) * f_{V_1}(A,B)) \\ &\quad +(DTW_2(A,B) * f_{V_2}(A,B))) \\ &\quad *f_C(A,B) \end{aligned} \tag{6.1}$$

where $DTW_1(A, B)$ is the distance between time series 1 of object $A$ and time series 1 of

object $B$, $f_{V_1}(A, B)$ is the variance factor for both time series 1 in the objects $A$ and $B$, $DTW_2(A, B)$ is the distance between time series 2 of object $A$ and time series 2 of object $B$, $f_{V_2}(A, B)$ is the variance factor for both time series 2 in the objects $A$ and $B$, and $f_C(A, B)$ is the correlation factor for objects $A$ and $B$.

The CV-DTW algorithm first calculates $DTW_1(A, B)$ and $DTW_2(A, B)$ applying the distance function DTW (line 1 of Alg. 3). Figure 6.1 shows a schematic representation for both of them. In real applications, objects $A$ and $B$ can represent two distinct regions of sugar cane crops, for example.

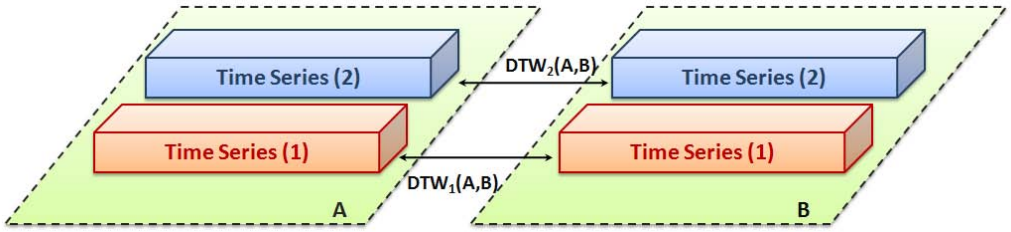



Figure 6.1: DTW calculus for two-dimensional objects $A$ and $B$.

In order to consider different distributions of each time series, CV-DTW calculates the variance for the time series 1 and 2 of objects $A$ and $B$ as $V_{A_1}$, $V_{B_1}$, $V_{A_2}$ and $V_{B_2}$ (line 2 of Alg. 3). Then, in the second step the algorithm uses these two variance values to calculate two variance factors $f_{V_1}(A, B)$ and $f_{V_2}(A, B)$ (line 3 of Alg. 3) that are given by Equations 6.2.

$$\begin{aligned} f_{V_1}(A, B) &= (1 - V_{A_1}) * (1 - V_{B_1}) \\ f_{V_2}(A, B) &= (1 - V_{A_2}) * (1 - V_{B_2}) \end{aligned} \tag{6.2}$$

where $V_{A_1}$ and $V_{A_2}$ are the variances for time series 1 and 2 of region $A$, and $V_{B_1}$ and $V_{B_2}$ are the variances for time series 1 and 2 of region $B$.

All variance values ($V_{A_1}$, $V_{B_1}$, $V_{A_2}$ and $V_{B_2}$) are subtracted from one to maintain the ascending order when CV-DTW is used to rank objects in the similarity search. In order to incorporate one measure that summarizes two time series 1 and 2 of each object, CV-DTW calculates the correlation between time series in both objects $A$ and $B$ as $C_A$ and $C_B$, respectively (line 4 of Alg. 3). Figure 6.2 illustrates a schematic representation for the correlation of an object. In the CV-DTW method, we used the Pearson's correlation, although depending on the time series used to represent each object, other measures of correlation might also be considered.

CV-DTW calculates the correlation factor, using these two correlation values $C_A$ and

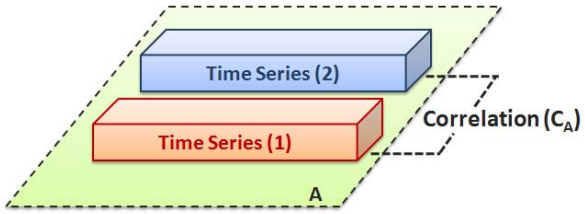



Figure 6.2: Example of correlation ($C_A$) between time series 1 and 2 of the object $A$.

$C_B$. The calculus of the correlation factor $f_C(A, B)$ (line 5 of Alg. 3) is given by Equation 6.3. Finally, the last step corresponds to the calculus of CV-DTW that indicates the similarity between objects A and B.

$$f_C(A, B) = C_A * C_B \tag{6.3}$$

where $C_A$ is the correlation value between two time series (1 and 2) for object $A$ and $C_B$ is the correlation value between two time series (1 and 2) for object $B$.

The computational cost of the CV-DTW method basically depends on the complexity of DTW. Although DTW has a time complexity of O($n^2$), there are codes that can effectively make DTW run in O($n$) (Keogh & Ratanamahatana, 2005). Thus, it is possible to calculate the CV-DTW method in linear time.

To evaluate the feasibility of the CV-DTW method, we have used this distance calculus with the kNN algorithm to measure the similarity of two objects. We consider the nearest-neighbor query to find the closest element to a query center, that is, given an element of interest (the center of the query), which are the elements of the dataset with smaller distances (higher similarities) to this element? Thus, given a query object $q_q$ and the set of data elements $T$, the nearest neighbor is the element of $T$ such that $NNQuery(q_q) = q_n \in T | \forall q_i \in T, d(q_q, q_n) \leq d(q_q, q_i)$. An example of a nearest neighbor query in a NDVI time series database is: "*find the time series in T that is the most similar to time series A*". The experimental results are presented in the next section.

### 6.2.1 Experimental Results

Experiments were performed with 10 datasets containing two time series (NDVI and WRSI) for regions of sugar cane fields in the São Paulo state (Araraquara, Araras, Jaboticabal, Jardinópolis, Jaú, Luis Antônio, Pitangueiras, Pontal, Ribeirão Preto and Sertãozinho). Each dataset corresponds to a two-dimensional object where NDVI and WRSI time series are related to a period from 2001 to 2008 with monthly measurements.

Three agrometeorologists individually classified the regions and ranked them considering one specific region as a query center. They generated a graph containing variables

NDVI and WRSI for each region they wanted to compare. Thus, they visually analyzed graphs to decide which regions were most similar to the query center. Finally, in a consensus way they prepared one ranking that was used in these experiments. The specialist's ranking is shown in Table 6.1. This ranking made by specialists was used as (ground truth) reference to access the fidelity provided by the automated result.

In order to validate the proposed method, we performed experiments employing two approaches that use:

1. *sumDTW*: sum of the DTW distances calculated for each series in different regions,

2. *CV-DTW*: weighting the DTW distance using correlation and variance factors.

### Jaboticabal as query center

The two approaches were used and generated a ranking with the most similar regions to the query center. Table 6.1 shows the results for the region of Jaboticabal as a query center. The methods sumDTW and CV-DTW presented different rankings for the same query, as is shown in Table 6.1. The rank proposed by the experts also appears in the same table.

Table 6.1: Comparative ranking using CV-DTW for similarity search in different regions (Jaboticabal as query center)

| Results for Jaboticabal as query center | | | | | |
|---|---|---|---|---|---|
| **Regions** | **Specialists** | **sumDTW** | | **CV-DTW** | |
| | ranking | ranking | values | ranking | values |
| Araraquara | 8 | 7 | 0.08680 | 8 | 0.02359 |
| Araras | 6 | 6 | 0.07875 | 6 | 0.02053 |
| Jardinopolis | 5 | 3 | 0.06788 | 4 | 0.01611 |
| Jau | 7 | 8 | 0.88589 | 7 | 0.02079 |
| Luis Antonio | 9 | 9 | 0.08899 | 9 | 0.02388 |
| Pitangueiras | 1 | 5 | 0.07071 | 5 | 0.01862 |
| Pontal | 2 | 1 | 0.01982 | 1 | 0.00609 |
| Rib. Preto | 4 | 2 | 0.06639 | 3 | 0.01484 |
| Sertaozinho | 3 | 4 | 0.06854 | 2 | 0.01370 |

In this experiments, CV-DTW presented more similar results to the ranking by the specialists than the other method. Dividing the list of regions in two groups, it can be seen that regions geographically closer to Jaboticabal appear in the top five ranking. The regions appearing in the latest ranking positions are geographically more distant and probably have small differences in climate that have been captured by the WRSI. The proposed method is closer to the results provided by specialists with one lag position in the ranking. When we just sum DTW values calculated for each series, the results do not

follow the specialists'. Figure 6.3 shows the top five positions in the ranking compiled by the specialist and by applying the CV-DTW distance function.

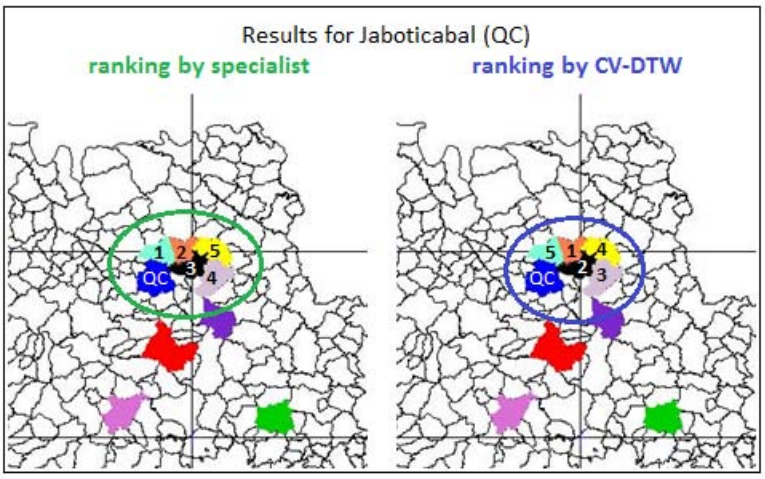



Figure 6.3: Visual presentation of the five top position of similarity search for Jaboticabal (query center) applying CV-DTW.

According to the experts, regions that appear in the top positions in the ranking have a climate more similar to Jaboticabal, the query center. Consequently, these regions have the same methods of planting and possibly the same cut-off date of sugar cane to the corresponding development stage of Jaboticabal. This explains the similarity among series. Thus, a method that approximates the ranking done by specialists can help identify similar regions in a given region with greater extentions, what can be difficult to do manually.

### Araraquara as query center

The two approaches were also used to generate a ranking to Araraquara as query center, which is presented in Table 6.2. The methods sumDTW and CV-DTW delivered different ranks for the same query, as is shown in Table 6.2. The ranking proposed by the experts also appears in the same table.

Although both methods have presented results very similar to the specialists' ranking, three positions exactly coincide and other two are inverted (2 and 4) in the method CV-DTW against only two coincident positions provided by the sumDTW method, considering the first five positions. Figure 6.4 shows the top five positions in the ranking compiled by the specialist and by applying the CV-DTW distance function.

According to specialists, Jaú, Araras and Luís Antônio are more similar to the Araraquara region due to climate conditions that define planting dates be more similar. Moreover, the values of sugar cane production released by IBGE for these regions are quite similar.

Table 6.2: Comparative ranking using CV-DTW for similarity search in different regions (Araraquara as query center)

| Results for Araraquara as query center | | | | | |
|---|---|---|---|---|---|
| **Regions** | **Specialists** | **sumDTW** | | **CV-DTW** | |
| | ranking | ranking | values | ranking | values |
| Araras | 2 | 3 | 0.09500 | 4 | 0.01658 |
| Jaboticabal | 6 | 7 | 0.08648 | 9 | 0.02424 |
| Jardinopolis | 7 | 8 | 0.09531 | 6 | 0.02046 |
| Jau | 3 | 2 | 0.08680 | 3 | 0.01481 |
| Luis Antonio | 1 | 1 | 0.06972 | 1 | 0.00552 |
| Pitangueiras | 9 | 6 | 0.08676 | 7 | 0.02098 |
| Pontal | 8 | 5 | 0.02197 | 8 | 0.02298 |
| Ribeirão Preto | 5 | 9 | 0.07365 | 5 | 0.01938 |
| Sertãozinho | 4 | 4 | 0.06921 | 2 | 0.01344 |

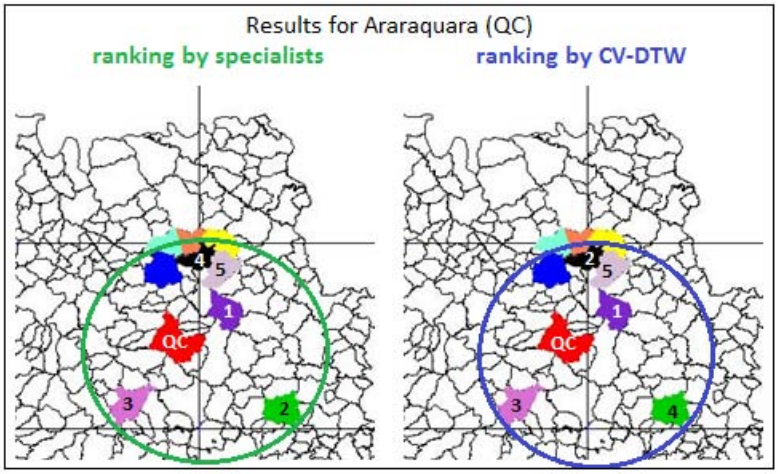



Figure 6.4: Visual presentation of the five top position of similarity search for Araraquara as query center with CV-DTW method.

Regions far away from Araraquara, such as Pitangueiras, Pontal, Jaboticabal and Jardinópolis have distinct climate characteristics with different amplitudes for temperature and dissimilar rainfall distribution. As a result, the first five positions indicated by specialists and detected by CV-DTW method have grouped regions with climate conditions more similar. As NDVI is directly influenced by climate, this variable did not change the result.

## 6.3 The FD-DTW method

When regions are represented as multidimensional objects (datasets) with more than two time series, the CV-DTW method becomes more complicated to be used due to the increasing of the possible combinations to calculate the correlation factor. The Pearson correlation method used in the previous proposal only calculates the correlation between

pairs of variables which makes the calculation of CV-DTW more complex.

As a solution, we propose a new method that takes advantage of the concept of intrinsic dimension. As explained in Chapter 3, the intrinsic dimension is related to the amount of information that the dataset represents, what can be estimated by the correlation fractal dimension from the dataset. In real applications, each sugar cane field can be modeled as a multidimensional object composed of heterogeneous time series from climate and remote sensing data. Thus, fractal dimension is an appropriate measurement to determine the correlation among all dimensions (climate and remote sensing time series) as demonstrated in Chapter 5 - Section 5.2.

The FD-DTW method combines DTW distance function with a factor based on fractal dimension in order to allow the similarity measure on two multidimensional objects (datasets) $A$ and $B$. FD-DTW weights the smaller DTW distance function between pairs of the same series in different objects by fractal dimension value calculated for each multidimensional object. The FD-DTW method is executed in five steps as presented by Algorithm 4.

**Algorithm 4** The steps of the FD-DTW Method

**Input:** Two objects (datasets) $A$ and $B$ composed of $n$ time series each one.
**Output:** The FD-DTW distance value.
**for** each $n$ pairs of time series for objects $A$ and $B$ **do**
2: Execute the DTW distance between correspondent series of datasets $A$ and $B$
**end for**
4: Calculate the fractal dimension of datasets $A$ and $B$ ($FD_A$ and $FD_B$)
Compute the fractal dimension factor $f_{FD}(A, B)$
6: Identify the minimum value from $DTW_1(A, B)$ to $DTW_n(A, B)$
Weight minimum DTW found by the fractal dimension factor

Let $A$ be a multidimensional object (dataset) represented by $n$ time series. Let $B$ be another multidimensional object also represented by $n$ time series. The FD-DTW method that calculates similarity between the objects $A$ and $B$ is given by Equation 6.4 that was empirically defined.

$$FD\_DTW(A, B) = min_{i=1}^{n}\{DTW_i(A, B)\} * f_{FD}(A, B) \tag{6.4}$$

where $DTW_i(A, B)$ is the distance between the same two time series $i$ of object $A$ and $B$, $f_{FD}(A, B)$ is the fractal dimension factor for objects $A$ and $B$.

First of all, the FD-DTW algorithm calculates $DTW_i(A, B)$ for all $i = \{1, \ldots, n\}$ time series, applying the distance function DTW (line 1-3 of Alg. 4). In this step, $n$ values of DTW are generated: $DTW_1(A, B), DTW_2(A, B), \ldots, DTW_n(A, B)$, where $A$ and $B$

refers to two multidimensional objects, as illustrated in Figure 6.5. For example, objects $A$ and $B$ can represent two distinct regions of sugar cane crops in real applications.

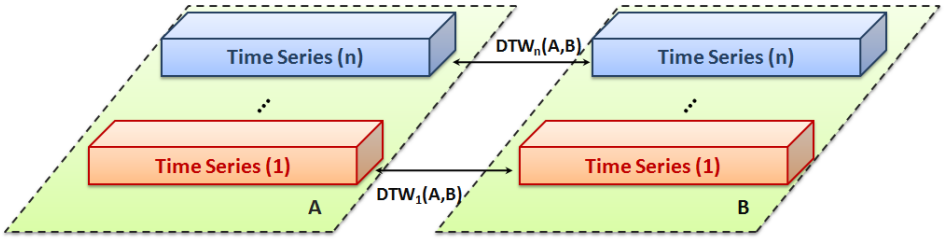



Figure 6.5: DTW calculus for multidimensional objects $A$ and $B$.

The second step corresponds to the computation of correlation fractal dimension between all series that represent a multidimensional object, as illustrated in Figure 6.6.

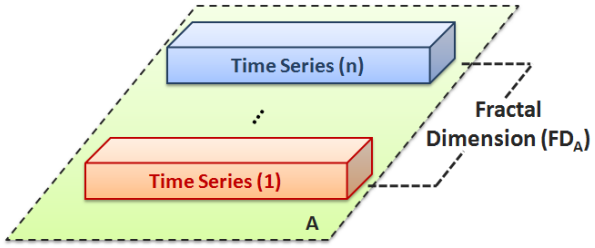



Figure 6.6: Example of fractal dimension ($FD_A$) involving $n$ time series of the object $A$.

In order to incorporate one measure that summarizes $n$ time series of each object, FD-DTW calculates the correlation fractal dimension among $n$ time series in both objects $A$ and $B$ as $FD_A$ and $FD_B$, respectively (line 4 of Alg. 4). The correlation fractal dimension indicates the real value that represents the object independent of the space in which the object is embedded (details in Chapter 3). FD-DTW calculates the correlation factor, using these two correlation fractal dimension values $FD_A$ and $FD_B$. The calculus of the correlation factor $f_{FD}(A, B)$ (line 5 of Alg. 4) is given by Equation 6.5.

$$f_{FD}(A, B) = FD_A + FD_B \tag{6.5}$$

where $f_{FD}(A, B)$ is the correlation factor, $FD_A$ is the correlation fractal dimension among $n$ time series of object $A$ and $FD_B$ is the correlation fractal dimension among $n$ time series of object $B$.

The last step of the method consists in the detection of the minimum $DTW_i(A, B)$ value to be multiplied by the correlation fractal dimension factor $f_{FD}(A, B)$ to generate the minimum distance between two multidimensional objects $A$ and $B$ (line 6 and 7 of Alg. 4). The computational cost of the FD-DTW method basically depends on the complexity of the fractal dimension estimation and the DTW calculation. Although DTW has a time

complexity of O($n^2$), there are improvements in the algorithm that can effectively make DTW run in O($n$) (Keogh & Ratanamahatana, 2005). Thus, as the fractal dimension was calculated by the Liboc() algorithm (linear cost on the number of elements in the dataset), it is possible to calculate the FD-DTW method in linear time.

To evaluate the feasibility of the FD-DTW method, we used this distance calculus with the kNN algorithm to measure the similarity of two objects, as aforementioned in Section 6.2. The experimental results are presented in the next Section.

### 6.3.1 Experimental Results

Experiments were performed with the same 10-producing regions of sugar cane in the São Paulo state in the same period, as previously described. However, in this experiment, the dataset for each region is composed of four variables: rainfall, maximum temperature, minimum temperature and NDVI. Experiments were accomplished considering two regions as query centers: Jaboticabal and Araraquara.

The assessment of the FD-DTW method was made ranking the most similar regions to the query center, which was compared to a ranking prepared by experts. The comparison of different regions represented by several variables is so difficult to accomplish without computational support. To classify the regions, experts have plotted all four series in a single graph for the 10 regions and visually analyzed the graphs in order to define regions with similar patterns. During this evaluation process, they consider different factors, such as climate conditions of the regions and production values for each region in the period of 2001-2008. This method used by experts is only feasible when the number of variables is small, but becomes impractical when the number of variables increases greatly. This ranking made by specialists was used as (ground truth) reference to access the accuracy provided by the automatic result.

#### Jaboticabal as query center

Considering Jaboticabal as the query center, we have executed the FD-DTW method and prepared a ranking that is presented in Table 6.3. We also show the ranking proposed by experts for comparison in Table 6.3.

FD-DTW presented results similar to the ranking given by the specialists in this experiment. The top five positions for both rankings is almost coincident, with a reversal in two positions: first and third. Figure 6.7 shows the top five positions in the ranking compiled by the specialist and by applying the FD-DTW distance function. These five regions are geographically closer to Jaboticabal, as it can be seen in Figure 6.7. The regions appearing in the latest ranking positions are more geographically distant and probably

Table 6.3: Comparative ranking using FD-DTW for similarity search in different regions (Jaboticabal as query center)

| Results for Jaboticabal as query center | | | |
|---|---|---|---|
| **Regions** | **Specialists** | **FD-DTW** | |
| | ranking | ranking | values |
| Araraquara | 9 | 9 | 0.11815 |
| Araras | 7 | 6 | 0.08497 |
| Jardinopolis | 5 | 5 | 0.07388 |
| Jau | 6 | 8 | 0.09877 |
| Luis Antonio | 8 | 7 | 0.09187 |
| Pitangueiras | 1 | 2 | 0.05667 |
| Pontal | 2 | 1 | 0.00001 |
| Ribeirão Preto | 3 | 4 | 0.06700 |
| Sertaozinho | 4 | 3 | 0.06140 |

there are differences in rainfall distribution and in the amplitude of temperature.

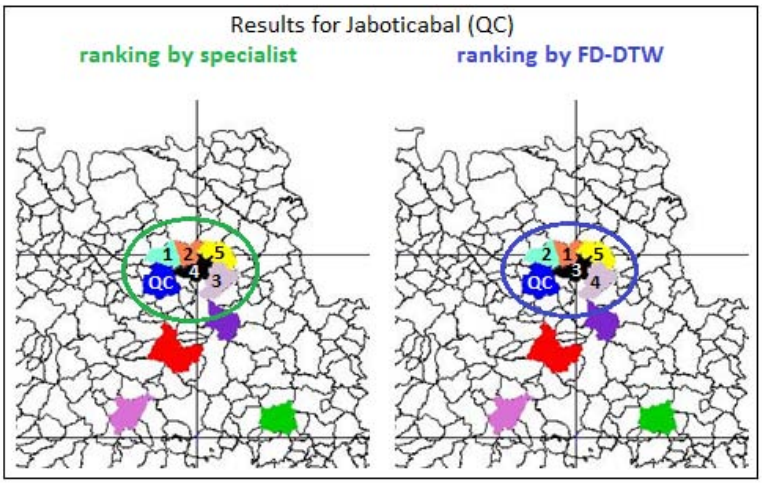



Figure 6.7: Visual presentation of the five top position of similarity query using Jaboticabal as the query center with FD-DTW.

According to specialists, the regions more similar to Jaboticabal have similar climatic conditions what make favorable the planting of the same variety of sugar cane during the same period of the year. Consequently, these regions have the same methods of planting and possibly the same cut-off date of sugar cane to the corresponding development stage of Jaboticabal (query center). This can explain the similarity among their series. Thus, a method that approximates the ranking done by specialists can aid identify similar regions in territories with great extensions what can be difficult to do manually.

### Araraquara as query center

In this experiment, we have used the Araraquara region as query center. This region is more geographically distant from Jaboticabal and is located in the center of the satellite

scene from where was extracted all time series. The FD-DTW method was executed to generate a ranking to be compared with the classification proposed by specialists. Both rankings are presented in Table 6.4.

Table 6.4: Comparative ranking using FD-DTW for similarity search in different regions (Araraquara as query center)

| Results for Araraquara as query center | | | |
|---|---|---|---|
| **Regions** | **Specialists** | **FD-DTW** | |
| | ranking | ranking | values |
| Araras | 2 | 2 | 0.09887 |
| Jaboticabal | 6 | 8 | 0.11815 |
| Jardinopolis | 7 | 6 | 0.11151 |
| Jau | 3 | 3 | 0.10524 |
| Luis Antonio | 1 | 1 | 0.00001 |
| Pitangueiras | 9 | 9 | 0.13683 |
| Pontal | 8 | 7 | 0.11622 |
| Ribeirão Preto | 5 | 5 | 0.11137 |
| Sertaozinho | 4 | 4 | 0.10721 |

FD-DTW presented results very similar to the ranking given by the experts in this experiment. The top five positions for both rankings is exactly the same. Figure 6.8 shows the top five positions in the ranking compiled by the specialist and by applying the FD-DTW distance function.

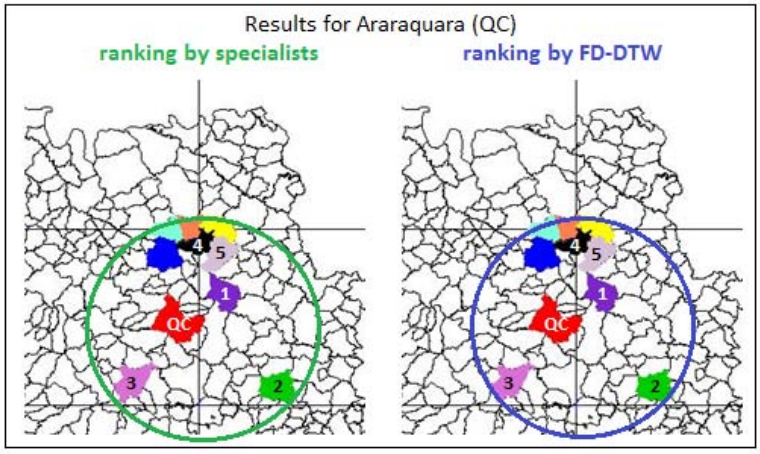



Figure 6.8: Visual presentation of the five top position of similarity query using Araraquara as the query center with FD-DTW.

The results indicated that Jaú, Araras and Luís Antônio are more similar to the Araraquara region, which was used as the query center in the experiment. Analyzing climate aspects of Araraquara and the other nine regions, the three regions that appear in the top position have a climate pattern very similar to Araraquara. The other regions geographically more distant present differences in characteristics of climate.

Focusing the analysis over NDVI we perceive that sugar cane is planted in equal dates using the same method of planting in all four regions (Jau, Araras, Luís Antônio and Araraquara). Moreover, the value in tons per year of sugar cane production is quite similar for these regions. Therefore, results of the proposed method are coherent and according to specialists' analysis.

Other regions are not similar to Araraquara since they are geographically more distant with different climate conditions. These regions (far from Araraquara) reach higher values of sugar cane production what corroborates that climate conditions should be carefully considered for sugar cane planting.

## 6.4 Summary

In this chapter, we presented different approaches based on a distance function and an algorithm to perform similarity search with the purpose of finding similar regions where sugar cane fields are cultivated. The feature of each region is represented by time series extracted from satellite images, generated by agrometeorological indexes and meteorological data.

The first method involves a weighting of DTW by correlation factor and variance since each region is now considered a bi-dimensional object. In the experiments, indexes that are used to show the spectral response of plants (NDVI) and the agrometeorological conditions to the proper growing of crops (WRSI) were used to represent each region. Applying the CV-DTW method, the results are similar to specialists' expectations. However, if the quantity of variables increases the calculus of the CV-DTW method becomes more complex due to the number of possible combinations to calculate the correlation factor.

The FD-DTW method uses the fractal dimension to weight DTW distance function in order to allow the similarity search of multidimensional objects. The fractal dimension measures the correlation in multidimensional datasets since it approximates the intrinsic dimension that represents the actual dimension of the object independent of its embedded dimension. Experiments were performed with four variables representing the plant and the climate of each region. Results are very similar (coinciding in many positions) to the ranking elaborated by experts.

The proposed methods allow specialists to assess wider regions in order to find areas of sugar cane with similar characteristics. Although it has achieved satisfactory results, tests with a larger number of regions represented by more variables must be considered.

# Chapter 7

# The CLIPSMiner Algorithm

## 7.1 Introduction

In the last decades, research in Climatology has indicated that climate is changing in the whole World. Meteorologists have analyzed large volumes of sensor data and outputs from global models in order to understand and to forecast extreme conditions and phenomena. Recent studies have indicated a disturbing situation regarding the temperature and precipitation in the Planet. Specifically, results from several analyses have showed that some extreme weather events have changed in frequency, duration and intensity over the last years (Meehl & Tebaldi, 2004; Vincent et al., 2005; Groisman et al., 2005; Goswami et al., 2006; Alexander et al., 2006; Ganguly & Steinhaeuser, 2008). Consequently, increased temperatures and regional changes in precipitation patterns can have adverse effects on natural and human systems.

Extreme precipitation events, such as heavy daily rainfall and many days with rainfall above the daily average can cause floods, which often result in devastating rural and metropolitan environments, as well as leading to loss of human lives. Thus, understanding trends of extreme events is so important to governments and communities to learn and to be prepared to mitigate the problem, and more importantly, to make decisions in a timely manner. Additionally, analyses of temperature time series indicate that it is crucial to define methods to reduce the emission of greenhouse gases and to adapt agricultural crops to the new conditions of increasing temperatures.

Climate data from ground-based stations, remote sensors, weather radars or sensor network have increased, yielding terabytes of data every week as mentioned earlier in Chapter 2. In addition, climate change models have been processed for different scenarios generating huge amounts of data. Consequently, experts need much more effort to analyze and to detect relevant patterns. Massive data volumes and processing complexity bring up

several problems and research challenges, such as forecasting of extreme events, correlation between climate and remote sensing time series, among others. Therefore, developing algorithms to retrieve relevant information for decision making and to extract interesting patterns is a needed endeavor.

Thus, to analyze large amounts of climate data from real measures or model outputs associated to remote sensing data and geographical information are important challenges to develop new data mining algorithms tailored to climate and agrometeorological data. In this context, we consider the problem of finding relevant patterns and extreme phenomena from climate time series. Generally, time series are converted into symbolic representation to simplify the analysis. However, we are interested in generating patterns considering continuous data, i.e., convert time series to a string sequence without losing information about data range.

As a solution, we present a new unsupervised algorithm, called CLIPSMiner (*CLImate PatternS Miner*) to discover relevant and extreme patterns in heterogeneous climate and remote sensing time series. This new algorithm works on multiple time series of continuous data, identifying all defined patterns or the relevant ones according to a relevance factor, which can be tuned by the user. To improve the analysis of long series, the CLIPSMiner algorithm allows the generation of patterns for given periods of time (for instance, years, months or days). Thus, meteorologists can examine and compare generated patterns in each period, considering its tendency, i.e., whether there was an increase/decrease on the number of patterns and/or if the maximum and minimum values in each pattern vary between periods.

This chapter presents in Section 7.2 a problem formalization of pattern mining. Section 7.3 details the CLIPSMiner algorithm. Section 7.4 presents and analyzes experimental results obtained by executing CLIPSMiner over synthetic data, as well as over real data. Finally, section 7.5 summarizes the Chapter.

## 7.2 Problem Formalization

Our focus is to generate patterns as discrete intervals that represent phenomena on climate time series. These patterns should be positive peaks, negative peaks or range of values with low variation. These patterns allow a quantization of time series, keeping the embedded semantic on data. Specifically, we address the following problems:

1. How to mine interesting climate patterns in time series of continuous data?

2. How to quantize time series retaining the temporal meaning of the patterns?

3. How to discover relevant patterns in datasets that combine heterogeneous time series?

4. How to mine different time series and detect time delay between them?

To better understand the proposed method, we define some important concepts as follows.

**Definition 7.1** ***Time series*** *$S$ is defined as a sequence of pairs $(v_i, t_i)$ with $i = 1, \ldots, n$, i.e. $S = [(v_1, t_1), \ldots, (v_i, t_i), \ldots, (v_n, t_n)]$, such that $(t_1 < \ldots < t_i < \ldots < t_n)$, where each $v_i$ is a data value and each $t_i$ is a time value in which $v_i$ occurs.*

Each pair $(v, t)$ is called an *event* $e$. A set of events contains $n$ events of type $(v_i, t_i)$ for $i = 1, \ldots, n$. Each $v_i$ is a continuous value. Each $t_i$ is a unit of time that can be given in days, months or years. Given two sequences $S_1$ and $S_2$, the values $t_i$ of both must be measured in the same time unit.

**Definition 7.2** ***The event sequence*** *$S_e$ is a set of consecutive events $e_i$, i.e. $S_e = (e_i, e_{i+1}, \ldots, e_k)$, where $e_i = (v_i, t_i)$ for $i \geq 1$ and $k \leq n$ and $k - i \geq q$, where $q$ is the minimum number of events in an event sequence.*

We are interested in extracting event sequences from a given sequence where the number of elements $e_i$ in the event sequence depends on the difference between events given by $d_i = (v_{i+1} - v_i)$ and a given $\delta$ parameter whose default value is set by the CLIPSMiner algorithm. The extracted event sequences comprise a period of events having the tendency to rise or fall, when plotted as a graph.

The value of $\delta$ is usually very small, tending to zero ($\delta \to 0$). The value of delta can also be defined by the user. Therefore, we define three exclusive types of event sequences.

**Definition 7.3** ***The ascending event sequence*** *$S_{ea}$ is a set of consecutive events $e_i$, such that $S_{ea} = (e_i, e_{i+1}, \ldots, e_k)$ where $\sum_i^k (d_i) > 0$, such that $\forall d_i$, $d_i > 0$ and $|d_{k-i}| < \delta$ to $(k - i) \leq$ parameter defined by the user.*

**Definition 7.4** ***The descending event sequence*** *$S_{ed}$ is a set of consecutive events $e_i$, such that $S_{ed} = (e_i, e_{i+1}, \ldots, e_k)$ where $\sum_i^k (d_i) < 0$, such that $\forall d_i$, $d_i < 0$ or $|d_{k-i}| < \delta$ to $(k - i) \leq$ parameter defined by the user.*

**Definition 7.5** ***The stable event sequence*** *$S_{es}$ is a set of consecutive events $e_i$, such that $S_{es} = (e_i, e_{i+1}, \ldots, e_k)$ where $\forall d_i$, $|d_i| < \delta$.*

The combination of different types of event sequences generates *patterns* that resemble peaks (negative and positive) and intervals with constant distribution.

A meaningful change or stability in the data distribution behavior should be monitored. For example, a variation from $0mm$ to $120mm$ in a short period of time in a rain series can mean an extreme phenomenon responsible for a flood at a given location. Thus, we define three types of patterns used to quantize a time series $S$.

**Definition 7.6** ***Valley patterns (V)*** *are defined as the concatenation of a descending event sequence and an ascending event sequence, i.e.* $V \Rightarrow S_{ed}S_{ea}$.

**Definition 7.7** ***Plateau patterns (P)*** *are defined as a stable event sequence, i.e.* $P \Rightarrow S_{es}$.

**Definition 7.8** ***Mountain patterns (M)*** *are defined as the concatenation of an ascending event sequence and a descending event sequence, i.e.* $M \Rightarrow S_{ea}S_{ed}$.

Figure 7.1(a) presents an example of a pattern $V$. In real data, a pattern $V$ can be observed when a sharp drop in the minimum temperature occurs, for example. For WRSI time series, a pattern $P$ can occur when $v_i$ has values closer to 1. This behavior in time series corresponds to the maximum soil water content, after a long period of rainfall, for example. In a real dataset, pattern $M$ occurs when there is a significant variation in the amplitude, such as a very heavy rain, for example. Figure 7.1(b) presents an interval in a time series and highlights a pattern $P$. Finally, Figure 7.1(c) presents an interval in a time series and highlights a pattern $M$.

Two thresholds ($\rho$ and $\lambda$) are defined to identify only the relevant patterns, acting as filters. The threshold $\rho$ is the relevance factor and it depends on the amplitude measure. The relevance factor is a measure for identifying whether a pattern $M$ or $V$ is relevant or not. The threshold $\lambda$ is the plateau length and is defined to identify relevant $P$ patterns. Both thresholds get a default value, that can be tuned by the user.

**Definition 7.9** ***Amplitude*** *($y$) is defined as the difference between the maximum and the minimum values of the time series,* $y = v_{max} - v_{min}$.

**Definition 7.10** ***Relevance factor*** *($\rho$) is a percentage of the amplitude value and is used to evaluate the height of an ascending ($S_{ea}$) and a descending ($S_{ed}$) event sequence.*

**Definition 7.11** ***Plateau length*** *($\lambda$) defines the length of an stable event sequence ($S_{es}$).*

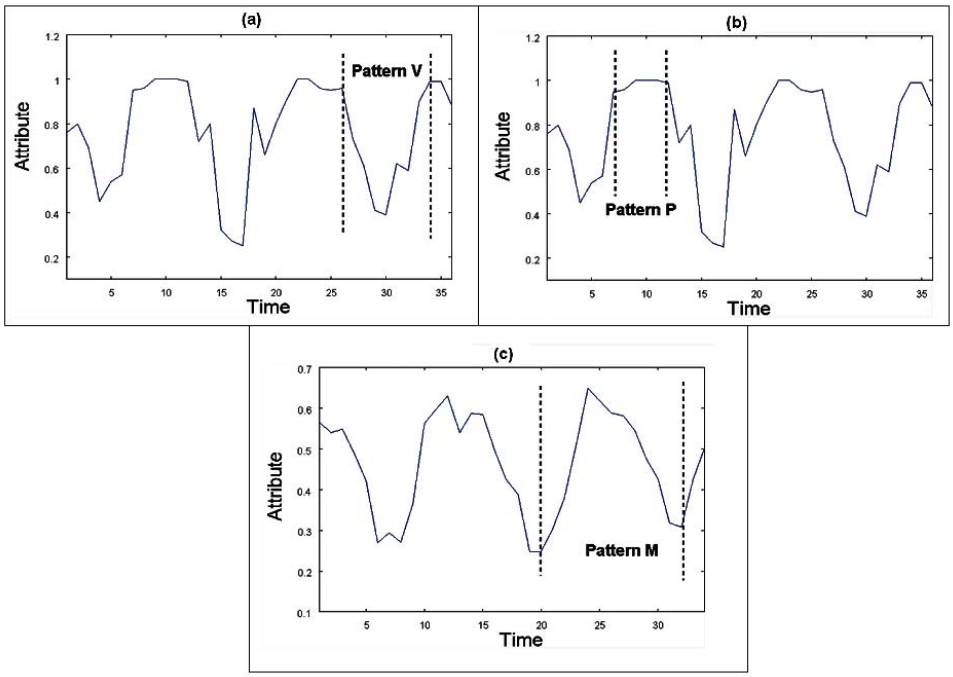



Figure 7.1: Examples of patterns detected by CLIPSMiner are presented in graphical format where the $y$ axis represents attribute value and time is given in $x$ axis. (a) Pattern of type $V$ is similar to negative peaks. (b) Pattern of type $P$ implies an interval in the time series with small variation. (c) Pattern of type M is equivalent to a positive peak.

For example, a pattern $M$ of daily rainfall ranging among (0, 5, 0) is not representative because it has a very small variation (only 5 mm), considering a range from 0 to around 150. However, an interval of daily rain that ranges from (0, 120, 0) is an extreme phenomenon that may cause disasters. In this case, the relevance factor indicates which patterns will be considered.

**Definition 7.12** ***Time delay*** $\tau$ *is the time interval between the beginning of the occurrence of a pattern in a time series and the beginning of the occurrence of a similar pattern in another time series. The time delay is measured in units of time.*

## 7.3 Description of the CLIPSminer algorithm

In this section, we present the CLIPSMiner algorithm that finds relevant and extreme patterns on time series. CLIPSMiner tracks time series of continuous data and sets control points as a quantization method. However, the algorithm considers the time occurrence of the events, organizing the pieces quantized in patterns that have a semantic related to weather events. Algorithm 5 summarizes the CLIPSMiner algorithm.

In the first step, it generates an array containing the differences between the previous and current values of the series, as it can be seen in Figure 7.2. For each time period ($p$)

**Algorithm 5** CLIPSMiner Algorithm

**Input:** Time series $S$; thresholds $\delta$, $\rho$, $\lambda$ and $p$
**Output:** Patterns $V$, $M$, $P$ and time delay $\tau$

1: **for** each time period $p$ of time series $S$ **do**
2: **for all** $v_i$ **do**
3: calculate array of differences $d_i = v_{i+1} - v_i$
4: **end for**
5: **for all** $d_i$ values **do**
6: Find $S_{ea}$ = Set of ascending event sequences
7: Find $S_{ed}$ = Set of descending event sequences
8: Find $S_{es}$ = Set of stable event sequences
9: **end for**
10: Prune $S_{ea}$ and $S_{ed}$ when $\sum d_i < \rho$
11: Prune $S_{es}$ when $\sum d_i < \lambda$
12: **for all** $S_e$ not pruned **do**
13: $V$ = concatenation of $S_{ed}S_{ea}$
14: $M$ = concatenation of $S_{ea}S_{ed}$
15: $P = S_{es}$
16: **end for**
17: Set of all patterns as $[e_{init}, e_i, e_{end}](t_{init}, t_{end})$
18: **for all** Patterns $V$, $M$, $P$ **do**
19: write $v_{init}, v_i, v_{end}, t_{init}, t_{end}$
20: **end for**
21: **for** each pair of patterns array **do**
22: calculate time delay between patterns in different array
23: write time delay $\tau$
24: **end for**
25: **end for**

of time series, CLIPSMiner first calculates an array composed of the differences between previous and current values, i.e. $d_i = v_{i+1} - v_i$ (lines 2 to 4). Setting the parameter $p$, CLIPSMiner divides the time series into $p$ pieces and discovers patterns in accordance with the relevance factor $\rho$ for each period $p_i$. Thus, it is possible to analyze the trend of the series in each period separately.

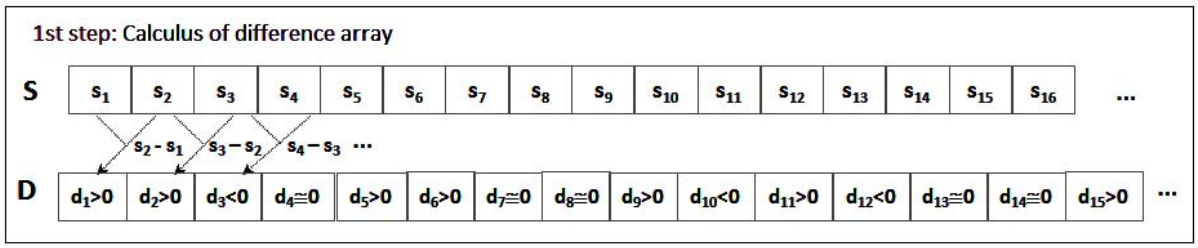



Figure 7.2: First step of CLIPSMiner algorithm. The values of $d_1, d_2, \ldots$ are only for illustration purposes (adapted from Romani et al. (2010d)).

Thus, by analyzing the array of $d_i$, it can be discovered if there is a tendency for rising or falling in the time series, what facilitates discovering the sequence of events. In the next step, the algorithm generates a set of sequences that can be ascending, descending or stable (lines 5 to 9) according to what can be seen in Figure 7.3. Thus, CLIPSMiner

prunes event sequences $S_{ea}$ and $S_{ed}$ smaller than $\rho$, and $S_{es}$ smaller than $\lambda$ (lines 10 and 11).

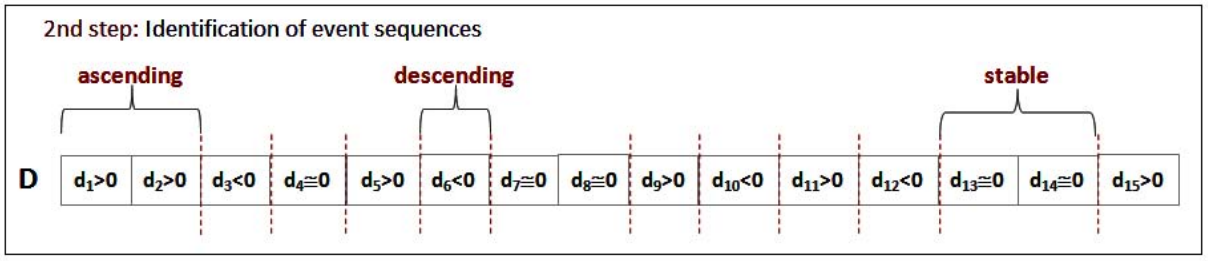



Figure 7.3: Second step of the CLIPSMiner algorithm, finding the ascending and descending patterns. Only one example of each pattern is highlighted in the Figure (adapted from Romani et al. (2010d)).

For each event sequence $S_{e_i}$ not pruned, the algorithm concatenates consecutive sequences $S_{ea}$ and $S_{ed}$ to generate an $M$ pattern, $S_{ed}$ and $S_{ea}$ to generate a $V$ pattern and $S_{es}$ to generate $P$ patterns (lines 12 to 16) as in Figure 7.4. CLIPSMiner stores the mined patterns in an array for each time series $S$. The format of the patterns is an event interval, such as $[v_{init}, v_{mid}, v_{end}]$, where $mid$ is an intermediate value, and the time interval $[t_{init}, t_{end}]$ where the event $e$ occurs (line 17).

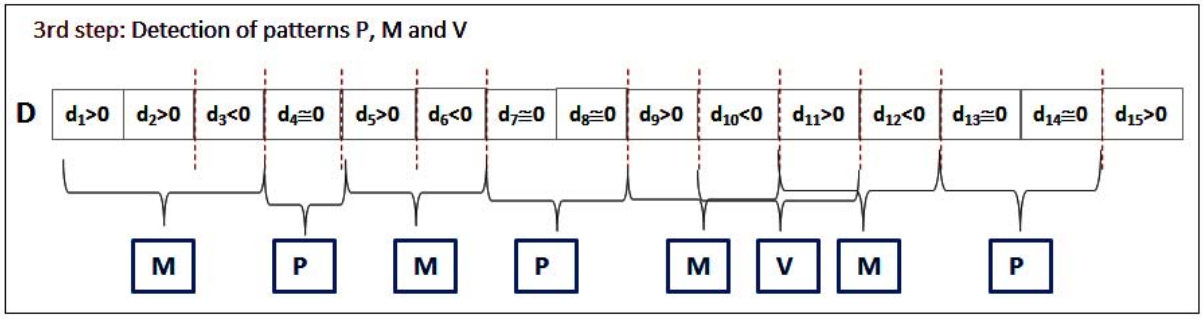



Figure 7.4: Third step of CLIPSMiner algorithm with examples of patterns M, V and P (adapted from Romani et al. (2010d)).

The last step (lines 21 to 24) corresponds to the calculus of the time delay between two time series $S_i$. The algorithm compares the occurrence time of the several intermediate values for the patterns $M$ and $V$ in two series, and calculates the difference between the values. The time delay $\tau$ is the mean value found.

### 7.3.1 Time Complexity

CLIPSMiner reads each of the $n$ events once, where $n$ is the length of the time series. When the events are read, an array of difference values are stored. Each of the $n-1$ values from the differences array are read to discover $M$, $V$ and $P$ patterns. This process is performed

on the $k$ time series stored in the dataset. Thus, the algorithm time complexity is $O(2nk)$ or simply $O(n)$.

## 7.4 Experimental Results

In this section, we discuss representative experiments performed on synthetic and real datasets. In this work, the default value for the parameter was defined empirically and corresponds to $\rho = y * 40/100$. The default value of $\lambda$ is 4 to allow discovering plateaus composed of four consecutive events, which was also empirically defined. The experiments were aimed at evaluating and validating the proposed algorithm. All experiments were made on a computer with 4GB of RAM, an Intel(R) Core(TM)2 Duo 2.66 GHz processor and the Microsoft Windows XP Professional.

### 7.4.1 Evaluating the results

The performance of the CLIPSMiner algorithm was evaluated based on measurements of the time required to process datasets of different sizes. In addition, we assessed the algorithm response to find patterns using several relevance factors.

To assess the quality of results, the CLIPSMiner algorithm was compared with two well-known statistical techniques usually employed by climatologists to analyze climate data: percentile and cross-correlation. Percentile is a measurement of the relative position of one value regarding all other values. The $p$th percentile has at least $p\%$ of the values below that point and at least $(100 - p)\%$ of the values above.

Percentile is widely used in climatology to determine high, very high and extreme values in climate time series. Therefore, this measure was used to compare results of 99th percentile with outputs of CLIPSMiner algorithm tuned to calculate extremes in time series. Figure 7.5 shows a schematic diagram with a comparison between Percentile method and CLIPSMiner algorithm. As it can be seen in Figure 7.5, when Percentile sorts the array in the first step, the algorithm loses the information about time of occurrence. At the end of the processing, Percentile detected only the extreme value. On the contrary, CLIPSMiner finds the extreme with two details: time of occurrence and the context in which this extreme event occurred.

The calculation of the time lag ($\tau$) made by our algorithm was compared with results presented by the cross-correlation method. This technique calculates the correlation between two time series, identifying how much one series must be shifted along the x-axis to make it similar to other one. A qualitative analysis is also presented, and results are analyzed for different values of the relevance factor $\rho$ and length of plateau $\lambda$ patterns.

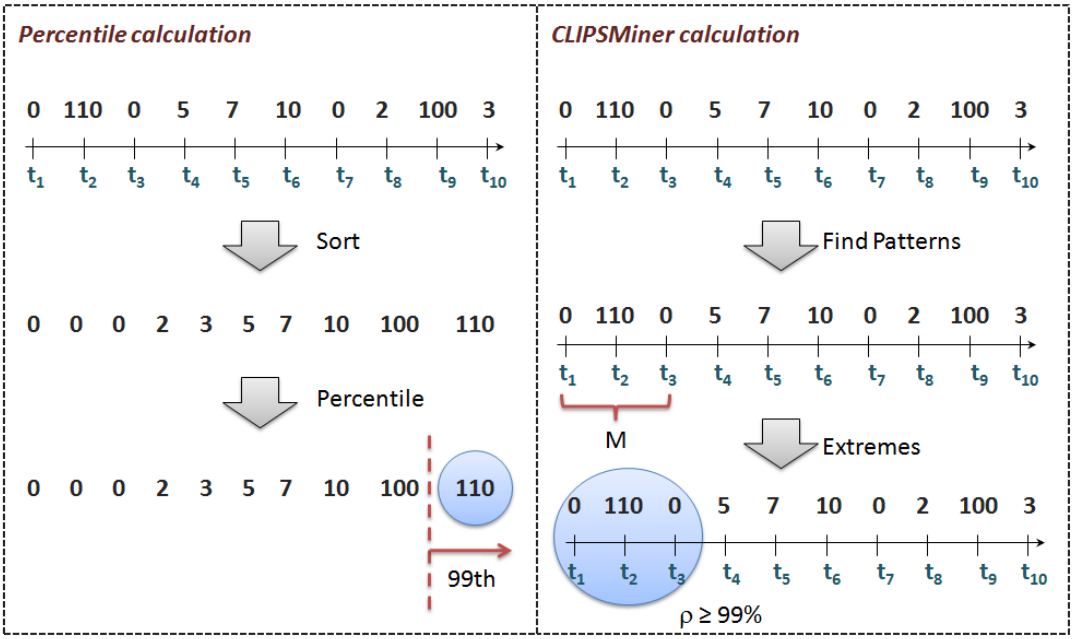



Figure 7.5: Comparison between Percentile and CLIPSMiner calculation through an example of execution.

### 7.4.2 Datasets Description (Synthetic and Real data)

We generated synthetic data (*Synth*) to simulate real climate datasets tendency. By generating synthetic data it was possible to control trends in the time series, which would not be possible if we have used outputs of climate forecasting models. *Synth* dataset is composed of three float attributes: $a_1$ that represents maximum temperature trend, $a_2$ that represents values of minimum temperature and $a_3$ that simulates daily rainfall values. Each attribute varies as follows: $a_1$ from 10 to 45, $a_2$ from -5 to 31 and $a_3$ from 0 to 150.

We have also used two real datasets (*Cps* and *FiveRegions*) composed of climate measures and remote sensing data. Table 7.1 contains the description of datasets used in the experiments.

Table 7.1: Datasets definition

| Dataset | Description | Source | E | N |
|---|---|---|---|---|
| *Synth* | Attributes with similar distribution to climate data | | 3 | 100,000 |
| *Cps* | Real dataset composed of measures of daily rainfall, maximum and minimum temperature from Campinas city (SP, Brazil) collected from 01/01/1890 to 01/31/2009 | IAC | 3 | 41,700 |
| *FiveRegions* | Real data composed of NDVI and WRSI values token from the 5 sugar cane productive areas of São Paulo State (Brazil) from 04/01/2001 to 03/31/2008 | Cepagri | 2 | ≅ 500 |

In the *Cps* dataset the three attributes correspond to daily rainfall value (*rain*), maximum ($t_{max}$) and minimum ($t_{min}$) temperatures measured over a period of 118 years at Campinas, Brazil. In the *FiveRegions* dataset each tuple corresponds to one NDVI mea-

surement per month, for a period of 7 years. There are months in the year, such as January, February and March, where there is no good images to analyze due to clouds coverage in Brazil. In this case, when the data is missing, we interpolate the values using the average. Thus, this dataset has approximately 500 events.

### 7.4.3 Results on the synthetic dataset

The Synth dataset was employed to show the results of our approach over reference data. We have run CLIPSMiner on synthetic dataset to find relevant and extreme patterns. Figure 7.6 shows the performance of the CLIPSMiner algorithm considering two aspects: number of events found and variation of the relevance factor. The graph of Figure 7.6(a) shows the execution time by the number of tuples that varies from 15,000 to 90,000 tuples. The execution time grows linearly from 0.6 milliseconds (ms) to 2.5 ms.

As the relevance factor is increased from 10% to 90%, the execution time decreases, as it can be seen in Figure 7.6(b). Similarly, the number of patterns decreases as the relevance factor increases, as expected (Figure 7.6(c)). When the relevance factor and the plateau length have higher values they make CLIPSMiner more sensitive. That is, the CLIPSMiner algorithm finds only the extreme patterns as $\rho$ and $\lambda$ factors increase.

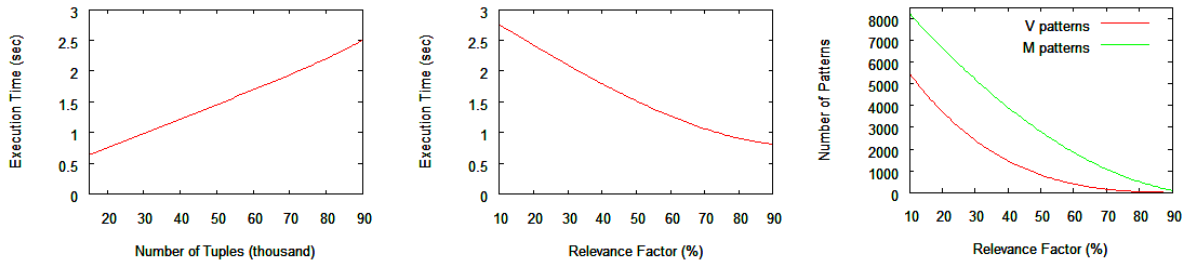



Figure 7.6: Performance of CLIPSMiner algorithm considering the number of patterns found and variation of the relevance factor: (a) execution time by number of tuples (b) execution time by relevance factor (c) number of patterns by relevance factor.

When parameters were set to $\rho = y * 70\%$ and $\lambda = 8$, the detected patterns of type $P$, $M$ and $V$ correspond to relevant patterns with a meaningful variation in the amplitude. For example, Figure 7.7(a) shows a small part of time series considering attribute $a_1$. CLIPSMiner detected the $V$ pattern $[37.06; 10.0; 43.66]$, highlighted in Figure 7.7(a). These patterns are more representative than the negative peaks in the period (27 to 32) and (33 to 36).

In real datasets, these relevant patterns correspond to periods with abrupt decrease in the maximum or the minimum temperature. Generally, small oscillations in temperature are not important to be monitored. However, sudden changes can be interesting to climate researchers. When a variation occurs in a period distinct from the one where such

variations usually occurs, it means that useful knowledge was mined.

When we increase the value of the relevance factor ($\rho$) to 95% of amplitude, CLIPSMiner becomes more restrictive and only very extreme patterns are detected. Figure 7.7(b) presents a graph with a pattern $M$ detected in the time series considering the attribute $a_3$. CLIPSMiner found a pattern $M$ in the period from 9 to 11, which has an amplitude greater than others in the time series. In real databases, this pattern is similar to an increase in the amount of rainfall, for example.

Many days without rain especially during plant growth stages, which are sensitive to water deficit can be worrying for farmers and government. Some agricultural crops can be damaged by water deficit and occasionally, it is necessary to use irrigation. Other crops may resist longer, but they also need monitoring. To monitor this pattern, agrometeorologists could find periods in which this phenomenon usually occurs, altering the $\lambda$ (plateau lenght) value. Increasing the $\lambda$ value allows to find the most extreme phenomena. In time series $a_3$, for example, when we increased $\lambda$, the number of $P$ patterns found decreased from 31 to 7. Consequently, the option to filter the result, by dynamically setting the parameters according to analysts needs is a big difference of the algorithm CLIPSMiner, as compared to others from the literature.

The CLIPSMiner algorithm allows to analyze time series for specific periods when the parameter $p$ is set. Thus, we defined periods of 50 years that generated 5 different periods. We executed CLIPSMiner with $\rho = y * 95\%$ of amplitude to return only the most extreme patterns. In the first two periods, patterns were generated with maximum values above 122 for attribute $a_3$. In the last two periods, the maximum values found in the $M$ patterns were above 120. Assuming that $a_3$ represents rainfall in real data, we evaluated the extreme daily rain for every 50-year period, and could perceive if changes occured in the time series trend. Any period of time could be set.

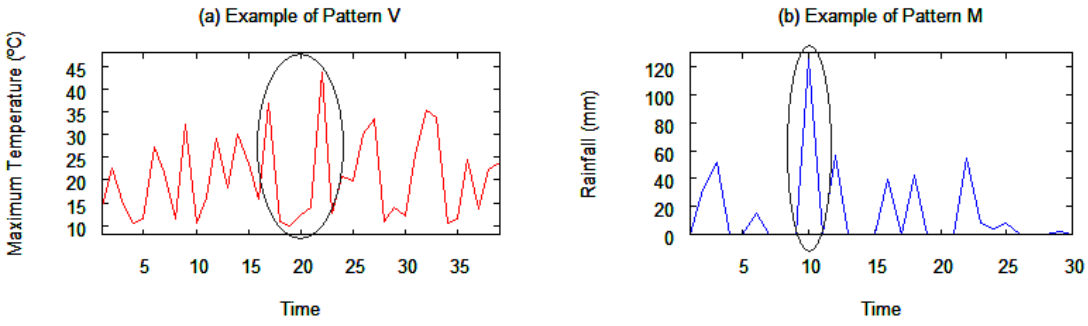



Figure 7.7: Example of extreme patterns: (a) V pattern similar to a negative peak (period 17 to 22) and (b) M pattern similar to a positive peak (9 to 11).

Comparing this result with the output generated by the percentile method, similarly we defined periods of 50 years that generated 5 different periods. We set the Percentile

algorithm to 99th to find extremes. This method detected only a value that represents extremes in each 5 parts of the time series. The Percentile method efficiently finds extremes, but does not provide information on the period where the extreme event occured, neither about preceding and subsequent events, as CLIPSMiner does.

The CLIPSMiner algorithm outputs showed that the majority of the extreme patterns occured in the range from 0 to 120 and back to 0. In the case of real data, this would be similar to the occurrence of heavy rain in a single day. In order to identify all possible extremes in a series, we included an option $-extreme$ to search beyond the standard M, P and V, situations in which extreme events occur. For example, if an interval as $[50, 130, 79]$ occurs, the algorithm does not return this event as far as the relevance factor would be higher than the difference found in this range. Thus, when we set $-extreme$ parameter, patterns equal to those of the example would also be returned as possible output.

Our algorithm also calculates the time delay between two time series, i.e., the time correlation between them. For the synthetic dataset, the $\tau$ value ranged from 0 to 11 time lags, because CLIPSMiner searches for lags in different parts of time series. Using the cross-correlation algorithm, no lags were found between the two time series.

### 7.4.4 Results on Real Data - The *Cps* dataset

The same process described for the *Synth* dataset was employed. Two experiments were executed assigning the default and maximum values for parameters $\rho$ and $\lambda$, in order to find the relevant and the extreme patterns from the time series. For both experiments, the $\delta$ value was set to 0.9. Figures 7.8(a) and (b) show the number of patterns discovered for the three time series of the *Cps* dataset when parameters were set to discover relevant patterns. The parameters values to discover relevant patterns were:

- $t_{min}$: $\rho = y * 45\%$ and $\lambda = 10$
- $t_{max}$: $\rho = y * 55\%$ and $\lambda = 20$
- $rain$: $\rho = y * 60\%$ and $\lambda = 30$

The parameters are different due to the range of variation of time series. Thereafter, we have increased the value of the parameters by 10 to find extreme patterns.

Analyzing the results presented in Figure 7.8, we can observe a meaningful decrease of patterns when $\rho$ and $\lambda$ values increase, i.e. they become more restrictive. Patterns of type $V$ in time series $t_{max}$ drop from 11 to 3. This pattern represents variations in maximum temperature in different periods of time. Table 7.2 shows the $V$ patterns found in time series $t_{max}$ using $\rho = y * 55\%$.

Table 7.2: $V$ patterns found for $t_{max}$ in Cps dataset considering relevance factor of 55%

| # patterns | $t_{max}$ values | date |
|---|---|---|
| 1: | [31.8; 13.6; 30.2] | [08/27/1912-09/07/1912] |
| 2: | [27.0; 10.0; 26.3] | [06/23/1918-07/01/1918] |
| 3: | [31.0; 14.5; 31.5] | [09/28/1918-10/06/1918] |
| 4: | [32.2; 13.4; 31.4] | [09/02/1933-09/13/1933] |
| 5: | [32.0; 15.6; 32.8] | [09/09/1948-09/17/1948] |
| 6: | [26.7; 09.7; 27.4] | [06/16/1952-06/24/1952] |
| 7: | [34.3; 16.1; 32.6] | [10/27/1959-11/04/1959] |
| 8: | [33.0; 14.2; 32.2] | [09/12/1990-09/19/1990] |
| 9: | [28.8; 12.8; 28.4] | [07/29/1993-08/03/1993] |
| 10: | [34.4; 15.8; 31.4] | [09/07/1999-09/14/1999] |
| 11: | [37.2; 19.2; 35.8] | [10/21/2007-10/29/2007] |

As it can be seen in Table 7.2, CLIPSMiner discovered relevant $V$ patterns in time series $t_{max}$, in the period between June to October, which is the Winter season and the beginning of Spring in South America. In general, such fall in the maximum temperature in Brazil is associated to cold fronts that come from the South. The proposed algorithm aims at mining a huge amount of data evidencing patterns according to a threshold that can be properly set by experts, in order to be more or less restrictive, depending on the analysts' intents.

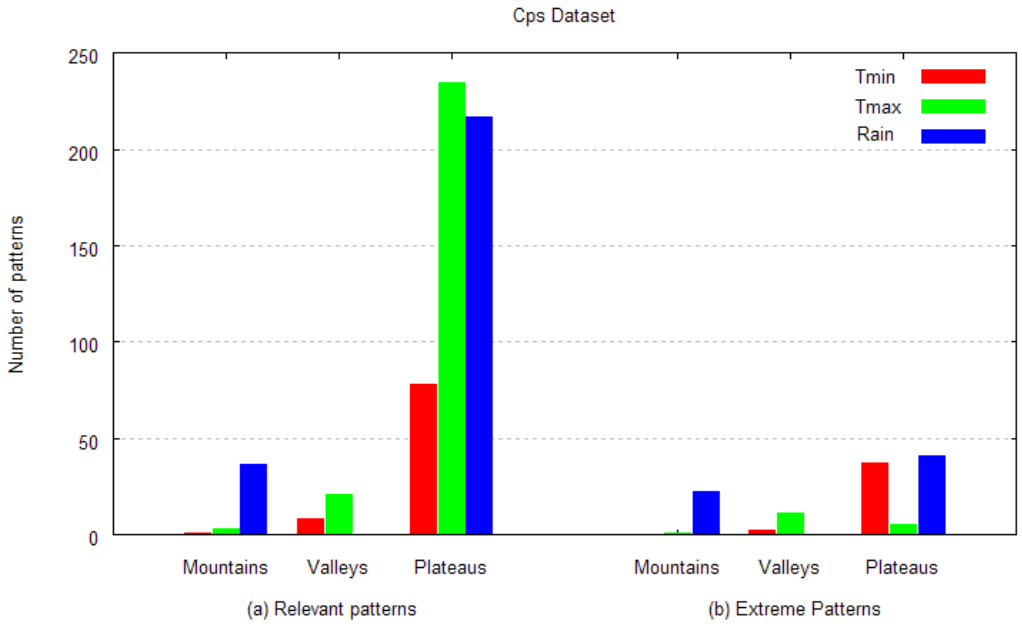



Figure 7.8: Results for *Cps* dataset: y-axis represents the number of patterns and the type of patterns are represented in x-axis. (a) number of relevant patterns and (b) quantity of extreme patterns.

Table 7.3 contains the $M$ patterns found in time series *rain* using $\rho = y * 70\%$. The results show a high increase in the rainfall volume in a short interval of time. This extreme phenomenon causes serious problems such as floods. Researchers are interested

Table 7.3: $M$ patterns found for *rain* in Cps dataset

| # patterns | *rain* values | date |
|---|---|---|
| 1: | [0.0; 103.0; 0.0] | [24/01/1899-27/01/1899] |
| 2: | [0.0; 119.0; 0.0] | [11/13/1923-11/15/1923] |
| 3: | [0.0; 142.4; 0.0] | [12/23/1925-12/26/1925] |
| 4: | [0.0; 127.7; 1.6] | [12/23/1949-12/25/1949] |
| 5: | [0.0; 107.0; 0.0] | [11/23/1951-11/27/1951] |
| 6: | [0.0; 115.7; 0.0] | [01/01/1982-01/04/1982] |
| 7: | [0.0; 108.3; 0.0] | [03/07/1987-03/12/1987] |
| 8: | [8.0; 138.2; 0.0] | [12/30/1989-01/04/1990] |
| 9: | [0.0; 107.6; 0.0] | [12/24/1997-12/26/1997] |
| 10: | [0.0; 144.7; 0.0] | [10/01/2001-10/04/2001] |
| 11: | [0.0; 138.5; 11.4] | [01/18/2005-01/21/2005] |

in finding out when such phenomena occurred in time series and the intensity of rainfall that occurred in a few days. Nowadays, these extreme events have occurred with greater frequency and seem to be associated to climate change (Alexander et al., 2006; Ganguly & Steinhaeuser, 2008).

According to the results, extreme rainfall started to reach values above 115 mm from 1923, which coincides with previous statistical analysis made by meteorologists using the Percentile method, as illustrated in Figure 7.9. This fact confirms the hypothesis of the researchers that the distribution of rainfall has increased in a fast pace, in the last decades.

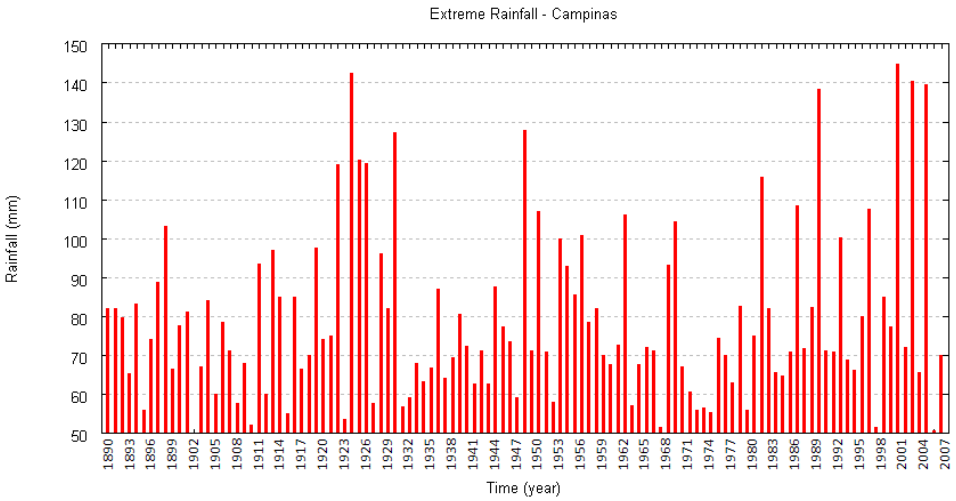



Figure 7.9: Graph with extreme rainfall per year in the Campinas region with 17 values above 100 mm.

Many $P$ patterns were found in all time series, specially for *rain*. This dataset is composed of daily values of temperature and rain. Thus, the algorithm detected periods with low variation in temperature or days without rain. Changing the values of parameters $\delta$ and $\lambda$, CLIPSMiner discovered prolonged droughts, that is a pattern studied by experts, because of their consequences for agriculture. Especially, when droughts occur in periods

when they are not expected. This type of pattern found by the CLIPSMiner algorithm is not detected by the percentile method. Experts commonly use other techniques such as the generation of indexes that determine extended dry periods and other extreme phenomena.

In order to analyze the trend differences in the beginning and at the end of the precipitation series, we defined the parameter $p$. In this experiment, we set $p$ to 10 years and $\rho$ to the maximum value (90%). Analyzing the results, we see that at the beginning of the time series (1901 to 1908) the maximum precipitation reached values of approximately 80 mm. After the 90s, these extreme values are above 130 mm, as seen in Figure 7.10.

| (a) Beginning of time series | | (b) End of time series | |
|---|---|---|---|
| *Rain* | *Date* | *Rain* | *Date* |
| [0.2; **81.0**; 0.0] | [12/02/1902-15/02/1902] | [8.0; **138.2**; 0.0] | [30/12/1989-04/01/1990] |
| [0.0; **84.0**; 0.0] | [03/01/1905-06/01/1905] | [0.0; **144.7**; 0.0] | [01/10/2001-04/10/2001] |
| [0.0; **81.0**; 0.0] | [05/07/1905-08/07/1905] | [0.0; **138.5**; 11.4] | [18/01/2005-21/01/2005] |

Figure 7.10: Extreme rain values for the beginning and the end of time series: (a) rainfall reached values of 80 mm approximately in the beginning of time series (1901 to 1905), (b) rainfall values increased to 130 mm at the end of time series (1989 to 2005).

To compare this result with the output generated by the percentile method, similarly we defined periods of 10 years. We set the Percentile algorithm to 99th to find extremes. This method detected only a value that represents extremes in each period of 10 years, as it can be seen in Figure 7.11. It found smaller rain values at the beginning of the time series than at the end, which shows that extreme climate phenomena has become more intense in recent decades.

Beyond getting the extreme events, the Percentile method does not provide information on the period where the extreme event occurred neither about preceding and following events, as CLIPSMiner does.

### 7.4.5 Results on Real Data - the *FiveRegions* Dataset

In this experiment, CLIPSMiner has detected more $M$ and $V$ patterns than plateaus ($P$), because the time interval was set to be monthly. Figures 7.12(a) and (b) summarize the patterns detected when the parameters were set to be smaller ($\rho = y * 10\%$ and $\lambda = 3$) and more sensitive ($\rho = y * 70\%$ and $\lambda = 5$), respectively.

CLIPSMiner found few $M$ patterns using default values for parameters in NDVI time series. The patterns detected were $[0.24; 0.64; 0.30]$ in $[09/2003 - 10/2004]$ for Jaboticabal,

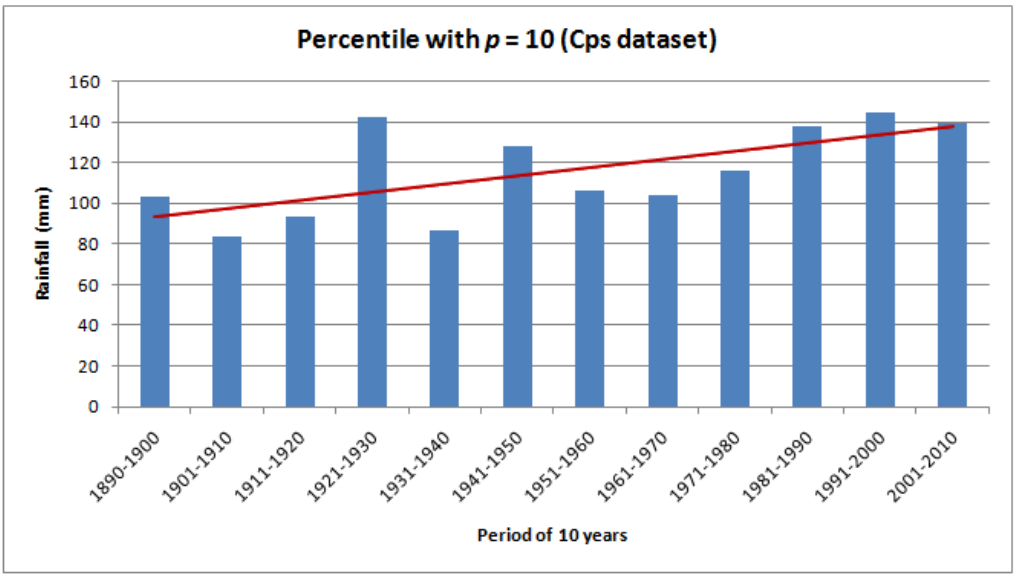



Figure 7.11: Extremes in each period of 10 years for rainfall in Campinas region (Cps dataset).

$[0.29; 0.61; 0.23]$ in $[10/2004 - 10/2005]$ and $[0.23; 0.61; 0.26]$ in $[10/2005 - 10/2006]$ for Jaú and $[0.26; 0.61; 0.26]$ in $[10/2002 - 09/2003]$ for Sertãozinho.

These $M$ patterns are related to periods when the green biomass reaches its highest values, before the sugar cane harvest that begins in April in the study area. $P$ patterns were found in WRSI time series. It corresponds to a small variation in the WRSI index, such as $[0.95; 1.0; 0.99]$ in $[10/2001 - 03/2002]$ found in the Jaboticabal dataset. This phenomenon occurs when the maximum soil water content is reached after a long period of rainfall.

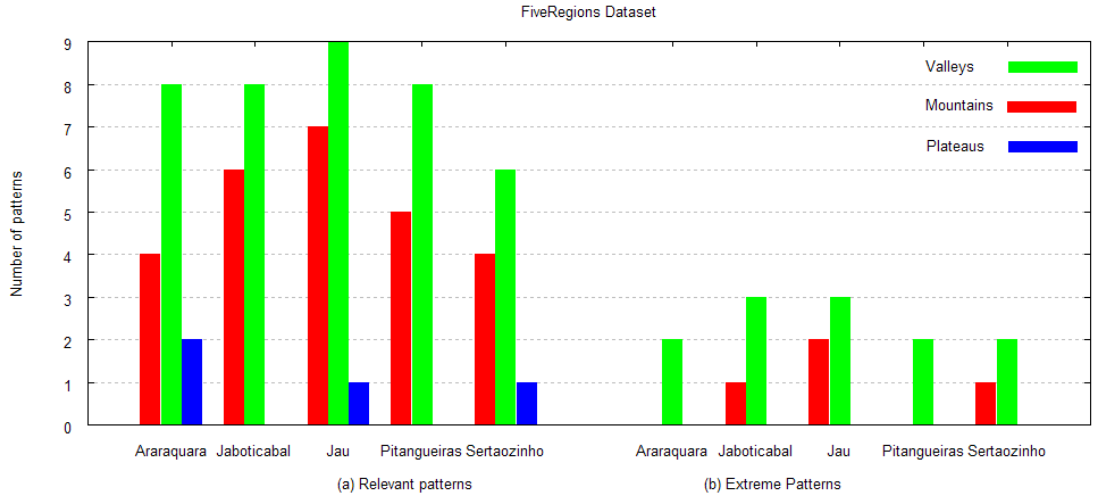



Figure 7.12: Results for *FiveRegion* dataset: y-axis represent the number of patterns and the type of patterns are represented in x-axis. (a) number of relevant patterns and (b) quantity of extreme patterns.

The cross-correlation method was calculated for two time series (NDVI and WRSI)

and presented two months of time lag. The CLIPSMiner algorithm showed that there are different time lag values along time series, as it can be seen in Figure 7.13.

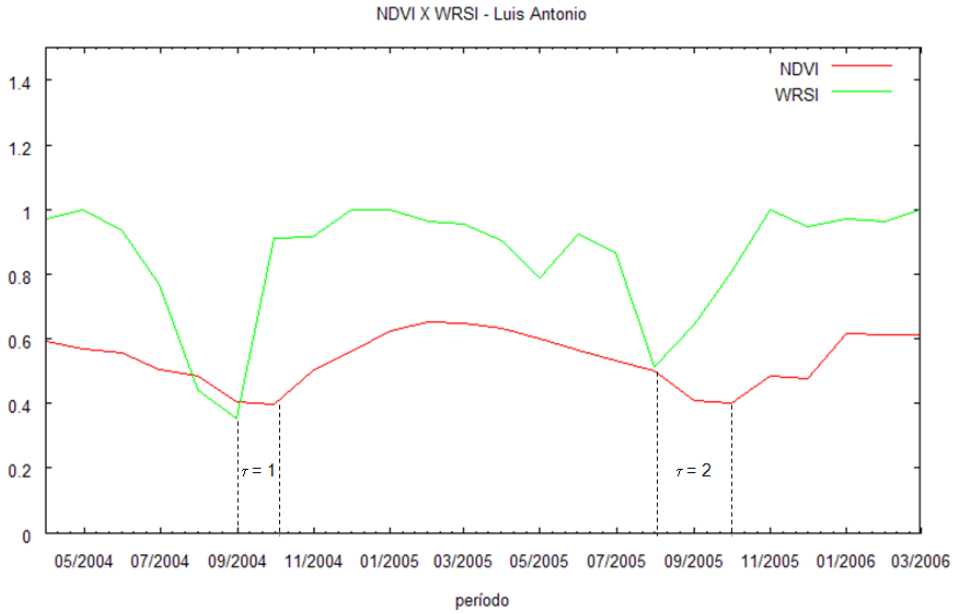



Figure 7.13: WRSI and NDVI time series from Luis Antônio with example of two different time lags ($\tau = 1$ and $\tau = 2$).

CLIPSMiner found $\tau$ equals 1, 2 or 3 depending on the period. It means that there are correlations between NDVI and WRSI with a delay of one, two or three months. The time delay is a relevant asset that can be used to mine different occurrences of correlations, spotting issues not expected by the specialists.

## 7.5 Summary

In this chapter, we presented CLIPSMiner, a new unsupervised algorithm to find relevant and extreme patterns in climate time series, as well as correlations between time series, showing the relationship between the series and when one affects the other. The experiments results show that CLIPSMiner is a powerful technique to analyze long and multidimensional climate time series. This algorithm works on multiple time series of continuous data, identifying sequential patterns defined with time constraints that are related to climate phenomena. The parameters can be dynamically tuned by the user, allowing the specialist to set the size or scale of the pattern to be mined in order to find extreme phenomena, or even to analyze parts of the series. Thus, CLIPSMiner gives more freedom and control to the user to analyze more closely the dataset.

The patterns detected preserve the semantics of climate events. Thus, the patterns $M$, $V$, $P$ can summarize the series and be used to index and to detect correlations be-

tween series, as well as to provide a simple way to discretize a long and continuous time series, keeping the temporal meaning of the patterns. The correlation among time series considering time windows are also provided by CLIPSMiner, what the traditional method of cross-correlation fails in providing to the specialists.

In summary, the results showed that the algorithm detects patterns known in climatology, which are manually detected by specialists and are time expensive. CLIPSMiner does that automatically, in linear time regarding the size of the dataset. Moreover, patterns detected using the highest relevance factor are coincident with extreme phenomena as many days without rain or heavy rain. This feature allows CLIPSMiner be used to compare real datasets with outputs of forecasting models in order to assist in climate change research.

# Chapter 8

# The CLEARMiner Algorithm

## 8.1 Introduction

In Chapter 5, we described the Apriori-FD method to find association rules in climate and remote sensing time series. Although the method has generated satisfactory rules for minimum and maximum temperatures, the rules generated for rainfall associated to the NDVI did not present a coherent association. As NDVI is an index that is related to the green biomass, the effect of rain on the crop growth could be detected by NDVI after a period of time. This fact is supported by studies described in (Gonçalves et al., 2009; Avila et al., 2009) that evidence the interest of agrometeorologists in associating remote sensing and climate data to better understand the influence of climate in development of agriculture crops. Statistical analysis performed by Gonçalves et al. (2009) showed that there are correlations between NDVI and rainfall with a delay of one or two months.

In this context, we focus on the association of local patterns in time series pairs with the purpose of improving yield forecasting of agricultural crops and increasing the sustainable usage of soil. Accordingly, we consider the problem of finding rules that associate patterns in a remote sensing time series to other patterns in climate series considering time delay. Examples of rules relating two or more time series could be "*a period of gradual increase in the WRSI values is followed by an increase in NDVI values*" or "*in years when El Niño is strong could occur rainfall above average in the Southern Brazilian region*".

As a solution, we propose a new unsupervised algorithm for mining association patterns on heterogeneous time series integrated to a remote sensing information system. The time series mining module was developed to generate rules considering a time lag. To do so, we define the constraint of time-window to find association rules that are extracted in two steps. First, the algorithm transforms multiple time series in a representation of patterns (*peaks*, *mountains* and *plateaus*), with discrete intervals that maintain the time occurrence

and represent phenomena on climate or remote sensing time series. In a second step, the algorithm generates rules that associate patterns in multiple time series with qualitative information.

This chapter is organized as follows. In Section 8.2 we present the information mining system where the CLEARMiner (ClimatE Association patteRns Miner) algorithm is embedded. Thus, we detail a theoretical formalization for association rules in Section 8.3. Section 8.4 presents and discusses the experimental results. Finally, Section 8.5 contains the conclusions.

## 8.2 Architecture of the RemoteAgri System

In order to support the mining of NOAA-AVHRR multi-temporal images associated to climate series to contribute to the advancement in agriculture research at a regional scale, we have developed the RemoteAgri system. Figure 8.1 shows a schematic diagram of the system prototype consisting of three major components: image geo-referencing module, time series extraction module, and time series mining method.

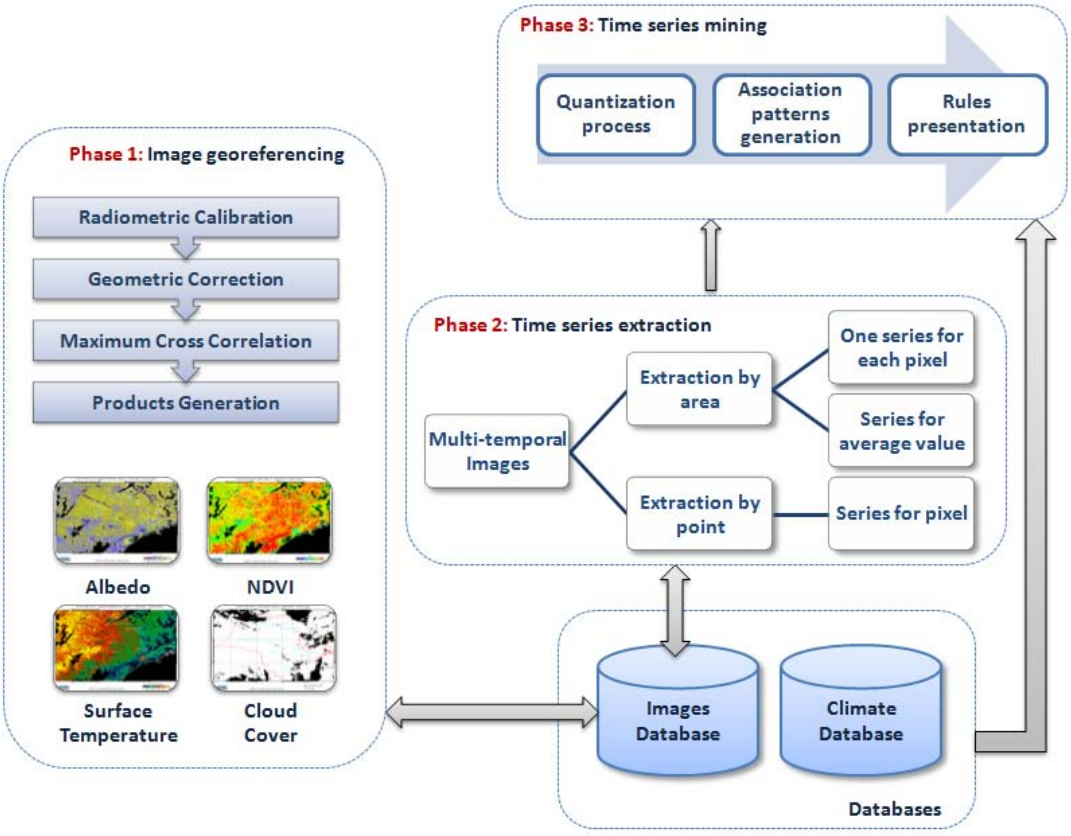



Figure 8.1: Schematic diagram of the multi-temporal image mining RemoteAgri system.

The first phase to be executed in the system corresponds to the image geo-referencing

module, that is presented in Figure 8.1. This module is executed calling the subroutines of NAV system (Emery et al., 1989; Esquerdo et al., 2006), in batch mode, to accomplish necessary tasks to geo-reference NOAA-AVHRR images. The geo-referencing module allows users to generate four different synthesis images: albedo, NDVI, surface temperature and cloud coverage for a specific region, as is illustrated in Figure 8.1.

As the volume of images is huge, an extraction module (phase 2 in the Figure 8.1) called *SatImagExplorer* was developed to automatically extract time series from images (Chino et al., 2010b,a). This module extracts values or computes indexes from a given image. Then, it generates a time series computing the index values for all images listed in the interface using the same coordinate (latitude/longitude) of the selected region in the original image. The user can select regions in the image through the mouse and/or a coordinates file. If the user needs a time series that represents a specific region, the system generates one time series for each pixel in the defined area and the average value considering all pixels. All time series extracted from the images are stored in the database. Details about the system are presented in Appendix A.

The last phase refers to time series mining module developed to associate climate data with indexes extracted from NOAA-AVHRR images. In the next section, we describe in details the three parts of this module.

## 8.3 Description of CLEARMiner

In this work, we present a new unsupervised algorithm, called CLEARMiner, to mine association patterns from time series extracted from NOAA-AVHRR. The process of time series mining was divided into three parts: *quantization process*, *association patterns generation*, and *rules presentation* as it can be seen in Figure 8.1.

First, time series are re-written in a symbolic representation that is more succinct and manageable than continuous data. We propose the use of patterns similar to positive and negative peaks, as well as plateaus that maintain the information about continuous data and time of occurrence, as was introduced in Chapter 7. The proposed algorithm renders a quantization process that preserves the time series semantics. The second part is related to the generation of rules from this symbolic representation. Finally, the third part corresponds to the presentation of association patterns in two formats: short and detailed.

### 8.3.1 Quantization Process

This process is similar to those described in Chapter 7 - Section 7.2. Figure 8.2 presents a summarization of the quantization process, which is divided in three steps in order to transform time series in a symbolic representation without losing semantic and temporal information.

As it can be seen in Figure 8.2, the V (valley) corresponds to a pattern defined as the concatenation of a descending event sequence and an ascending event sequence (i.e., $V = S_{ed}S_{ea}$). P (plateau) represents a kind of pattern described as a stable event sequence (i.e., $P = S_{es}$), while M (mountain) indicates a pattern generated by the concatenation of an ascending event sequence and a descending event sequence (i.e., $M = S_{ea}S_{ed}$).

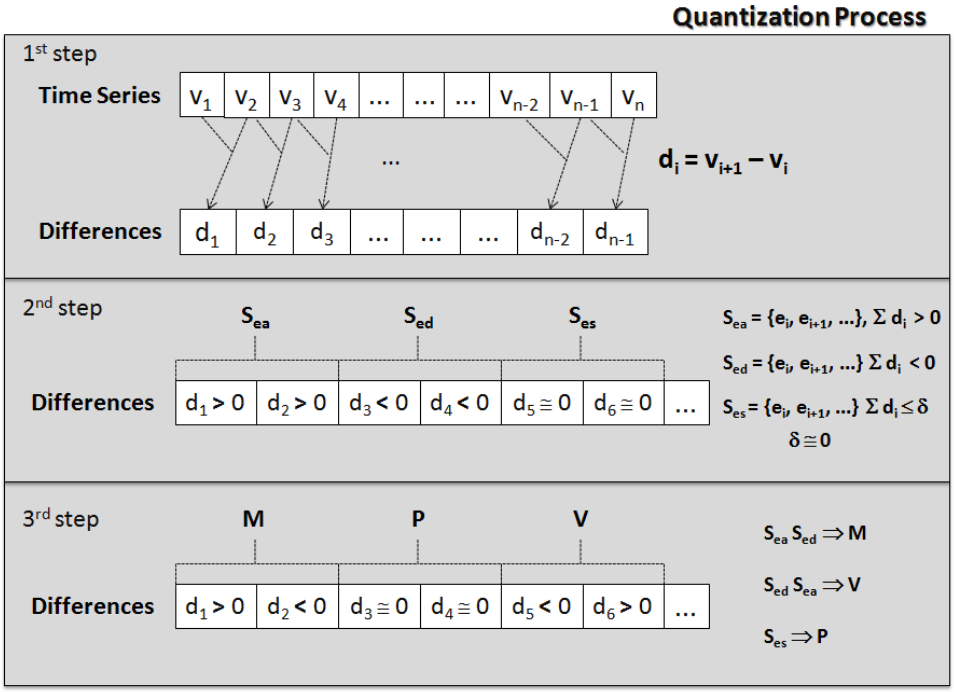



Figure 8.2: Representation of the three steps of quantization process. (1) Calculation of differences between previous and current values of time series, (2) Identification of ascending, descending and stable event sequences and (3) Detection of patterns M, V and P.

### 8.3.2 Association Patterns Generation

After the quantization process, time series are converted into a set of patterns V, M and P, but the complete format with data value and time is preserved. To understand the relationship between several time series, we define an association rule as:

*if one pattern occurs at period i in time series 1, then*
*another or the same pattern occurs at period j in time series 2.*

It means that a pattern in one time series can be associated to patterns in other time series. We consider an association rule as an expression of the form $S_i[\alpha] \Rightarrow S_j[\beta]$, where $S_i$ and $S_j$ are different time series, $\alpha$ and $\beta$ are frequent patterns.

The frequency of a pattern is the number of times the pattern occurs in the time series and is denoted by $fr(S_i[< pattern >])$. We have defined two metrics to assess the rules: *support* and *confidence*. Support of $S_i[\alpha] \Rightarrow S_j[\beta]$ represents the frequency of occurrences and is given by Equation 8.1.

$$support = \frac{fr(S_i[\alpha] \cup S_j[\beta])}{T} \tag{8.1}$$

where $fr(S_i[\alpha] \cup S_j[\beta])$ corresponds to the total number of input-patterns in the dataset that contains $\alpha \cup \beta$, and T is the total number of patterns in the dataset.

Differently from traditional algorithms of association rule mining that consider T as the total number of transactions in a database, we define T as a function of the number of patterns in the time series converted into a sequence of symbolic patterns. Thus, we define T by Equation 8.2.

$$T = \sum_{i=0}^{m-1}(n - i) \tag{8.2}$$

where $m$ and $n$ correspond to the size of the first time series and of the last one, respectively, converted into a sequence of symbolic patterns, for all $(n - i) > 0$.

The rules can be generated for the complete series, which greatly increases the number of generated rules or considering a sliding window of size $w$ that is defined by the number of patterns. This parameter can be changed by the user, depending on how far he/she wants to analyze. In general, the value of $w$ is small because specialists are more interested in knowing the correlation between two series in a short period of time to understand the correlation between specific episodes in different series. If rules are calculated for a window of size $w$ (same size for all series), we have T calculated by Equation 8.3.

$$T = \sum_{i=0}^{w-1}(n - i) * \frac{m}{w} \tag{8.3}$$

where $m$ and $n$ correspond to the size of the first time series and the last one, respectively, converted into a sequence of symbolic patterns, for all $(n - i) > 0$ and $w$ is the size of window defined by user.

For example, if a dataset contains 96 patterns and 45 patterns correspond to $S_1[V] \cup S_2[M]$, the $support(S_1[V] \cup S_2[M]) = 0.46$ (46%). Given a minimum support (*min_sup*) specified by the user, we say that a pattern is frequent if it occurs more than *min_sup*

times. Thus, frequent patterns are used to generate rules as described.

The confidence measure indicates the chance of $S_j[\beta]$ occurring if $S_i[\alpha]$ also occurs. The confidence for the rule $S_i[\alpha] \Rightarrow S_j[\beta]$ is given by Equation 8.4.

$$conf = \frac{fr(S_i[\alpha] \cup S_j[\beta])}{fr(S_i[\alpha])} \tag{8.4}$$

Given a user-specified minimum confidence (*min_conf*), rules are generated if they satisfy the conditions $support \geq min\_sup$ to discover frequent patterns and $conf \geq min\_conf$.

The CLEARMiner algorithm first converts time series into a sequence of three types of patterns (V, M and P) that are relevant and meaningful to agrometeorological researches. In the same time, the algorithm considers the occurrence time of events, organizing the pieces quantized in patterns that have a semantic related to weather events. After, CLEARMiner generates rules for the full time series or by window of size $w$. Algorithm 6 shows a pseudo-code for CLEARMiner.

**Algorithm 6** CLEARMiner Algorithm

**Input:** Dataset $A$ of $k$ time series structured as $\{e_1, e_2, ..., e_n\}$ where $e_i$ is an event of time series $S_i$; thresholds $\delta$, $\rho$, $\lambda$ and $w$
**Output:** The mined rules
1: Scan dataset $A$
2: **for** each time series $S_i$ **do**
3: PatternsFind($S_i$, $\delta$, $\rho$, $\lambda$)
4: **end for**
5: $F_1$ ={1-frequentPattern ($S_i[< pattern >]$)}
6: **for** $p = 2; p \leq m; p = p + 1$ **do**
7: $C_p$ = Set of candidate p-frequentPattern
8: ($S_i[< pattern >]S_j[< pattern >]$ and so on)
9: **for all** input-frequentPatterns in the dataset **do**
10: increment count of all p-frequentPattern $\in C_l$
11: **end for**
12: $F_p = \{frequentPattern \in C_p |$
13: $sup(frequentPattern) \geq min\_sup\}$
14: **end for**
15: **for all** $w$ **do**
16: RuleGenerate($F_p$, min_conf)
17: **end for**

The *PatternsFind* module is called to find patterns and to generate an array of patterns for all series. The pseudo-code for *PatternsFind* was already presented in Alg. 5 - Chapter 7.

The CLEARMiner algorithm calculates j-frequentPatterns for each time series. For example, if a dataset contains three time series, a 2-frequentPattern time series can be $S_1[P]S_2[V]$ or $S_1[V]S_2[M]$, i.e., a frequentPattern combines patterns of different time series. The algorithm only stores j-frequentPatterns greater than the *min_sup* threshold defined by the user (lines 5 to 14 - Alg. 6). This step is illustrated in Figure 8.3(a).

To calculate the support of the pattern $S_1[M] \cup S_2[P]$, we first calculate the frequency of the pattern $S_1[M] \cup S_2[P]$, counting the number of times that $S_1[M]$ is associated to $S_2[P]$, that is equal to 9. Then, we calculate T by Equation 8.2, where $m = n = 8$, i.e. both time series $S_1$ and $S_2$ have the same number of patterns. As a consequence, $T = \sum_{i=0}^{8-1}(8-i) = \sum_{i=0}^{7}(8-i) = 36$. Using Equation 8.1, the support $sup = fr(S_1[M] \cup S_2[P])/T = 9/36 = 0.25 = 25\%$ to the pattern $S_1[M] \cup S_2[P]$. As the $sup \geq min\_sup$, the pattern $S_1[M] \cup S_2[P]$ is selected as frequent.

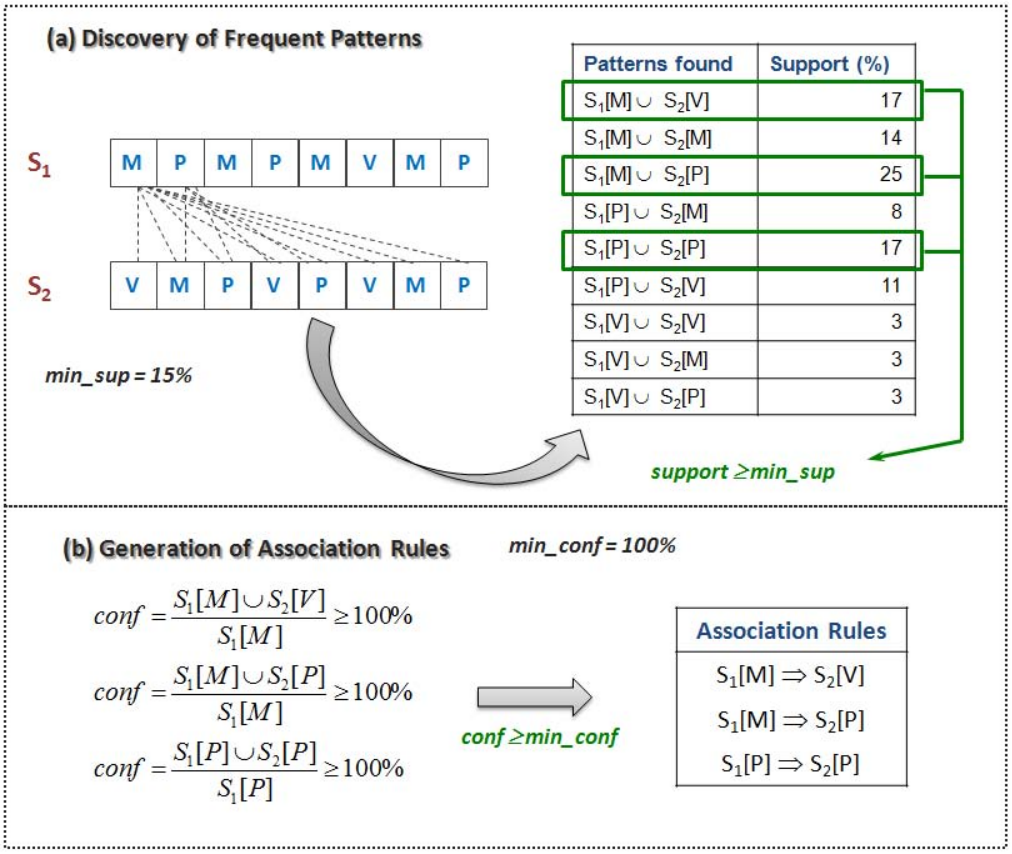



Figure 8.3: Diagram illustrating the steps for rules generation (a) Example of the Frequent Patterns Discovery Process. (b) Example of Association Rules Generation.

**Algorithm 7** RuleGenerate Method

**Input:** $F_p$ and min_conf
**Output:** The mined rules
**for all** frequentPattern $S_i[\alpha]$ and $S_j[\beta] \in F_p$ **do**
2: $conf = fr(S_i[\alpha] \cup S_j[\beta])/fr(S_i[\alpha])$
**if** $conf \geq min\_conf$ **then**
4: output the rule $S_i[\alpha] \Rightarrow S_j[\beta]$ and $conf$
**end if**
6: **end for**

For each frequent pattern in $F$, the algorithm calculates, via the *RuleGenerate* method, the confidence value (line 2 - Alg. 7). If confidence is greater than *min_conf*, it generates

rules (lines 3 to 5 - Alg. 7), as it can be seen in Figure 8.3(b).

### 8.3.3 Rules Presentation

To better visualize the rules, the algorithm presents them in two formats: short (a simple and succinct way to represent the rules) and extended (all the details are provided).

The short format is aimed at a fast and easy analysis of the series. However, it contains no information about the context in which the phenomenon occurred. Examples of rules that can be generated are:

$$S_1[V] \Rightarrow S_2[M]$$

In the first example, the rule indicates that a decrease in the time series 1 is associated to an increase in the other series ($S_2$). In addition to the rules in short format, the CLEARMiner algorithm generates association rules in extended format as well. An example is:

$$S_1[v_i, v_k, v_n](t_{init_1} - t_{end_1}) \Rightarrow S_2[v_j, v_l, v_m](t_{init_2} - t_{end_2})$$

This rule indicates that the pattern $[v_i, v_k, v_n]$ occurred in the period $(t_{init_1} - t_{end_1})$ for the time series $S_1$, which is associated to the pattern $[v_j, v_l, v_m]$ occurred in the period $(t_{init_2} - t_{end_2})$ for the series $S_2$ with $t_{init_1} \leq t_{init_2}$ and $t_{end_1} \leq t_{end_2}$. Thus, the user can analyze rules in the short format to verify correlations between time series and use the extended format to obtain more details. An example with real data is presented in Figure 8.4.

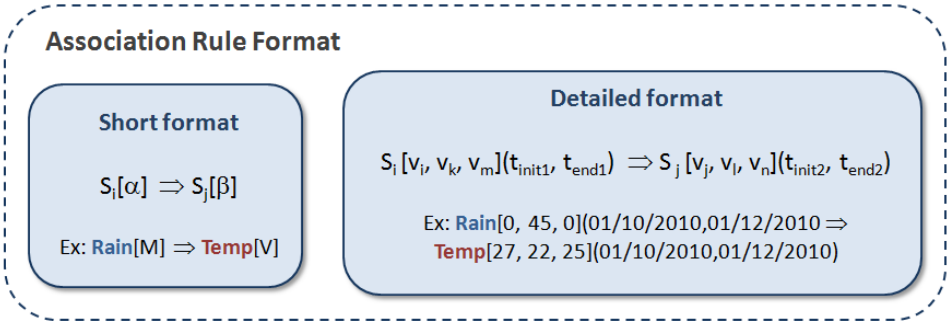



Figure 8.4: Examples of rules in short and extended format that represent a peak of rain from 0 to 45 and returning to 0 occurred between the period of 01/10/2010 and 01/12/2010, which is associated to negative peak of temperature from 27°C to 22°C returning to 25°C.

### 8.3.4 Time Complexity

CLEARMiner reads each of the $n$ events once, where $n$ is the length of the time series. When the events are read, an array of difference values are stored. Each of the $n-1$ values from the differences array are read to discover $M$, $V$ and $P$ patterns. This process is performed on the $k$ time series stored in the dataset. Afterwards, the association rules are generated for each two series, regarding the number of patterns (m < n) and window size (w). Thus, the complexity of the algorithm is $n$ multiplied by a small constant ($kmw$). Consequently, the algorithm runs in linear time ($O(n)$).

## 8.4 Experimental Results

Experiments were performed on real datasets, and the results from two of them are detailed as follows. The results from such experiments followed the specialists' expectations and helped on tuning the algorithms' parameters. Table 8.1 presents a summary of the datasets used, giving their dimensions ($E$) and the size of time series ($N$).

Table 8.1: Datasets definition

| Name | Description | Source | E | N |
|---|---|---|---|---|
| *Sugar Cane* | Real data composed of NDVI and WRSI values token from 5 sugar cane productive areas of Sao Paulo State, Brazil from 04/01/2001 to 03/31/2008 | Cepagri | 2 | $\cong 500$ |
| *El Niño* | Real data composed of temperature and anomalies for 4 regions in the Pacific Ocean and rainfall of Quaraí, Brazil | CPTEC | 9 | 500 |

### 8.4.1 Experiment 1: *Sugar Cane* dataset

In this experiment, CLEARMiner has mined more $M$ and $V$ patterns than plateaus ($P$), because the time interval employed is monthly. The $M$ patterns detected from NDVI time series are given by their actual amplitude values, regarding the three regions analyzed, Jaboticabal, Jaú and Sertãozinho, as follows:

- Jaboticabal: [0.24; 0.64; 0.30] in [09/2003-10/2004]
- Jaú: [0.29; 0.61; 0.23] in [10/2004-10/2005] and [0.23; 0.61; 0.26] in [10/2005-10/2006]
- Sertãozinho: [0.264471; 0.611832; 0.269969] in [10/2002-09/2003]

The $M$ patterns in NDVI are related to periods when green biomass reaches its high values, before the sugar cane harvest that begins each May. $P$ patterns were found

in WRSI time series. It corresponds to a small variation in WRSI index, such as [0.95; 1.0; 0.99] [10/2001−03/2002] found in the Jaboticabal time series. This phenomenon occurs when the maximum soil water content is reached after a long rainy season.

We have used window of size 2 patterns ($w = 2$) because the values are monthly and the number of patterns found was not very large. The thresholds *min_sup* and *min_conf* were set to 20% and 90% respectively. Thus, the algorithm found the rules as follows:

Table 8.2: Rules generated from NDVI and WRSI time series

| **Examples of Association Rules** | |
|---|---|
| NDVI[V] ⇒ WRSI[V] | WRSI[P] ⇒ NDVI[M] |
| WRSI[V] ⇒ NDVI[V] | WRSI[V] ⇒ NDVI[M] |
| WRSI[M] ⇒ NDVI[M] | WRSI[M] ⇒ NDVI[V] |

The rules from Table 8.2 show that when a negative peak occurs in the WRSI time series, the same pattern occurs in the NDVI time series, as for example the rule:

Example 1:

*Short Format:* WRSI[V] ⇒ NDVI[V]

*Extended Format:* WRSI[0.8; 0.27; 0.87](05/2002 - 09/2002) ⇒
NDVI[0.54; 0.27; 0.63](05/2002 - 02/2003)

However, observing the rules in extended format we can see that pattern V occurs in NDVI time series with a time lag. This time lag calculated by the algorithm is 3 months, considering the inflection point of the two curves (WRSI and NDVI). This information is not evident in short format but it is important to better understand the context in which the phenomenon occurs.

Moreover, when a plateau with maximum values for WRSI occurs, there is the default type for the positive peak NDVI, as for example the rule:

Example 2:

*Short Format:* WRSI[P] ⇒ NDVI[M]

*Extended Format:* WRSI[1.0; 1.0; 0.99](10/2001 - 03/2002) ⇒
NDVI[0.37; 0.55; 0.32](10/2001 - 08/2002)

These rules indicate that there is a dependency (lagged correlation) between NDVI and WRSI. They confirm the expectations of researchers in agrometeorology, because high values of WRSI indicate that there was enough rain to make the soil wet. NDVI measures the green biomass and the index increases if the plant has more green biomass.

### 8.4.2 Experiment 2: *El Niño* dataset

We also have used the CLEARMiner algorithm for mining patterns in heterogeneous time series of meteorological data (rainfall) and anomalies related to the El Niño phenomenon. The El Niño dataset is composed of monthly temperatures and anomalies from four regions in the Pacific Ocean from 1966 to 2008. The warming of the Pacific Ocean can occur in three or four regions and the values of temperature were measured in these regions. Both phenomena El Niño and La Niña influence the climate in South and Southeast of South America (details in Chapter 2).

In this experiment, CLEARMiner has detected $M$ and $V$ patterns. No plateau pattern was found in the El Niño dataset, probably because the data was measured monthly, and the anomaly series has very small variation and amplitude. Here also, we have used a window of size equal to two (patterns), because the values are monthly and the number of patterns found was not very large. Thresholds *min_sup* and *min_conf* were also set to 20% and 90% respectively. Examples of mined rules from El Niño dataset are presented in Table 8.3.

Table 8.3: Association Rules to El Niño dataset

| **Examples of Association Rules** | |
|---|---|
| Rain[V] ⇒ Temp[V] | Rain[M] ⇒ Temp[V] |
| Rain[M] ⇒ Temp[M] | Rain[V] ⇒ Temp[M] |
| Rain[V] ⇒ Anom[M] | Rain[V] ⇒ Anom[V] |
| Rain[M] ⇒ Anom[M] | Rain[M] ⇒ Anom[V] |
| Temp[M] ⇒ Anom[M] | Anom[V] ⇒ Temp[V] |
| Anom[M] ⇒ Temp[V] | Anom[M] ⇒ Temp[M] |

These association patterns indicate that an increase in a series (Rain, Temp, Anom) led to an increase in the other series (Temp, Anom) in previous or subsequent time. It also found that a decrease in the values observed in a series (Rain, Anom) led to a decrease in another series (Temp, Anom). It was also observed that an increase in a series can lead to a decrease in the other series analyzed, or vice versa. CLEARMiner detected several practical rules, exemplified as follows.

Example 1:

*Short Format:* Anom[M] ⇒ Rain[M]

*Extended Format:* Anom[-1.27; -0.55; -0.84] (01/05/1966 - 01/09/1966) ⇒ Rain[43.0; 241.6; 18.8](01/05/1966 - 01/08/1966)

When an increase occurred in anomalies [-1.27, -1.03, -0.84] in the period between (05/01/1966 and 09/01/1966), the rain increased in South region of Brazil [43.0, 241.6, 18.8] in the period between (01/05/1966 and 01/08/1966).

Example 2:

*Short Format:* Rain[M] ⇒ Anom[M]

*Extended Format:* Rain[44.8; 355.2; 70.0](12/01/1982 - 04/01/1983) ⇒ Anom[-1.15; 3.33; 2.13](03/01/1982 - 02/01/1983)

This result is a useful rule that was found by the algorithm: when an increase occurred in anomalies series, rain increased during Spring/Summer in the South Region of Brazil. This association pattern highlights a strong El Niño occurred in 1983 as it can be seen in Figure 2.4 of Chapter 2 - Section 2.2.

### 8.4.3 Performance Evaluation

In this section we show results by comparing the CLEARMiner algorithm with two classical and baseline algorithms, Apriori (Agrawal et al., 1993a) and the Generalized Sequential Pattern (GSP) algorithm (Srikant & Agrawal, 1996). Both algorithms were performed in the Weka platform[1] and CLEARMiner was developed in Java.

As the two algorithms (Apriori and GSP) work only with discrete data, we were only able to compare the rules generation. The datasets used to run Apriori and GSP were quantized by CLEARMiner to avoid distortions that could be caused by different quantization processes.

The Apriori algorithm mined few rules and did not consider time of occurrences. Setting confidence to 0.8, the Apriori algorithm generated only three rules from the dataset with NDVI and WRSI values for Jaú region as follows:

1. WRSI = V ⇒ NDVI = V conf:(1)
2. WRSI = M ⇒ NDVI = M conf:(1)
3. NDVI = V ⇒ WRSI = V conf:(0.86)

The GSP algorithm scans the database several times to generate a set of candidate k-sequences and to calculate their support. We executed the GSP algorithm with $min\_sup = 0.2$, which generated the sequences presented in Table 8.4 for the dataset of NDVI and WRSI values for Jaú region. For *min_sup* values above 0.2, the GSP algorithm in Weka did not work properly.

The sequences mined by GSP are similar to rules generated by CLEARMiner. However, both algorithms (Apriori and GSP) do not keep information about the occurrence

[1] http://www.cs.waikato.ac.nz/ml/weka/

Table 8.4: Sequences generated by GSP

| 1-sequences | 2-sequences |
|---|---|
| {NDVI[V],WRSI[V]} | {NDVI[M],WRSI[P]}{NDVI[V],WRSI[V]} |
| {NDVI[M],WRSI[P]} | {NDVI[M],WRSI[M]}{NDVI[V],WRSI[V]} |
| {NDVI[M],WRSI[M]} | {NDVI[V],WRSI[P]}{NDVI[V],WRSI[V]} |
| {NDVI[V],WRSI[P]} | |

time of the events. CLEARMiner generates rules in an extended format, which can be used to obtain more details about the correlation between time series.

Another advantage of our method is the quantization process that is executed as a first part. This quantization generates a representation that encompasses the semantics meaningful for climate and agroclimate time series. The criteria to quantize time series is based on phenomena that are observed by meteorologists and agrometeorologists and impacts the environment.

## 8.5 Summary

In this chapter, we presented a new unsupervised algorithm to mine association patterns in climate and remote sensing time series, integrated in a remote sensing information system produced to improve the monitoring of sugar cane fields. CLEARMiner presents rules in two formats: short and extended. Short rules are easier to understand, but they are not sufficient to visualize the peak amplitudes and the length of the plateaus. Therefore, the algorithm also presents rules in extended format including details of the values variation and time intervals.

The mined rules for the relevance patterns indicate a relation between series, allowing these patterns (phenomena) happen in different intervals of time. Summarizing, the main contributions of our algorithm are:

1. Include a process of discretization that preserves the semantic meaning of data regarding time;

2. Keep the discretized continuous intervals with their respective times of occurrence to generate the rules;

3. Consider the time lag when it generates rules that associate different time series.

Then, this new method can be used by agrometeorologists to mine and discover knowledge from their long time series of past and forecasting data, being a valuable tool to support their decision making process. In the next chapter, we present the conclusions, main contributions and further work.

# Part III

# Conclusions and Further Work

# Chapter 9

# Conclusions and Further Work

## 9.1 Introduction

After decades of gathering and storing data, many historical series were generated and can be studied in several areas of knowledge, such as economy, health, telecommunication, weather, geoscience and more recently in remote sensing. Time series provide valuable information to comprehend different phenomena along time. Statisticians have contributed to time series analysis proposing innumerous methods for forecasting, monitoring and extreme detections. Recently, time series begins to be studied by researchers in Data Mining proposing several techniques of indexing time series, querying time series, discovering sequential patterns and mining association rules from them.

Analyses of historical series from different countries contribute to understand natural phenomena, which are connected to global warming. Consequently, the measured increase in the average temperature has impelled researches for collaborative development involving meteorologists, mathematicians, statisticians and computer scientists, in order to assess the real impact of such increases as well as on how to deal with it. Understanding what are the main sources for this phenomenon and what are its consequences to the life and to the earth environment is a great challenge for researchers in the whole World. Computer science has an important opportunity to contribute with solutions that use techniques from temporal, spatial and spatio-temporal data mining, for instance.

In this thesis, techniques from data mining were extended and new ones proposed to support analyses of climate and remote sensing time series to collaborate in the understanding of agricultural crops, such as sugar cane, which are used as an important source of renewable energy in Brazil. The experiments accomplished with time series from producing regions of sugar cane in São Paulo state spotted correlations between climate conditions and vegetative indexes obtained from satellite with time lag. Results

presented in this thesis confirm Ganguly's observations (Ganguly & Steinhaeuser, 2008) about relatively simple data mining methods properly employed can result in scientific insight with social impacts.

## 9.2 Main Contributions

In this thesis, we proposed methods based on fractal theory and time series mining. These methods help climatologists and agrometeorologists during the analysis process of the data gathered through ground-based stations and remote sensing images. It improves research in Agriculture, especially in a country of great territorial extensions, such as Brazil.

In order to monitor evolving climate data and to highlight where the specialists should pay more attention during the analysis process, we proposed a fractal-based method associated to a statistical analysis module (Romani et al., 2009a). The method measures the fractal dimension along time spotting trend changes and the attributes responsible for the change behavior.

To aid comparing different sugar cane regions represented by multiple time series we proposed two methods based on weighting the DTW distance function by correlation factors (Romani et al., 2009d, 2010a), where one of them takes advantage of the correlation fractal dimension.

Collaborating in the detection of high and extreme phenomena on climate data without loss of semantical information about time and context of occurrence, we proposed the CLIPSMiner algorithm (Romani et al., 2009c, 2010d). CLIPSMiner is a new unsupervised algorithm to find relevant and extreme patterns in time series, as well as correlations between time series, showing the relationship between the series and when one affects the other.

Finally, we used two approaches to mine rules from time series. The first one is the Apriori-FD method (Romani et al., 2008) that combines techniques of feature selection, discretization and association rules to discover patterns and knowledge from climate data and remote sensing images. As Apriori-FD is not able to deal with the time lag in the generation of rules, we proposed the CLEARMiner algorithm (Romani et al., 2010c). CLEARMiner is a unsupervised algorithm to mine rules, which associates patterns in a time series to patterns in other series considering a time lag. CLEARMiner first converts time series into a symbolic representation and in the second step discovers association patterns between series. The algorithm considers a time-window constraint to reduce the search space and the number of generated rules.

The proposed methods were compared with other similar techniques and assessed by

a group composed of meteorologists, agrometeorologists and remote sensing specialists. All methods reached satisfactory results as declared by the specialists. The development of this thesis also allowed the accomplishment of research in subjects correlated to this thesis:

- creation and organization of the NOAA-AVHRR database, which was employed for the experiments and is being used in the CEPAGRI-UNICAMP;
- collaboration in the development of SatImagExplorer system to automatically extract time series from multiple images from satellites (details in Appendix A);
- development of a new unsupervised discretizer to preprocess climate series through a non-fixed statistical discretization that takes in account the mean and standard deviation values and the Chebyshev's Inequality (Traina et al., 2010);
- designing an environment that allows users to browse a dataset in its tabular format, visualize such data, select query centers, perform similarity queries and have the results of the queries drawn into visualization workspaces that co-exist in the system (Rodrigues Jr. et al., 2010);
- contributing to the proposition of a research project (number 09/53153-3) to Microsoft Research Institute - FAPESP, called "AgroDataMine: Development of Algorithms and Methods of Data Mining to Support Research on Climate Changes Regarding Agrometeorology."

Therefore, we consider that the main contribution of this doctorate program was the proof that well-tailored fractal correlation and data mining techniques can be employed in a satisfactory way to improve the agriculture monitoring, helping agricultural entrepreneurs on decision making.

## 9.3 Publications

Papers published during the doctorate period are also considered as a valuable contribution of this thesis. The list of these publications is presented as follows:

**International Journal**

[1] Gonçalves, R. R. V.; Zullo Jr., J.; Romani, L. A. S.; Nascimento, C. R. and Traina, A. J. M. - "Analysis of NDVI time series using cross-correlation and forecasting methods for monitoring sugar cane fields in Brazil". *International Journal of Remote Sensing*, 20p. (to appear in 2011)

**National Journal**

[1] Chino, D. Y. T.; Romani, L. A. S. and Traina, A. J. M. - "Construindo Séries Temporais de Imagens de Satélite para Sumarização de Dados Climáticos e Monitoramento de Safras Agrícolas". *Revista Eletrônica de Iniciação Científica - REIC* (Online), v. 10, p. 1-16, 2010.

[2] Romani, L. A. S.; Ávila, A. M. H.; Zullo Jr., J.; Traina Jr., C. and Traina, A. J. M. - "Mining Relevant and Extreme Patterns on Climate Time Series with CLIPSMiner". *Journal of Information and Data Management*, v. 1(2), June 2010, p. 245-260, 2010.

**International Book Chapter**

[1] Romani, L. A. S.; Sousa, E. P.; Ribeiro, M. X.; Ávila, A. M. H.; Zullo Jr., J.; Traina Jr., C. and Traina, A. J. M. - "Mining Climate and Remote Sensing Time Series to Improve Monitoring of Sugar Cane Fields". In: Prado, H. A.; Luiz, A. J. B.; Chaib Filho, H. (Org.). *Computational Methods Applied to Agricultural Research: Advances and Applications.* 1 ed. Hershey: IGI Global, p. 1-25, 2010.

**International Conferences - Full papers**

[1] Romani, L. A. S.; Ávila, A. M. H.; Zullo Jr., J.; Chbeir, R.; Traina Jr., C. and Traina, A. J. M. - "CLEARMiner: a new algorithm for mining association patterns on heterogeneous time series from climate data". In: 25th ACM Symposium on Applied Computing (SAC), 2010, Sierre. *Proceedings of the SAC 2010.* New York: ACM Press, v. 1, p. 901-906, 2010.

[2] Romani, L. A. S.; Gonçalves, R. R. V.; Zullo Jr., J.; Traina Jr., C. and Traina, A. J. M. - "New DTW-based Method to Similarity Search in Sugar Cane Regions Represented by Climate and Remote Sensing Time Series". In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010), 2010, Honolulu. *Proceedings of the IGARSS 2010.* Los Alamitos: IEEE Society, v. 1, p. 355-358, 2010.

[3] Rodrigues Jr., J. F.; Romani, L. A. S.; Traina, A. J. M. and Traina Jr., C. - "Combining Visual Analytics and Content Based Data Retrieval Technology for Efficient Data Analysis". In: 14th International Conference Information Visualisation, 2010, London, UK. *Proceedings of the IV 2010*, v. 1, p. 61-67, 2010.

**International Conferences - Extended short papers**

[1] Traina, A. J. M.; Ribeiro, M. X.; Cordeiro, R.; Romani, L. A. S.; Sousa, E. P.; Ávila,

A. M. H.; Zullo Jr., J.; Traina Jr., C. and Rodrigues Jr., J. F. - "How to Find Relevant Patterns in Climate Data: An Efficient and Effective Framework to Mine Climate Time Series and Remote Sensing Images". In: SIAM Annual Meeting, 2010, Pittsburg, USA. *Proceedings of the SIAM-AN 2010*. New York : SIAM, v. 1, p. 124-125, 2010.

**International Workshops - Full papers**

[1] Romani, L. A. S.; Sousa, E. P.; Ribeiro, M. X.; Zullo Jr., J.; Traina Jr., C. and Traina, A. J. M. - "Employing Fractal Dimension to Analyze Climate and Remote Sensing Data Streams". In: SIAM Multimedia Data Mining Workshop 2009 - SDM 2009, 2009, Sparks, Nevada. *Proceedings of the MDM/SDM 2009*, v. 1, p. 1-15, 2009.

[2] Romani, L. A. S.; Zullo Jr., J.; Nascimento, C. R.; Gonçalves, R. R. V.; Traina Jr., C. and Traina, A. J. M. - "Monitoring sugar cane crops through DTW-based method for similarity search in NDVI time series". In: Fifth International Workshop on the Analysis of Multi-temporal Remote Sensing Images, 2009, Groton, Connecticut, USA. *Proceedings of Multitemp 2009*, v. 1, p. 171-178, 2009.

**National Conferences - Full papers**

[1] Chino, D. Y. T.; Romani, L. A. S. and Traina, A. J. M. - "Extração de Séries Temporais de Imagens de Satélite para Monitoramento de Safras Agrícolas e de Dados Climáticos". In: XXIX Concurso de Trabalhos de Iniciação Científica da SBC, 2010, Belo Horizonte. *Anais do CTIC'2010*. Porto Alegre: Sociedade Brasileira de Computação - SBC, v. 1, p. 137-144, 2010. (Work awarded as the second best paper of the event.)

[2] Romani, L. A. S.; Ávila, A. M. H.; Zullo Jr., J.; Traina Jr., C. and Traina, A. J. M. - "Mining Climate and Remote Sensing Time Series to Discover the Most Relevant Climate Patterns". In: XXIV Brazilian Symposium on Databases, 2009, Fortaleza, CE. *Proceedings of the SBBD 2009*. Porto Alegre: Sociedade Brasileira de Computação, v. 1, p. 181-195, 2009.

[3] Romani, L. A. S.; Ávila, A. M. H.; Traina Jr., C. and Traina, A. J. M. - "Detecting Extreme in Climate Time Series using Data Mining Techniques". In: III Simpósio Internacional de Climatologia, 2009, Canela - RS. *Anais do III SIC*. Rio de Janeiro: SBMET, 2009. v. 1. p. 1-6.

[4] Romani, L. A. S.; Traina, A. J. M.; Sousa, E. P.; Zullo Jr., J.; Ávila, A. M. H.; Rodrigues Jr., J. F. and Traina Jr., C. - "Computational framework to analyze

agrometeorological, climate and remote sensing data: challenges and perspectives". In: XXXVI Seminário Integrado de Software e Hardware (in XXIX Congresso da Sociedade Brasileira de Computação), 2009, Bento Gonçalves - RS. *Anais do SEMISH 2009*. Porto Alegre: SBC, p. 323-337, 2009.

[5] Romani, L. A. S.; Sousa, E. P.; Traina Jr., C.; Zullo Jr., J.; Traina, A. J. M. - "Aplicação de Método Baseado em Fractais para Detecção de Correlações entre Imagens AVHRR-NOAA e Dados Climáticos para Regiões Produtoras de Cana-de-açúcar". In: XIV Simpósio Brasileiro de Sensoreamento Remoto (SBSR), 2009, Natal, RN. *Anais do XIV SBSR*. São José dos Campos: Editora do INPE, v. 1, p. 403-410, 2009.

**National Conference - Short paper**

[1] Nunes, S. A.; Romani, L. A. S.; Ávila, A. M. H.; Traina Jr., C.; Sousa, E. P. and Traina, A. J. M. - "Análise baseada em fractais para identificação de mudanças de tendências em múltiplas séries climáticas". In: XXV Brazilian Symposium on Databases, 2010, Belo Horizonte - MG. *Proceedings of the SBBD 2010 - Short Paper Session*. Porto Alegre: SBC, p. 65-72, 2010.

**National Workshops - Full paper**

[1] Romani, L. A. S.; Traina, A. J. M.; Ribeiro, M. X.; Sousa, E. P.; Zullo Jr., J. and Traina Jr., C. - "Aplicação de Técnicas de Mineração em Dados Climáticos e de Satélite para Auxiliar no Acompanhamento das Safras de Cana-de-Áçucar". In: IV Workshop em Algoritmos e Aplicações de Mineração de Dados, 2008, Campinas. *Anais do WAAMD 2008*. Porto Alegre: Sociedade Brasileira de Computação, v. 1, p. 87-92, 2008.

## 9.4 Further Work

Knowledge discovery in agricultural data has a wide variety of subjects still to be explored, especially with regard to the analysis of climate data associated to remote sensing data. Although this thesis has contributed with new methods in this direction, some further works to complement this research can be:

- Development of techniques to find correlations between attributes from datasets characterized by the presence of clusters;
- Definition of methods to associate time series in different time scales with semantic information obtained from experts to improve the analysis process considering

distinct information source;

- Assessment of similarity measures proposed in this work with algorithms of clustering, aimed at extending the applicability of the methods;
- Methods to analyze forecasting data, aimed at discovering possible model errors that can be analyzed and employed in order to tune climate change forecasting models;
- Exploration of broader correlations between data elements that can lead to more refined outliers detection;
- Inclusion of methods to spatial-temporal data mining in order to support spatial analysis;
- Development of classification methods for satellite images using clustering to automatically define classes;
- Development of visual analytics methods to detect missing data, outliers and patterns in heterogeneous time series.

Therefore, we argue that the research started in this thesis can be the basis to a large research field. This can also motivate new researchers into this fascinating and thought-provoking area, integrating computer science and agrometheorology.

# Bibliography


Abdel-Rahman, E. M. and Ahmed, F. B. (2008). The application of remote sensing techniques to sugarcane (saccharum spp. hybrid) production: a review of the literature. *International Journal of Remote Sensing*, 29(13):3753–3767.

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., and Renz, M. (2006). Similarity search on time series based on threshold queries. In *Proceedings of the 10th International Conference on Extending Database Technology (EDBT'2006)*, pp. 276–294, Munich, Germany.

Aggarwal, C. C. (2003). A framework for diagnosing changes in evolving data streams. In *Proceedings of the International Conference on Management of Data (SIGMOD'2003)*, pp. 575–586, San Diego, California, USA. ACM Press.

Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. (2004a). A framework for projected clustering of high dimensional data streams. In *Proceedings of the 30th International Conference on Very Large Databases (VLDB'2004)*, pp. 852–863, Toronto, Canada. Morgan Kaufmann.

Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. (2004b). On demand classification of data streams. In *Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining (KDD'2004)*, pp. 503–508, Seattle, WA, USA. ACM Press.

Aggarwal, C. C. and Yu, P. S. (2008). Locust: An online analytical processing framework for high dimensional classication of data streams. In *Proceedings of the 24th International Conference on Data Engineering (ICDE'2008)*, pp. 426–435, Cancún, México. IEEE Computer Society.

Agrawal, R., Faloutsos, C., and Swami, A. (1993a). Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO'1993)*, pp. 69–84, Chicago, USA. Springer-Verlag.

Agrawal, R., Imielinski, T., and Swami, A. N. (1993b). Mining association rules between sets of items in large databases. In Buneman, P. and Jajodia, S., editors, *Proceedings of the 19th International Conference on Management of Data (SIGMOD'1993)*, v. 1, pp. 207–216, Washington, USA. ACM Press.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, (VLDB'1994)*, pp. 487–499, Santiago de Chile, Chile. Morgan Kaufmann.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering (ICDE'1995)*, pp. 3–14, Taipei, Taiwan. IEEE Computer Society.

Ahola, J. (2001). Mining sequential patterns. Research Report TTE1-2001-10, VTT Information Technology.

Alexander, L., Zhang, X., Peterson, T., Caesar, J., Gleason, B., Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzadech, F., Tagipour, A., Kumar, K. R., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D., Burn, J., Aguilar, E., Brunet, M. Taylor, M., New, M., Zhai, P., Rusticucci, M., and Vasquez-Aguirre, J. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, 111:1–22.

Alfonsi, R. R., Pedro Jr., M. J., Brunini, O., and Barbieri, V. (1987). Condições climáticas para a cana-de-açúcar. In Paranhos, S. B., editor, *Cana-de-açúcar: cultivo e utilização*, v. 1, pp. 42–55. Fundação Cargill, Campinas.

André-Jönsson, H. and Badal, D. Z. (1997). Using signature files for querying time-series data. In *Proceedings of the 1st European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'1997)*, pp. 211–220, Trondheim, Norway.

Antunes, J. F. G. (2005). *Aplicação de lógica fuzzy para estimativa de área plantada da cultura de soja utilizando imagens AVHRR-NOAA*. Master, Unicamp.

Anyamba, A. and Tucker, C. J. (2005). Analysis of sahelian vegetation dynamics using noaa-avhrr ndvi data from 1981-2003. *Journal of Arid Environments*, 63:596–614.

Assad, E. D., Pinto, H. S., Zullo Jr., J., and Marin, F. (2007). Mudanças climáticas e agricultura: uma abordagem climatológica. *Ciência e Ambiente*, 34:169–182.

Assad, E. D. and Sano, E. E. (1998). *Sistema de informações geográficas: aplicações na agricultura*. Embrapa-SPI/Embrapa Cerrados, Brasília, DF, Brasil.

Avila, A. M. H., Gonçalves, R. R. d. V., Pinto, H. S., and Zullo, J. J. (2009). Relação entre a precipitação e o ndvi em imagens avhrr/noaa para a cana-de-açúcar, no estado de são paulo. In *Anais do XIV Simpósio de Sensoriamento Remoto (SBSR'2009)*, pp. 553–560, Natal, RN, Brasil. INPE.

Bajgiran, P. R., Darvishsefat, A. A., Khalili, A., and Makhdoum, M. F. (2008). Using avhrr-based vegetation indices for drought monitoring in the northwest of iran. *Journal of Arid Environments*, 72:1086–1096.

Barbará, D. and Chen, P. (2003). Using self-similarity to cluster large data sets. *Data Mining and Knowledge Discovery*, 7(2):123–152.

Barbará, D., Chen, P., and Nazeri, Z. (2004). Self-similar mining of time association rules. In *Proceedings of the 8th Pacific-Asia Conference Advances in Knowledge Discovery and Data Mining (PAKDD'2004)*, v. 3056, pp. 86–95, Sydney, Australia. Springer.

Berlato, M. A. and Fontana, D. C. (2003). *El Niño e La Niña: impactos no clima, na vegetação e na agricultura do Rio Grande do Sul; aplicações de previsões climáticas na agricultura*. Editora da UFRGS, Porto Alegre, RS, Brasil.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the Knowledge Discovery in Databases - KDD Workshop (KDD'1994)*, pp. 359–370, Seattle, Washington, USA. ACM Press.

Bettini, C., Wang, X. S., and Jajodia, S. (1998). Mining temporal relationships with multiple granularities in time sequences. *Data Engineering Bulletin*, 21:32–38.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the International Conference on Management of Data (SIGMOD'1997)*, pp. 255–264, Tucson, Arizona, USA. ACM Press.

Cai, Y. and Ng, R. T. (2004). Indexing spatio-temporal trajectories with chebyshev polynomials. In *Proceedings of the International Conference on Management of Data (SIGMOD'2004)*, pp. 1–12, Paris, France. ACM Press.

Camara, G. M. S. (1993). Ecofisiologia da cultura de cana-de-açúcar. In Camara, G. M. S., editor, *Produção de cana-de-açúcar*, pp. 31–64. FEALQ, Piracicaba.

Casagrande, A. A. (1991). *Tópico de morfologia e fisiologia de cana-de-açúcar*. Funep, Jaboticabal, SP, Brasil.

Chakrabarti, D. and Faloutsos, C. (2002). F4: large-scale automated forecasting using fractals. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'2002)*, v. 1, pp. 2–9, McLean, VA, EUA. ACM Press.

Chan, K. and Fu, A. W. (1999). Efficient time series matching by wavelets. In *Proceedings of the 15th International Conference on Data Engineering (ICDE'1999)*, pp. 126–133, Sydney, Australia. IEEE Computer Society.

Chen, L. and Ng, R. T. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB'2004)*, pp. 792–803, Toronto, Canada. Morgan Kaufmann.

Chen, L., Özsu, M. T., and Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the International Conference on Management of Data (SIGMOD'2005)*, pp. 491–502, Baltimore, USA. ACM Press.

Chen, Q., Chen, L., Lian, X., Liu, Y., and Yu, J. X. (2007a). Indexable pla for efficient similarity search. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'2007)*, pp. 435–446, Vienna, Austria. Morgan Kaufmann.

Chen, Y., Nascimento, M. A., Ooi, B. C., and Tung, A. K. H. (2007b). Spade: On shape-based pattern detection in streaming time series. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE'2007)*, pp. 786–795, Istanbul, Turkey. IEEE Computer Society.

Chino, D. Y. T., Romani, L. A. S., and Traina, A. J. M. (2010a). Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas. *Revista Eletrônica de Iniciação Científica*, 10:1–16.

Chino, D. Y. T., Romani, L. A. S., and Traina, A. J. M. (2010b). Extração de séries temporais de imagens de satélite para monitoramento de safras agrícolas e de dados climáticos. In *Anais do XXIX Concurso de Trabalhos de Iniciação Científica in XXX Congresso da Sociedade Brasileira de Computação (CSBC'2010)*, v. 1, pp. 137–144, Belo Horizonte, MG, Brasil. SBC.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. In *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB Endowment)*, v. 1, pp. 1542–1552, Auckland, New Zealand. Morgan Kaufmann.

Doorenbos, J. and Kassam, A. H. (1979). *Yield response to water*. FAO: Irrigation and drainage, n.33. FAO, Rome, Italy.

Doorenbos, J. and Kassam, A. H. (1994). *Efeito da água no rendimento das culturas*. Universidade Federal da Paraíba, Campina Grande, PB, Brasil.

Dufek, A. S. and Ambrizzi, T. (2008). Precipitation variability in são paulo state, brazil. *Theoretical and Applied Climatology*, 93:167–178.

Emery, W., Baldwin, D. G., and Matthews, D. (2003). Maximum cross correlation automatic satellite image navigation and attitude corrections for open ocean image navigation. *IEEE Transactions on Geoscience and Remote Sensing*, 41:33–42.

Emery, W. J., Brown, J., and Novak, Z. P. (1989). Avhrr image navigation: summary and review. *Photogrammetric Engineering and Remote Sensing of Environment*, 55:1175–1183.

Esquerdo, J. C. D. M., Antunes, J. F. G., Baldwin, D. G., Emery, W. J., and Jr, J. Z. (2006). An automatic system for avhrr land surface product generation. *International Journal of Remote Sensing*, 27:3925–3942.

Everinghama, Y. L., Smyth, C. W., and Inman-Bamber, N. G. (2009). Ensemble data mining approaches to forecast regional sugarcane crop production. *Agricultural and Forest Meteorology*, 149:689–696.

Faloutsos, C. and Kamel, I. (1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *Proceedings of the 13th Symposium on Principles of Database Systems (PODS'1994)*, pp. 4–13, Minneapolis, Minnesota, USA. ACM Press.

Faloutsos, C., Ranganathan, M., and Manolopulos, Y. (1994). Fast subsequence matching in time-series databases. In *Proceedings of the International Conference on Management of Data (SIGMOD'1994)*, v. 23, pp. 419–429, Minneapolis, Minnesota, USA. ACM Press.

Fauconier, R. and Bassereau, A. H. (1975). *La canã de azucar*. Blume, Barcelona, Spain.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, USA.

Ferraro, D. O., Rivero, D. E., and Ghersa, C. M. (2009). An analysis of the factors that influence sugarcane yield in northern argentina using classification and regression trees. *Field Crops Research*, 112:149–157.

Ferrer-Troyano, F., Aguilar-Ruiz, J. S., Riquelme, J. C., and Jose, C. (2006). Data streams classication by incremental rule learning with parameterized generalization. In *Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC'2006)*, pp. 657–661, Dijon, France. ACM Press.

Fontana, D. C., Potgieter, A., and Apan, A. (2005). Relação entre a precipitação pluvial e índice de vegetação em imagens multitemporais modis. In de Agrometeorologia (SBAgro), S. B., editor, *Anais do XVI Congresso Brasileiro de Agrometeorologia (CBAGRO'2005)*, Campinas, SP, Brasil.

Fortes, C. and Demattê, J. A. M. (2006). Discrimination of sugarcane varieties using landsat 7 etm+ spectral data. *International Journal of Remote Sensing*, 27(7):1395–1412.

Frentzos, E., Gratsias, K., and Theodoridis, Y. (2007). Index-based most similar trajectory search. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE'2007)*, pp. 816–825, Istanbul, Turkey. IEEE Computer Society.

Gama, J., Medas, P., and Rodrigues, P. (2005). Learning decision trees from dynamic data streams. In *Proceedings of the 20th Annual ACM Symposium on Applied Computing (SAC'2005)*, pp. 573–577, Santa Fe, New Mexico. ACM Press.

Ganguly, A. R. and Steinhaeuser, K. (2008). Data mining for climate change and impacts. In *Proceedings of the 8th International Conference on Data Mining Workshops (ICDM'2008)*, pp. 385–394, Pisa, Italy. IEEE Computer Society.

Garcia, E. and Vieira, M. T. P. (2008). Estudo de caso de mineração de dados multirelacional: Aplicação do algoritmo connetionblock em um problema da agroindústria. In *Proceedings of the 23rd Brazilian Symposium on Databases (SBBD'2008)*, v. 1, pp. 224–237, Campinas, SP, Brazil. SBC.

Garofalakis, M. N., Rastogi, R., and Shim, K. (1999). Spirit: Sequential pattern mining with regular expression constraints. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB'1999)*, pp. 223–234, Edinburgh, Scotland. Morgan Kaufmann.

Garofalakis, M. N., Rastogi, R., and Shim, K. (2002). Mining sequential patterns with regular expression constraints. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):530–552.

Geurts, P. (2001). Pattern extraction for time series classification. In Raedt, L. D. and Siebes, A., editors, *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'2001)*, pp. 115–127, Freiburg, Germany. Springer.

Godoy, O. P. and Toledo, F. F. (1972). *Plantas extrativas: cana-de-açúcar, amendoim, girassol, mamona, soja*. ESALQ, Piracicaba, SP, Brasil.

Goldemberg, J. (2007). Ethanol for a sustainable energy future. *Science*, 315:808–810.

Goldemberg, J., Coelho, S. T., and Guardabassi, P. (2008). The sustainability of ethanol production from sugarcane. *Energy Policy*, 36:2086–2097.

Gonçalves, R. R. V. (2008). *Relação entre a resposta espectral da cana-de-açúcar, registrada nas imagens dos satélites AVHRR/NOAA, em São Paulo, e dados agroclimáticos, no período de 2001 a 2008*. Mestrado, Unicamp.

Gonçalves, R. R. V., Nascimento, C. R., Zullo Jr., J., and Romani, L. A. S. (2009). Relationship between the spectral response of sugar cane, based on avhrr/noaa satellite images, and the climate condition, in the state of são paulo (brazil), from 2001 to 2008. In Civco, D. L., editor, *Proceedings of the 5th International Workshop on the Analysis of Multi-temporal Remote Sensing images (Multitemp'2009)*, pp. 315–322, Groton, Connecticut, USA.

Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Addison Wesley, USA.

Goswami, B., Venugopal, V., Sengupta, D., Madhusoodanan, M., and Xavier, P. (2006). Increasing trend of extreme rain events over india in a warming environment. *Science*, 314(5804):1442–1445.

Grimm, A. and Tedeschi, R. G. (2008). Enso and extreme rainfall events in south america. *Journal of Climate*, 22:1589–1609.

Groisman, P., Knight, R., Easterling, D., Karl, T., Hegerl, G., and Razuvaev, V. N. (2005). Trends in intense precipitation in the climate record. *Journal of Climate*, 18:1326–1350.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528.

Guyot, G. (1990). Optical properties of vegetation canopies. In Steven, M. D. and Clark, J. A., editors, *Application of Remote Sensing in Agriculture*, pp. 19–44. Butterworths, London, UK.

Hajj, M. E., Bégué, A., Guillaume, S., and Martiné, J.-F. (2009). Integrating spot-5 time series, crop growth modeling and expert knowledge for monitoring agricultural practices - the case of sugarcane harvest on reunion island. *Remote Sensing of Environment*, 113:2052–2061.

Han, J. and Kamber, M. (2001). *Data Mining - Concepts and Techniques*. Morgan Kaufmann Publishers, New York, NY, USA, 1st edition edition.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the International Conference on Management of Data (SIGMOD'2000)*, pp. 1–12, Dallas, USA. ACM Press.

Holben, B. N. (1986). Characteristics of maximum value composite images from temporal avhrr data. *International Journal of Remote Sensing*, 7:1417–1435.

Holben, B. N., Tucker, C. J., and Cheng-Jeng, F. (1980). Spectral assessment of soyabean leaf area and leaf biomass. *Photogrammetric Engineering and Remote Sensing of Environment*, 46:651–656.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91.

Houtsma, M. and Swami, A. N. (1993). Set-oriented mining of association rules. Research Report RJ 9567, IBM Almaden Research Center.

Huete, A. R., Liu, H. Q., Batchily, K., and Leeuwen, W. V. (1997). A comparison of vegetation indices over a global set of tm images for eos-modis. *Remote Sensing of Environment*, 59:440–451.

IPCC, w. g. (2007). Climate change 2007: Fourth assessment report (ar4). Technical report, Intergovernmental Panel on Climate Change.

Jensen, M. E. (1968). Water consumption by agricultural plants. In Koslowsky, T. T., editor, *Water deficits and Plant Growth*, v. 2. Academic Press, New York.

Jin, C., Qian, W., Sha, C., Yu, J. X., and Zhou, A. (2003). Dynamically maintaining frequent items over a data stream. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'2003)*, pp. 287–294, New Orleans, Louisiana, USA. ACM Press.

Kampel, M. (2004). Características gerais dos satétiles noaa: histórico, instrumentos e comunicação de dados. In Ferreira, N. J., editor, *Aplicações ambientais brasileiras dos satélites NOAA e TIROS-N*. Oficina de textos, São Paulo.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons.

Keogh, E. (2001). A tutorial on indexing and mining time series data. In IEEE, editor, *Proceedings of the International Conference on Data Mining (ICDM'2001)*, pp. 1–2, San Jose, California, USA. IEEE Computer Society.

Keogh, E., Chakrabarti, K., Pazzani, M. J., and Mehrotra, S. (2001a). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286.

Keogh, E. J., Chakrabarti, K., Mehrotra, S., and Pazzani, M. J. (2001b). Locally adaptive dimensionality reduction for indexing large time series databases. In Aref, W. G., editor, *Proceedings of the International Conference on Management of Data (SIGMOD'2001)*, pp. 151–162, Santa Barbara, USA. ACM Press.

Keogh, E. J. and Pazzani, M. J. (2001). Derivative dynamic time warping. In *Proceedings of the 1st SIAM International Conference on Data Mining (SDM'2001)*, pp. 1–11, Chicago, IL, USA. SIAM.

Keogh, E. J. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge Information System*, 7(3):358–386.

Kerber, R. (1992). Chimerge: Discretization of numeric attributes. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI'1992)*, pp. 123–128, San Jose, California, USA.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB'2004)*, pp. 180–191, Toronto, Canada. Morgan Kaufmann.

Kim, S.-W., Park, S., and Chu, W. W. (2001). An index-based approach for similarity search supporting time warping in large sequence databases. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'2001)*, pp. 607–614, Heidelberg, Germany. IEEE Computer Society.

Kleinberg, J. M. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Korn, F., Jagadish, H., and Faloutsos, C. (1997). Efficiently supporting ad hoc queries in large datasets of time sequences. In *Proceedings of the International Conference on Management of Data (SIGMOD'1997)*, pp. 289–300, Tucson, USA. ACM Press.

Kurgan, L. A. and Cios, K. J. (2004). Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(2):145–153.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the Workshop on Research Issues in Data Mining and Knowledge Discovery in ACM SIGMOD'2003*, pp. 2–11, San Diego, California, USA. ACM Press.

Lin, J., Keogh, J., Wei, L., and Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2).

Liu, H. and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the 7th International Conference on Tools with Artificial Intelligence (ICTAI'1995)*, pp. 388–391, Washington, DA, USA. IEEE Computer Society.

Liu, W. T. and Kogan, F. (2002). Monitoring brazilian soybean production using noaa/avhrr based vegetation condition indices. *International Journal of Remote Sensing*, 23:1161–1179.

Lucas, A. A. and Schuler, C. A. B. (2007). Análise do ndvi/noaa em cana-de-açúcar e mata atlântica no litoral norte de pernambuco, brasil. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 11:607–614.

Luo, C., Zhao, Y., Cao, L., Ou, Y., and Zhang, C. (2008). Exception mining on multiple time series in stock market. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WIC'2008)*, v. 3, pp. 690–693, Sydney, Australia. IEEE Computer Society.

Macedo, I. C., Seabra, J. E., and Silva, J. E. (2008). Green house gases emissions in the production and use of ethanol from sugarcane in brazil: The 2005/2006 averages and a prediction for 2020. *Biomss and Bioenergy*, 32:582–595.

Maignan, F., Bréon, F. M., Bacour, C., Demarty, J., and Poirson, A. (2008). Interannual vegetation phenology estimates from global avhrr measurements: Comparison with in situ data and applications. *Remote Sensing of Environment*, 112:496–505.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature: updated and augmented.* W. H. Freeman and Company, New York, NY, USA.

Manjhi, A., Shkapenyuk, V., Dhamdhere, K., and Olston, C. (2005). Finding (recently) frequent items in distributed data streams. In *Proceedings of the 21st International Conference on Data Engineering (ICDE'2005)*, pp. 767–778, Tokyo, Japan. IEEE Computer Society.

Mannila, H., Toivonem, H., and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Journal of Data Mining and Knowledge Discovery*, 1:259–289.

Marengo, J. A., Nobre, C. A., Salati, E., and Ambrizzi., T. (2007). Caracterização do clima atual e definição das alterações climáticas para o território brasileiro ao longo do século xxi: Sumário técnico. Technical report, MMA, SBF, DCBio.

Martinelli, L. A. and Filoso, S. (2008). Expansion of sugarcane ethanol production in brazil: environmental and social challenges. *Ecological Applications*, 18(4):885–898.

McNicholas, P. D., Murphy, T. B., and O'Regan, M. (2008). Standardizing the lift of an association rule. *Computational Statistics & Data Analysis*, 52(10):4712–4721.

Meehl, G. and Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, 305(5686):994–997.

Moreira, M. A. and Shimabukuro, Y. E. (2004). Cálculo do índice de vegetação a partir do sensor avhrr. In Ferreira, N. J., editor, *Aplicações ambientais brasileiras dos satélites NOAA e TIROS-N* 271, p. Oficina de textos, São Paulo.

Morettin, P. A. and Toloi, C. M. C. (2006). *Análise de séries temporais.* Editora Edgard Blücher Ltda., São Paulo, SP, Brasil, 2ª edição edition.

Morinaka, Y., Yoshikawa, M., Amagasa, T., and Uemura, S. (2001). The l-index: An indexing structure for efficient subsequence matching in time sequence databases. In *Proceedings of the Workshop on Mining Spatial and Temporal data (5th Pacific-Asia Conference on Knowledge Discovery and Data Mining)*, pp. 51–60, Hong Kong, China.

Morse, M. D. and Patel, J. M. (2007). An efficient and accurate method for evaluating time series similarity. In *Proceedings of the International Conference on Management of Data (SIGMOD'2007)*, pp. 569–580, Beijing, China.

Nascimento, C. R., Zullo Jr., J., Romani, L. A. S., and Rodrigues, L. H. A. (2009). Identification of sugar cane fields in the state of sao paulo using a time series of avhrr/noaa satellite images. In *Proceedings of the The 5th International Workshop on the Analysis of Multi-temporal Remote Sensing images (Multitemp'2009)*, pp. 104–111, Groton, Connecticut, USA.

Nass, L. L., Pereira, P. A. A., and Ellis, D. (2007). Biofuels in brazil: An overview. *Crop Science*, 47(Nov-Dec):2228–2237.

Novo, E. M. L. M. (1992). *Sensoriamento remoto: princípios e aplicações*. Edgard Blucher, São Paulo, SP, Brasil, 2 edition.

Nunes, S. A., Romani, L. A. S., Ávila, A. M. H., Traina Jr., C., Sousa, E. P., and Traina, A. J. M. (2010). Análise baseada em fractais para identificação de mudanças de tendências em múltiplas séries climáticas. In *Anais do XXV Simpósio Brasileiro de Banco de Dados (SBBD'2010)*, pp. 1–8, Belo Horizonte, MG, Brasil. SBC.

Oates, T., Schmill, M. D., Jensen, D., and Cohen, P. R. (1997). A family of algorithms for finding temporal structure in data. In *Proceedings of the 6th International Workshop on AI and Statistics*, pp. 371–378, Fort Lauderdale, Florida, USA.

Ometto, J. C. (1988). *Freqüência de irrigação em cana-de-açúcar*. FEALQ, Piracicaba, SP, Brasil.

Omiecinski, E. R. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69.

Papadimitriou, S., Brockwell, A., and Faloutsos, C. (2004). Adaptive, unsupervised stream mining. *VLDB Journal*, 13(3):222–239.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii regression, heredity and panmixia. *Philosophical Transactions of the Royal Society*, 187:253–318.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'2001)*, pp. 215–224, Heidelberg, Germany. IEEE Computer Society.

Pellegrino, G. Q. (2001). *Utilização de Dados Espectrais do Satélite NOAA14/AVHRR como Fonte de Dados para Modelos Matemáticos de Estimativa da Fitomassa da Cana-de-Açúcar*. Phd, Universidade Estadual de Campinas.

Pereira, A. R., Angelocci, L. R., and Sentelhas, P. C. (2002). *Agrometeorologia: fundamentos e aplicações práticas*. Livraria e Editora Agropecuária, Guaíba, RS, Brasil.

Pinto, H. S. and Assad, E. D. (2008). Aquecimento global e cenários futuros da agricultura brasileira. Technical report, British Embassy – Brasil.

Ponzoni, F. (2001). Comportamento espectral da vegetação. In Meneses, P. R. and Netto, J. S. M., editors, *Sensoriamento remoto: reflectância de alvos naturais*, v. 1, pp. 157–199. UnB; Embrapa Cerrados, Brasilia; Planaltina.

Popivanov, I. and Miller, R. J. (2002). Similarity search over time series data using wavelets. In *Proceedings of the 18th International Conference on Data Engineering (ICDE'2002)*, pp. 212–221, San Jose, USA. IEEE Computer Society.

Ribeiro, M. X., Traina, A. J. M., and Traina Jr., C. (2008). A new algorithm for data discretization and feature selection. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC'2008)*, pp. 953–954, Fortaleza, CE, Brazil. ACM Press.

Rodrigues, P., Gama, J., and Pedroso, J. P. (2008). Hierarchical clustering of time-series data streams. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):615–627.

Rodrigues Jr., J. F., Romani, L. A. S., Traina, A. J. M., and Traina Jr., C. (2010). Combining visual analytics and content based data retrieval technology for efficient data analysis. In *Proceedings of the 14th International Conference Information Visualisation (IV'2010)*, v. 1, pp. 61–67, London, UK.

Romani, L. A. S., Gonçalves, R. R. V., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2010a). New dtw-based method to similarity search in sugar cane regions represented by climate and remote sensing time series. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'2010)*, v. 1, pp. 1–4, Honolulu, Hawaii, USA. IEEE Geoscience and Remote Sensing Society.

Romani, L. A. S., Sousa, E. P., Ribeiro, M. X., Ávila, A. M. H., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2010b). Mining climate and remote sensing time series to improve monitoring of sugar cane fields. In Prado, H. A., Luiz, A. J. B., and Chaib Filho, H., editors, *Computational Methods Applied to Agricultural Research: Advances and Applications*, pp. 1–25. IGI Global, Hershey, 1ª edition.

Romani, L. A. S., Sousa, E. P., Ribeiro, M. X., Zullo Jr., J., and Traina Jr., C. (2009a). Employing fractal dimension to analyze climate and remote sensing data streams. In *Proceedings of the SIAM Multimedia Data Mining Workshop (MDM/SDM'2009)*, pp. 1–15, Sparks, Nevada, USA. SIAM.

Romani, L. A. S., Traina, A. J. M., Ribeiro, M. X., Sousa, E. P., Zullo Jr., J., and Traina Jr., C. (2008). Aplicação de técnicas de mineração em dados climáticos e de satélite para auxiliar no acompanhamento das safras de cana-de-Açucar. In *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD'2008)*, pp. 87–92, Campinas, SP, Brasil. SBC.

Romani, L. A. S., Ávila, A. M. H., Traina Jr., C., and Traina, A. J. M. (2009b). Detecting extreme in climate time series using data mining techniques. In *Anais do III Simpósio Internacional de Climatologia (SIC'2009)*, v. 1, pp. 1–6, Canela, RS, Brasil. SBMET.

Romani, L. A. S., Ávila, A. M. H., Zullo Jr., J., Chbeir, R., Traina Jr., C., and Traina, A. J. M. (2010c). Clearminer: a new algorithm for mining association patterns on heterogeneous time series from climate data. In *Proceedings of the 25th ACM Symposium on Applied Computing (SAC'2010)*, v. 1, pp. 901–906, Sierre, Switzerland. ACM Press.

Romani, L. A. S., Ávila, A. M. H., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2009c). Mining climate and remote sensing time series to discover the most relevant climate patterns. In *Anais do XXIV Simpósio Brasileiro de Banco de Dados (SBBD'2009)*, v. 1, pp. 181–195, Fortaleza, CE, Brasil. SBC.

Romani, L. A. S., Ávila, A. M. H., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2010d). Mining relevant and extreme patterns on climate time series with clipsminer. *Journal of Information and Data Management (JIDM)*, 1(2).

Romani, L. A. S., Zullo Jr., J., Nascimento, C. R., Gonçalves, R. R. V., Traina Jr., C., and Traina, A. J. M. (2009d). Monitoring sugar cane crops through dtw-based method for similarity search in ndvi time series. In *Proceedings of the 5th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multitemp'2009)*, pp. 171–178, Groton, Connecticut, USA.

Rosborough, G. W. and Baldwin, D. G. E. W. J. (1994). Precise avhrr image navigation. *IEEE Transactions on Geoscience and Remote Sensing*, 32:644–657.

Rosseti, L. (2001). Zoneamento agrícola em aplicações de crédito e securidade rural no brasil. *Revista Brasileira de Agrometeorologia*, 9(3):386–399.

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. (1973). Monitoring vegetation systems in the great plains with erts. In NASA, editor, *Proceedings of the Earth Resources Technology Satellite*, pp. 309–317, Washington, DC, USA. NASA.

Rudorff, B. F. T., Adami, M., Aguiar, D. A. d., Gusso, A., Silva, W. F. d., and Freitas, R. M. d. (2009). Temporal series of evi/modis to identify land converted to sugarcane. In *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'2009)*, v. IV, pp. 252–255, Cape Town, South Africa. IEEE Geoscience and Remote Sensing Society.

Rudorff, B. F. T., Aguiar, D. A. d., Silva, W. F. d., Sugawara, L. M., Adami, M., and Moreira, M. A. (2010). Studies on the rapid expansion of sugarcane for ethanol production in são paulo state (brazil) using landsat data. *Remote sensing*, 2:1057–1076.

Rudorff, B. F. T. and Batista, G. T. (1990). Yield estimation of sugarcane based on agrometeorological-spectral models. *Remote Sensing of Environment*, 33:183–192.

Sakurai, Y., Faloutsos, C., and Yamamuro, M. (2007). Stream monitoring under the time warping distance,. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE'2007)*, pp. 1046–1055, Istanbul, Turkey. IEEE Computer Society.

Savarese, A., Omiecinski, E., and Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Data Bases (VLDB'1995)*, pp. 432–444, Zurich, Switzerland. Morgan Kaufmann.

Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W. H. Freeman and Company, New York, NY, USA.

Shatkay, H. and Zdonik, S. B. (1996). Approximate queries and representations for large data sequences. In Su, S. Y. W., editor, *Proceedings of the 12th International Conference on Data Engineering (ICDE'1996)*, pp. 536–545, New Orleans, Louisiana, USA. IEEE Press.

Simoff, S. J., Böhlen, M. H., and Mazeika, A. (2008). Visual data mining: An introduction and overview. *Visual Data Mining LNCS*, 4404:1–12.

Song, J. and Gao, W. (1999). An improved method to derive surface albedo from narrowband avhrr satellite data: narrowband to broadband conversion. *Journal of Applied Meteorology*, 38:239–249.

Sousa, E. P. M. (2006). *Identificação de correlações usando a Teoria dos Fractais*. Phd thesis, USP.

Sousa, E. P. M., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. (2007a). Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing Journal - Special Issue on Knowledge Discovery from Data Streams*, 25:33–59.

Sousa, E. P. M. d., Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2007b). A fast and effective method to find correlations among attributes in databases. *Data Mining and Knowledge Discovery*, 14(3):367 – 407.

Srikant, R. and Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology (EDBT'1996)*, v. 1057, pp. 3–17, Avignon, France.

Thornthwaite, C. W. and Mather, J. R. (1955). The water balance. *Climatology*, 8(1):104.

Townshend, J. R. G. (1994). Global data sets for land applications from the advanced very high resolution radiometer: an introduction. *International Journal of Remote Sensing*, 15:3319–3332.

Traina, A. J. M., Ribeiro, M. X., Cordeiro, R., Romani, L. A. S., Sousa, E. P., Ávila, A. M. H., Zullo Jr., J., Traina Jr., C., and Rodrigues Jr., J. F. (2010). How to find relevant patterns in climate data: An efficient and effective framework to mine climate time series and remote sensing images. In *Proceedings of the SIAM Annual Meeting (SIAM-AN'2010)*, v. 1, pp. 124–125, Pittsburg, USA.

Traina, A. J. M., Traina Jr., C., Papadimitriou, S., and Faloutsos, C. (2001). Tri-plots: Scalable tools for multidimensional data mining. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'2001)*, pp. 184–193, San Francisco, CA, USA. ACM Press.

Traina Jr., C., Sousa, E. P. M., and Traina, A. J. M. (2005). Using fractals in data mining. In Kantardzic, M. M. and Zurada, J., editors, *New Generation of Data Mining Applications*. Wiley/IEEE Press.

Traina Jr., C., Traina, A. J. M., and Faloutsos, C. (2010). Fast feature selection using fractal dimension - ten years later. *Journal of Information and Data Management*, 1(1):17–20.

Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2000). Fast feature selection using fractal dimension. In *Proceedings of the XV Brazilian Symposium on Databases (SBBD'2000)*, pp. 158–171, João Pessoa, PB, Brazil. SBC.

Velasco, I. and Fritsch, J. (1987). Mesoscale convective complexes in the americas. *Journal of Geophysical Research*, 92:9591–9613.

Vianello, R. L. and Alves, A. R. (1991). *Meteorologia básica e aplicações*, v. 1. Universidade Federal de Viçosa - imprensa universitária, Viçosa, Minas Gerais, Brasil.

Vincent, L., Peterson, T., Barros, V., Marino, M., Rusticucci, M., G., C., Ramirez, E., Alves, L., Ambrizzi, T., Berlato, M., Grimm, A., Marengo, J., Molion, L., Moncunill, D., Rebello, E., Anunciação, Y., Quintana, J., Santos, J., Baez, J., Coronel, G., Garcia, J., Trebejo, I., Bidegain, M., Haylock, M., and Karoly, D. (2005). Observed trends in indices of daily temperature extremes in south america 1960-2000. *Journal of Climate*, 18:5011–5023.

Vlachos, M., Gunopulos, D., and Kollios, G. (2002). Discovering similar multidimensional trajectories. In *Proceedings of the 18th International Conference on Data Engineering (ICDE'2002)*, pp. 673–684, San Jose, CA, USA. IEEE Computer Society.

Wang, J., Price, K. P., and Rich, P. M. (2003). Temporal responses of ndvi to precipitation and temperature in the central great plains. *International Journal of Remote Sensing*, 24:2345–2364.

Wang, Q., Adiku, S., Tenhunen, J., and Granier, A. (2005). On the relationship of ndvi with leaf area index in a deciduous forest site. *Remote Sensing of Environment*, 94:244–255.

Wang, Q., Megalooikonomou, V., and Faloutsos, C. (2010). Time series analysis with multiple resolutions. *Information Systems*, 35:56–74.

Wei, W. W. S. (2006). *Time Series Analysis: univariate and multivariate methods*. Pearson Addison Wesley, USA, second edition edition.

Winograd, S. (1976). On computing the discrete fourier transform. *Mathematics*, 73(4):1005–1006.

Xavier, A. C., Rudorff, B. F. T., Shimabukuro, Y. E., Berka, L. M. S., and Moreira, M. A. (2006). Multi-temporal analysis of modis data to classify sugarcane crop. *International Journal of Remote Sensing*, 27:755–768.

Xavier, A. C. and Vettorazzi, C. A. (2004). Mapping leaf area index through spectral vegetation indices in a subtropical watershed. *International Journal of Remote Sensing*, 25:1661–1672.

Yamamoto, C. H., Oliveira, M. C. F., Rezende, S. O., and Nomelini, J. (2008). Including the user in the knowledge discovery loop: Interactive itemset-driven rule extraction. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC'2008)*, v. 2, pp. 1212–1217, Fortaleza, CE, Brazil. ACM Press.

Yi, B. and Faloutsos, C. (2000). Fast time sequence indexing for arbitrary lp norms. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB'2000)*, pp. 385–394, Cairo, Egypt. Morgan Kaufmann.

Yi, B.-K., Jagadish, H. V., and Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. In *Proceedings of the 14th International Conference on Data Engineering (ICDE'1998)*, pp. 201–208, Orlando, Florida, USA. IEEE Computer Society.

Yoo, J. S. and Shekhar, S. (2009). Similarity-profiled temporal association mining. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1147–1161.

Zaki, M. J. (2000). Sequences mining in categorical domains: Incorporating constraints. In *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM'2000)*, pp. 422–429, McLean, VA, USA. ACM Press.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42:31–60.

Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W. (1997). New algorithms for fast discovery of association rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'1997)*, pp. 283–286, Newport Beach, California, USA. ACM Press.

Zhai, P., Baethgen, W. E., Cerda, M. S., Davey, M., Goolaup, P., Kontongomde, H., Kousky, V. E., Llansó, P., Ropelewski, C. F., and Reid, P. (2005). Guidelines on climate watches. Technical report, World Meteorological Organization. GuidelinesonClimateWatches.pdf.

Zhang, C., Yang, Q., and Liu, B. (2005). Guest editors' introduction: Special section on intelligent data preparation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(9):1163–1165.

Zhu, Y. and Shasha, D. (2003). Efficient elastic burst detection in data streams. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD'2003)*, pp. 336–345, Washington, DC, USA. ACM Press.

# Part IV

# Appendix

# Appendix A

# The *SatImagExplorer* System

## A.1 Introduction

Nowadays, data recorded by satellites are more accessible and there are appropriate technology (software and hardware) to receive, distribute, manipulate and process long time series of images. However, the majority of this technology is for commercial use and is not suitable for processing long series of images. Examples of software developed to manipulate remote sensing images are Erdas[1], Idrisi[2] and Spring[3].

The development of the NavPro system (described in Chapter 2) allowed to enlarge the scope of research that involves NOAA/AVHRR multi-temporal images. However, the process of extraction, storage and visualization of time series from these images is very costly to be done manually, as previously mentioned. Thus, it is crucial the development of a new technique to perform these tasks automatically in order to allow the accomplishment of new research activities by the remote sensing specialists.

In general, time series are built by extracting measurements or performing calculations on data from regions of interest (ROIs) from the images that were gathered sequentially along a period of time. The possibility of automatically extracting time series was much desired by experts, but was not available on commercial systems, as the ENVI[4] (The Environment for Visualizing Images) system, whose module IDL (Interactive Data Language) allows to build time series from images. However, IDL is a script language not trivial to use. Remote sensing specialists have two alternatives to deal with the problem of extracting time series from multi-temporal images: learn how to develop scripts in IDL or request the support of an expert in IDL language. In fact, this task is very difficult to be executed without an automatic process.

As a contribution of this work, it was proposed the development of the *SatImagExplorer* system that handles sequences of images automatically, extrapolating the values extraction from each pixel to all images of the same sequence. Consequently, this system makes the task of obtaining time series of a huge number of multi-temporal images more agile and efficient. *SatImagExplorer* has been implemented by an undergraduate student[5] as part of his initiation in research program in the scope of this doctorate project.

[1] http://www.erdas.com/
[2] http://www.clarklabs.org/
[3] http://www.dpi.inpe.br/spring/portugues/index.html
[4] http://www.ittvis.com/ProductServices/ENVI.aspx
[5] Daniel Y. T. Chino, that is supported by AgroDataMine project - number 09/53153-3

The *SatImagExplorer* system was developed using open source software what allows broader contributions to its improvement in the future. The system was coded in C++ language with the open source library GDAL[6] (Geospatial Data Abstraction Library) for the manipulation of images in GeoTIFF[7] format. The GeoTIFF format is an extension of TIFF images (Tagged Image File Format) for geospatial imagery. Images in GeoTIFF format have information about the image itself, such as dimension, color depth, etc., geographic/cartographic information and values measured by sensors on board satellites.

## A.2 Description of the *SatImagExplorer* system

The *SatImagExplorer* user interface is simple and intuitive. Through the mouse interaction, the user defines any polygonal region, which will be used for the analysis of the image series. The system architecture is modular, as it can be seen in Figure A.1.

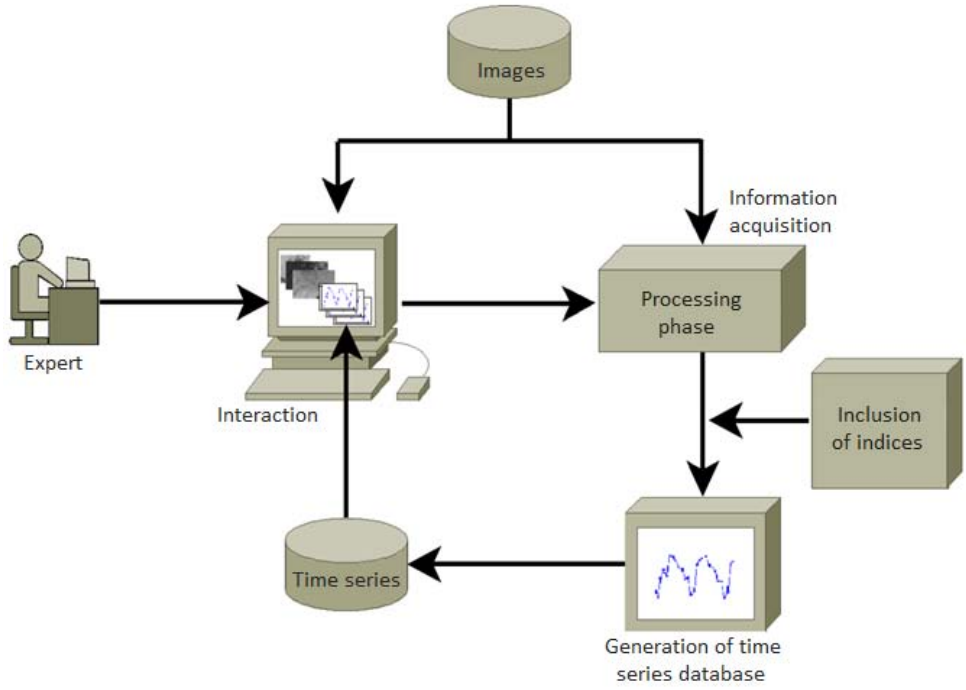



Figure A.1: Architecture schema of the *SatImagExplorer* system (adapted from (Chino et al., 2010a)).

### A.2.1 Interaction Module

The *SatImagExplorer* interface is composed of windows, icons, menus, and a pointing device which allows direct manipulation. The user opens a list of image files from which are extracted values of the same pixel, or region, of all images generating time series. Dragging the mouse, the user indicates a polygonal region which will be used to extract time series per pixel and receives visual feedback as an indication of the marked region.

Alternatively, the system has option for determining the region of interest through a list of coordinates (latitude and longitude) in text files. Thus, it can be pointed out

[6]http://gdal.org/
[7]http://trac.osgeo.org/geotiff

precisely the regions of interest that will be analyzed, for example the exact locations where there is sugar cane crop fields. That is, if specialists want to track the performance of a specific planting from a farm or business, they can use a coordinates file, which provides exactly the region of interest.

### A.2.2 Processing Module

*SatImagExplorer* extracts values associated to pixels in all images opened in the system, generating a time series for each pixel for the region of interest defined by user. Time series are stored in a database allowing users to process further analyses.

Thus, the user, or the specialist in the field, can manipulate images and time series extracted from them in order to acquire knowledge that assists him/her in the analysis task. The *SatImagExplorer* system allows simultaneously opening several satellite images in GeoTIFF format. In general, these images have already been submitted to the georeferencing process. *SatImagExplorer* opens several format of images, such as: raw, NDVI, surface temperature and GOES satellite, which are low and medium resolution images. Figure A.2 shows a screenshot of the *SatImagExplorer* system.

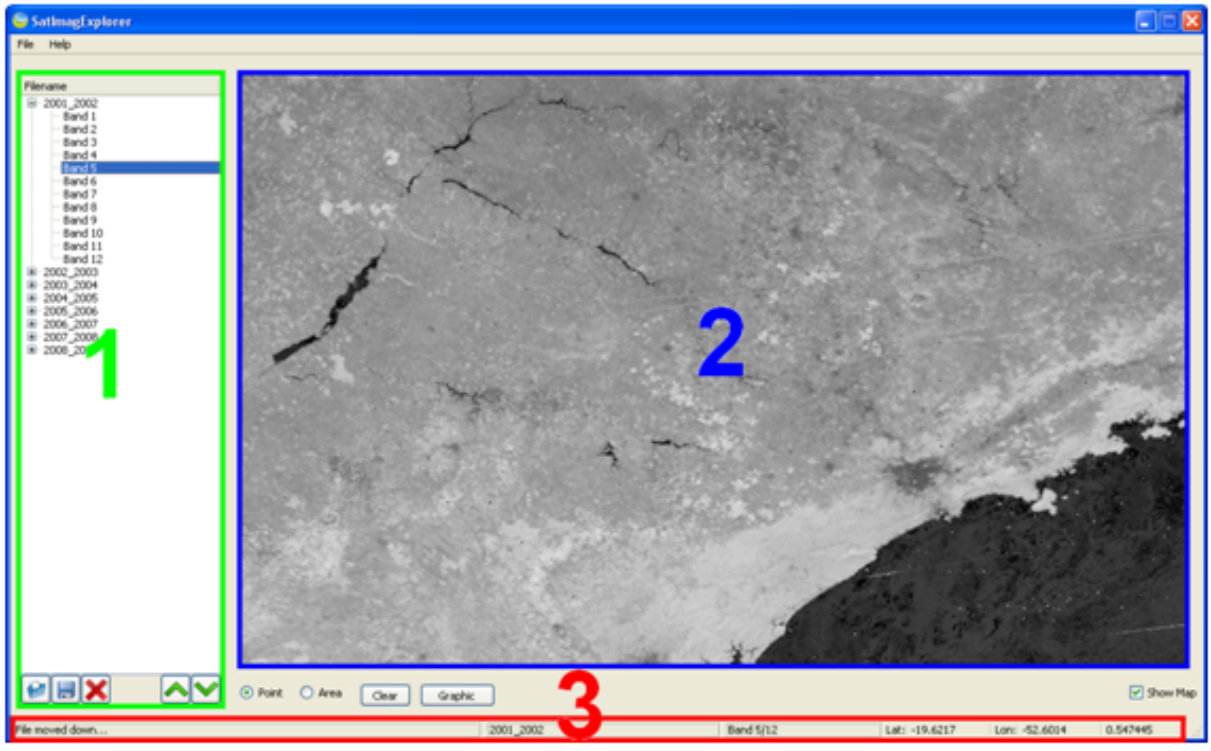



Figure A.2: *SatImagExplorer* interface where 1 corresponds to an area that lists the names of images, 2 is the area of the interface in which the image is presented and 3 shows the status bar.

Once the images are opened they are displayed in the list located in the left side of the screen. Then, the user can select which image will be presented in the right side of the screen. In the status bar, the system presents information about latitude and longitude coordinates of a specific point or region that had been indicated by the mouse position and its value.

Figure A.3 (a) shows an area selected by the user through the mouse. In Figure A.3 (b), the marked area was selected using a coordinates file, which highlights sugar cane

fields in São Paulo state.

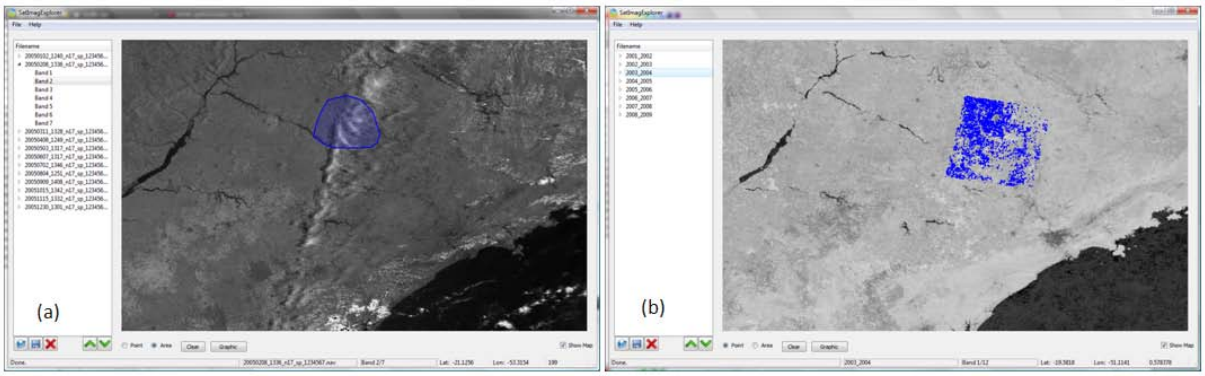



Figure A.3: Images with regions selected by users using two approaches: (a) selection through mouse; (b) selection through coordinates file.

## A.3 Extraction and Inclusion of Indexes

First, the user defines coordinates in the images using both approaches: mouse or file. Thus, *SatImagExplorer* reads the matrix of pixels to generate time series from each pixel in the region of interest for all images opened. The system also allows visualizing time series in the usual graphical format. In the abscissa axis is represented the time scale. Indexes or raw values of each pixel are plotted in the ordinate axis. When an area is selected, charts display the values for all selected coordinates, represented by lines in gray and the average values for region are represented by the solid line in blue as illustrated in Figure A.4.

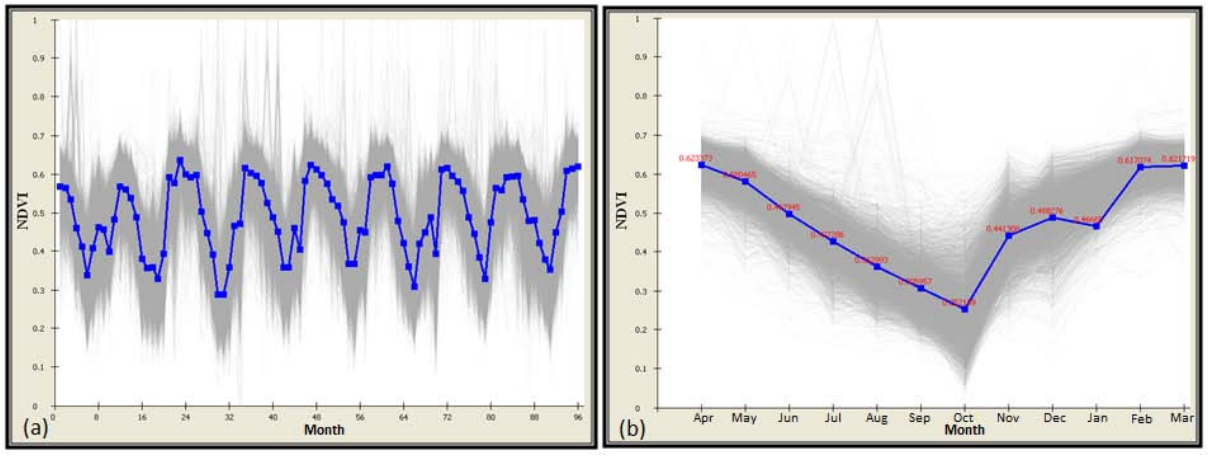



Figure A.4: Graphs show NDVI time series of sugar cane region in the São Paulo state. (a) series of 8 years, (b) series corresponds to year/season of 2006/2007 from April to March.

Figures A.4(a) and A.4(b) show the time series graphs of the same region. In both graphs, the abscissa axis represents time measured in months and the ordinate axis indicates NDVI values. As it can be seen in Figure A.4(a), there are multi-temporal images that correspond to eight years of sugar cane crop seasons (April 2001 to March 2009),

generating a large amount of data whose indexes are very difficult to be manually extracted. Thus, the development of a system that is able to automatically manipulate multi-temporal satellite images is decisive to improve the specialists' job.

Figure A.4(b) shows an example of the standard curve of the sugar cane crop for the crop season of 2006/2007. The NDVI values have increased in November when sugar cane crops are in the beginning of the vegetative development. From December to April, NDVI values have been higher than in other months because sugar cane crops are on top of development and present more biomass. When sugar cane crops begin the senescence phase in May until maturation phase in July, NDVI values decrease.

This analysis can aid improving regional systems of crop season forecasting, since they can be used to confirm (with certain limitations) the beginning and end of crop growing in some predetermined areas.

Through both charts, we can notice that most of coordinates measures are close to average values, but there are outliers in the graph. Values that are outside the range of common value were probably generated by errors in measurements caused by noise or even belong to pixels that do not correspond to sugar cane fields.

Users can type mathematical formulas in the *SatImagExplorer* system in order to calculate and to extract indexes from raw images. These formulas can already be developed as NDVI or be under development/testing to generate new indexes using the channels present in a raw satellite image. The formulas are calculated at runtime and can be changed according to the expert needs.

To avoid typographical errors and mistakes in formulating, the formulas are validated considering a grammar $G = ((< exp >, < term >< factor >, < op_1 >, < op_2 > , < op_un >)(number, band, variable, (, ), +, -, *, /,' sin, cos, exp, log, sqrt), P, < exp >)$, where the set of production rules P are described in Figure A.5 as Backus-Naur extended normal form.

```
<exp> ::= <term> { <op_1> <term> }
<term> ::= <factor> { <op_2> <factor> }
<factor> ::= number | band | variable
| <op_un> ( <factor> ) | ( <exp> )
<op_1> ::= + | -
<op_2> ::= * | / | ^
<op_un> ::= sin | cos | exp | log | sqrt
```

Figure A.5: Grammar used to produce formulas for the indexes calculation.

The mathematical formulas are verified through a descending parser encoded by recursive procedures. After verification, the formulas are converted from infix notation of input to an intermediate format in Reverse Polish Notation and are solved using a heap structure and respecting the precedence of operators. With this procedure, the system ensures that the equation types by user in the system is complete.

The specialist can define variables using the channels in each image, numbers, the four arithmetic basic operations, as well as unary and power operands. These user-defined variables can also be used in other variables, as illustrated in Figure A.6(a).

The ability to add new indexes and measures in an automated and simple way brings flexibility to the process of analyzing and understanding the behavior of regions of interest

in a specific time period. This characteristic makes the *SatImagExplorer* powerful enough to aid in understanding different phenomena and to support decision making.

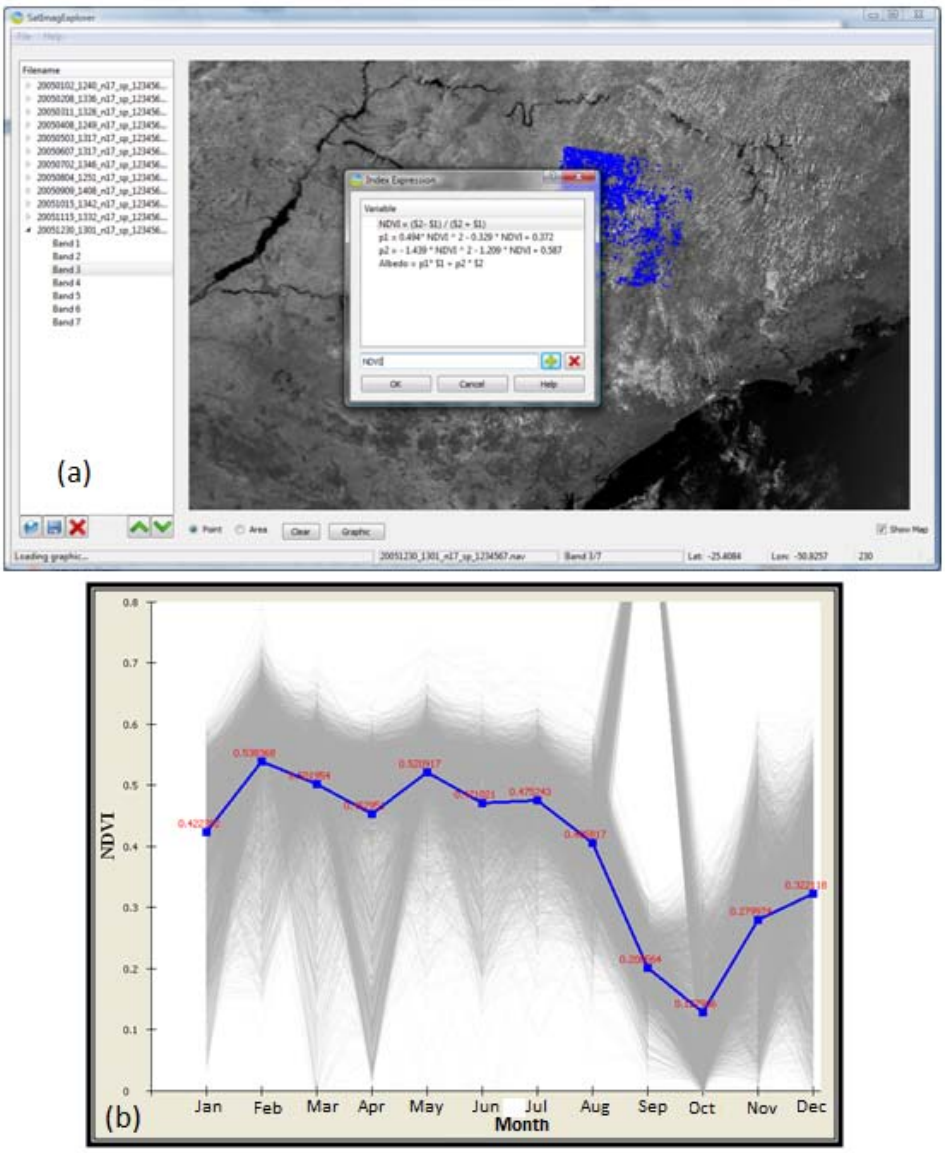



Figure A.6: Option of inputing new indexes in the *SatImagExplorer* system: (a) definition of formulas by users, (b) NDVI time series extracted from raw the NOAA/AVHRR images.

Figure A.6(a) presents the NDVI equation defined by expert to be submitted to raw NOAA/AVHRR images for sugar cane crop fields. Figure A.6(b) shows the NDVI trend in a dataset of multi-temporal images of 2005, from January to December. Equations in conventional notation is typed by users directly in the pop-up window of *SatImagExplorer*.

The outlier value exhibited in September (ninth value on the horizontal axis) of Figure A.6(b) was caused due to lack of data in the region of interest, as presented in Figure A.7. This failure occurs when the satellite is not positioned on the nadir for the receiving antenna at the time of image capture.

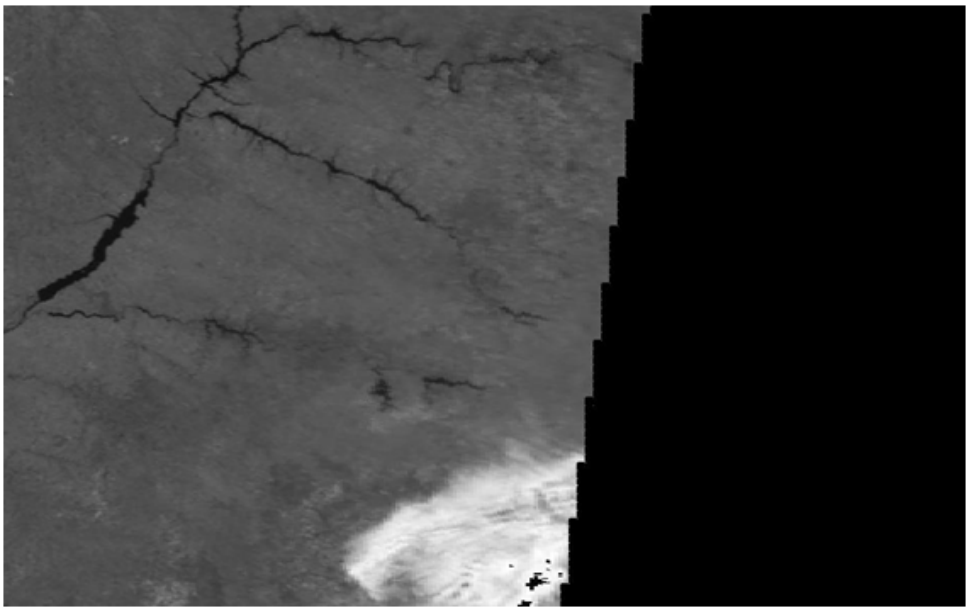


Figure A.7: Image of São Paulo state with noise caused by the positioning of satellite.

## A.4 Validation Process

In order to validate the extraction process performed by *SatImagExplorer* we compared the outputs of our system with a traditional method employed by specialists. In general, they write a set of scripts in IDL language of Envi software to extract values that are associated to pixels in NOAA/AVHRR images.

Figure A.8 shows the data extracted from an eight-year time series of sugar cane crops (April 2001 to March 2009). The blue curve represents the data taken with the ENVI system and the curves in red and green refer to data extracted through *SatImagExplorer*. The data show noise in the periods of January 2002 and March 2008, because this time interval corresponds to a period of considerable rainfall in the São Paulo state. Thus, many NDVI images are discarded during periods with large amounts of clouds since is impossible to obtain correct measurements of NDVI indice.

Both systems (ENVI and *SatImagExplorer*) show similar curves of sugar cane crops with gaps around 20% in the values (January, 2007). These disparities occur due to differences in methods used to accomplish the calculations on different systems, which will be more carefully examined in future work.

Another difference between the systems is the value assigned in the case of noise or lack of data in the image. The ENVI software assigns null values and *SatImagExplorer* uses an invalid negative value.

*SatImagExplorer* also allows that data be submitted to a process of linear interpolation, which gives the specialist a more homogeneous view of the NDVI average values extracted from the images, as it can be seen in Figure A.8.

## A.5 Further Work

In this Section we presented the *SatImagExplorer* system that is being developed to manipulate heterogeneous remote sensing images. The system extracts information associated

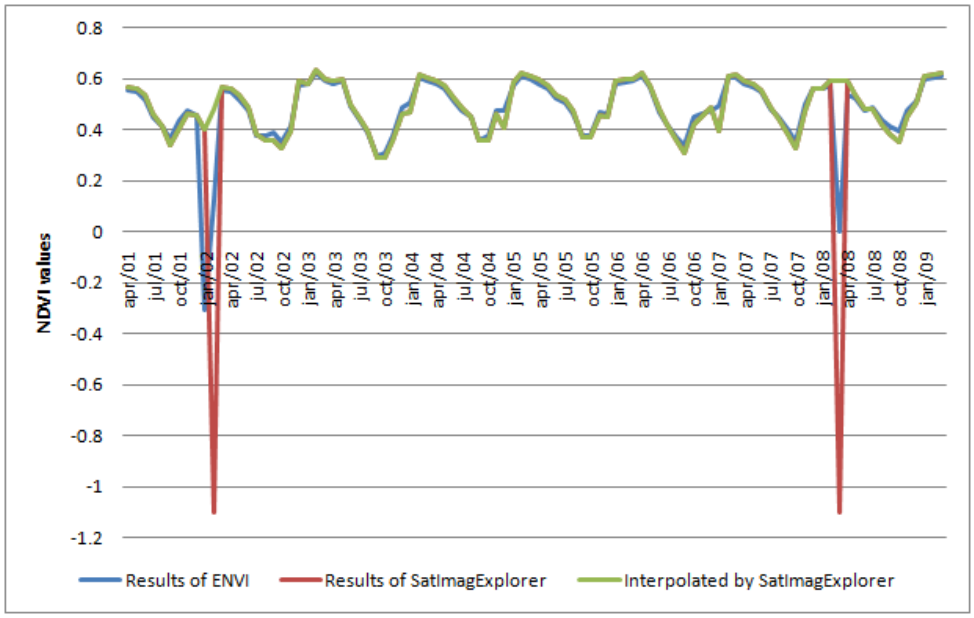



Figure A.8: Time series extracted through IDL scripts of ENVI software in blue and *SatImagExplorer* in red and green.

to each pixel for multi-temporal images generating time series of different indexes.

As further work, we intend to incorporate time series mining techniques in the system, such as association rules and clustering. In addition, the system will be linked to an information visualization system to facilitate and increase the working capacity of the users.

# Appendix B

# Fractal-based analysis of multiple time series

## B.1 Introduction

In general, meteorological data are useful only if associated with a geographic system that allows identifying and relating them to characteristics of the region from where they are collected. Thus, considering the major goal of geographic information systems of providing subsidies for analysts to determine the spatial and temporal evolution of geographic phenomena and their inter-relationship, the information provided by these systems can be a useful tool to be associated with climate data analysis.

For years, meteorologists have studied historical data to understand and characterize the planet's climate and predict possible future scenarios in different regions of the planet. IPCC suggests an increase in the average global temperature what may lead to acceleration of the hydrological cycle and consequent intensification of extreme events (IPCC, 2007).

In this scenario, it is important to understand the trends of extreme phenomena in order to be prepared for such adverse situations, creating conditions to mitigate the problems and make decisions. In this context, the following tasks can be highlighted:

1. Efficient analysis of multiple climate time series to find patterns and trend changes,

2. Identification of climate extremes that indicate regional or global climate changes.

In order to help the domain specialists to perform these tasks, we propose a process for climate time series analysis as an initial improvement of the method described in Chapter 5 - Section 5.4. The implementation of the adjustments in the method has been made by an undergraduate student[1] as part of his initiation in research program in the scope of this doctorate project.

Our approach deals with multiple time series as a multidimensional data stream, such that each time series defines an attribute of the stream. Thus, it is possible to integrate multiple climate variables in a unified process.

The analysis process we propose combines:

1. fractal data stream monitoring for pattern discovery and behavior changes detection,

[1]Santiago Augusto Nunes, that is supported by AgroDataMine project - number 09/53153-3

2. clustering to find similar (or distinct) patterns revealed when data are analyzed with different time granularities and,

3. statistical analysis to identify individual variable changes associated with significant overall behavior variations.

Experimental studies carried out on real climate time series from different regions of Brazil indicate that our approach can be a useful tool to assist specialists in analyzing large amounts of climate data.

## B.2 Proposed Analysis Process

This section presents an analysis process to detect patterns of interest in multiple time series by applying a combination of different techniques, namely fractal data stream monitoring, data clustering and statistical analysis.

Figure B.1 illustrates the whole process. The first step is to associate multiple (georeferenced) climate time series to compose a multidimensional data stream, i.e., each time series determines a stream attribute. For instance, temperature and rainfall time series can be aggregated into a bi-dimensional data stream. This approach allows an integrated analysis of different climate variables in order to discover overall behavior changes over time. In other words, it makes possible to evaluate how the variables are correlated and how these correlations vary, especially when significant behavior changes are identified.

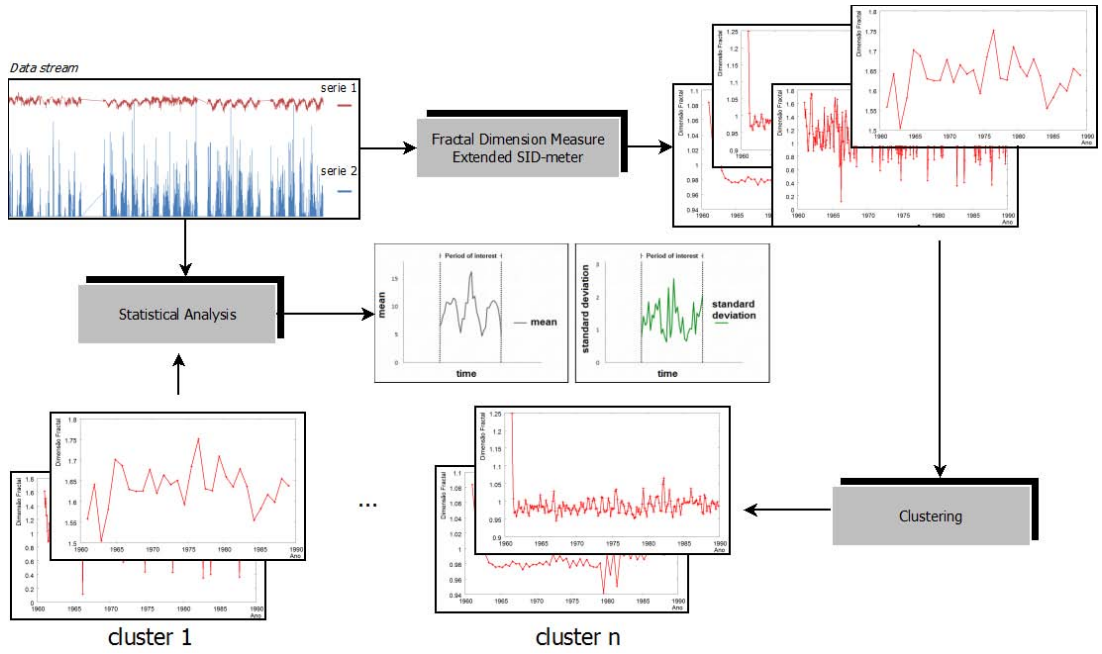



Figure B.1: Analysis process of multiple time series.

The second step is the fractal analysis of the data stream, applying the *SID-meter* approach. The *SID-meter* algorithm was originally proposed to work with a single sliding window of a predetermined size (details in Section 3.4). An extension of the *SID-meter* (Nunes et al., 2010) was designed to handle simultaneous multiple-sized windows over a data stream in a single-read processing. As a result, we can find similar or distinct temporal patterns occurring in different time granularities, e.g., monthly or annually.

The extended *SID-meter* generates sliding windows of different sizes based on initialization parameters. Mainly, the narrowest and the widest windows must be determined by setting:

- the minimum and the maximum values of $n_c$ (counting periods) and $n_i$ (number of events per period);
- increment values of $n_c$ and $n_i$.

For instance, consider a data stream composed of daily measures: by setting $n_c$ to vary from 3 to 6 with an increment of 1, and $n_i$ to vary from 30 (1 month per counting period, approximately) to 365 (1 year per counting period), we define windows ranging from $3 * 30 = 3$ months to $6 * 365 = 6$ years. These parameters should be determined considering the temporal granularity of the data and the purposes of the analysis to be carried out. Thus, domain specialists can be very helpful on this task.

The output of the extended *SID-meter* is a set of $D_2$ (Correlation Fractal dimension) graphs generated for windows of different sizes. Thus, in the third step of the analysis process, we map each graph into a time series of $D_2$ measures. The set of $D_2$ time series is then clustered in order to find similar patterns appearing in different time granularities and distinct patterns, which are detected only in some specific time windows. Moreover, the centroids of the clusters provide some additional information to the domain specialists on how to choose windows of interest for further analysis, i.e., a centroid represents the general pattern of the cluster and therefore the corresponding time window (with a specific time granularity) can be analyzed in more details. The clustering task combines *K-Medoids* (Kaufman & Rousseeuw, 1990), which is a partition-based clustering method, with Dynamic Time Warping (DTW) to measure the similarity among the $D_2$ time series. They are both simple and widely used techniques and presented satisfactory results on our initial empirical studies.

The last step of the process is the statistical analysis, which can be applied considering only the time windows represented by the centroids of the clusters. Basic statistical measures, such as mean and standard deviation, can indicate the variation of individual climate variables in periods of significant alterations in the fractal dimension. The main goal is to discover how each climate variable influences correlation changes, in particular those related to extreme climate events.

We applied this whole process to long, georeferenced climate time series collected from different regions of Brazil, aimed at identifying meaningful behavior changes on the data over time and to associate them with relevant and extreme climate events. The results are presented in Section B.3.

## B.3 Experimental Results

We have performed several experiments on real time series to validate our approach. Two datasets from different regions of Brazil were used to conduct the studies:

1. South: real climate series, provided by Agritempo[2], consisting of daily measurements of rainfall and average temperature measured by 19 meteorological stations in the period from 1994 to 2008.

[2]www.agritempo.gov.br/

2. Southeastern: real climate series, also provided by Agritempo, consisting of daily measurements of rainfall and average temperature measured by 10 meteorological stations of the São Paulo state, in the period from 1995 to 2009.

The time series analyzed in the experiments are georeferenced, as the geographical position of the source meteorological stations affects the results significantly. In general, the climate exhibits different characteristics in geographically distinct regions. Figure B.2 shows the location of the meteorological stations in the South and in São Paulo. It is worth mentioning that the São Paulo state has a significant number of stations when compared to the other states of the Southeastern region and therefore it is analyzed separately.

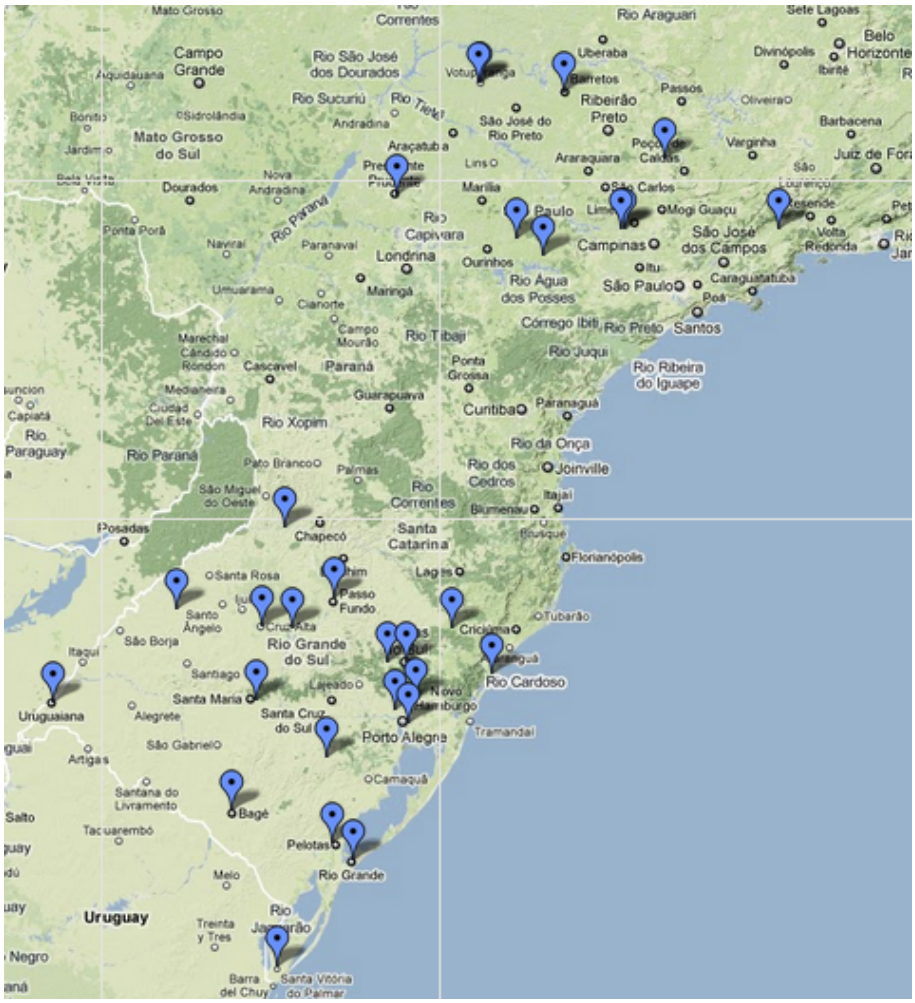



Figure B.2: Geographical coordinates of the meteorological stations in South region and in São Paulo state.

From each dataset we create a bi-dimensional data stream composed of the attributes rainfall and average temperature. As initial parameters of the extended *SID-meter* we set the number of count periods ($n_c$) in the range 2-5 and the movement step ($n_i$) of 1 month to 1 year. In other words, we defined windows ranging from two months to five years.

The graphs depicting the fractal dimension $D_2$ measured over time by the extended *SID-meter* for the two data streams have significantly different trends, although they are related to the same two climate variables and the same time interval. The differences in patterns observed in the graphs evidence the distinct climate characteristics of the two

corresponding regions. Also notice that, for both datasets, the value of $D_2$ remained around 1, which indicates that the average temperature and rainfall variables are correlated. This behavior is consistent with the experts' expectations, since the correlation between these variables ranges from a stronger correlation in a given period to a weaker correlation in others.

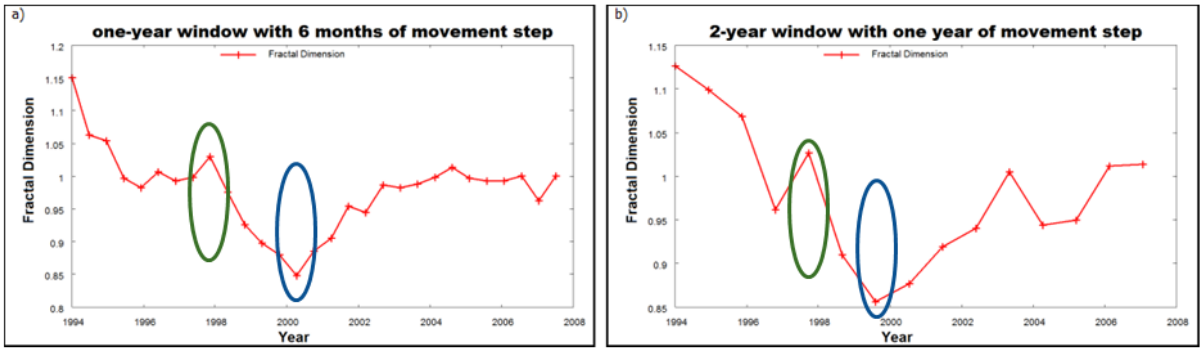



Figure B.3: Variation of $D_2$ for the South region of Brazil. The patterns found show the occurrence of El Niño (green) and La Niña (blue): (a) one-year window with 6 months of movement step; (b) two-year window with one year of moment step.

Figure B.3 shows two $D_2$ graphs for the South region, considering two time windows of different sizes. We can see that, although the graphs are related to different time granularities, there is a similar pattern of $D_2$ variation. Both graphs show a peak in 1998 and a decline in the value of fractal dimension around the year 2000. By assessing the fluctuation of mean and standard deviation for the variables rainfall and temperature during the 1997/1998 and 2000/2001, we can observe higher values of standard deviation in precipitation (varying from 14.7 to 19.8) in the first period and higher values of standard deviation in temperature (varying from 6.2 to 8.9) in the second period. According to specialists and temperature data records, the winters of 2000 and 2001 were marked by low temperatures, very cold weather and snowfall in both years (2000/2001) in Rio Grande do Sul (a state in the South region). Additionally, the 1999/2001 period coincides with the occurrence of La Niña.

The period from 1997 to 1998 suffered the occurrence of a very strong El Niño. Moreover, the Meso-scale Convective Complexes (MCCs) strongly influence the maximum rainfall over northwestern of Rio Grande do Sul during the spring (Velasco & Fritsch, 1987). These systems are a result of the interaction between the subtropical high-level jet during this season and the low-level jet coming from the north. They move towards the east of Southern Brazil after they are originated over the west (Grimm & Tedeschi, 2008). During El Niño phenomenon, extreme events may move to the south or to the north, but both indicate the importance of the western part of Southern Brazil in the increase of extreme events during El Niño years. MCCs are frequent in this area and these systems take advantage of the circulation anomalies during El Niño episodes (Grimm & Tedeschi, 2008)).

Figure B.4 shows two $D_2$ graphs for the São Paulo state, considering two time windows of different sizes. They also present the variation of $D_2$ with a similar pattern of behavior.

Figure B.4 (a) corresponds to a two-year window for which the value of fractal dimension is updated every year (one-year of movement step). We can observe that the value of $D_2$ decreases in the period 2000 to 2002, which coincides with the occurrence of La

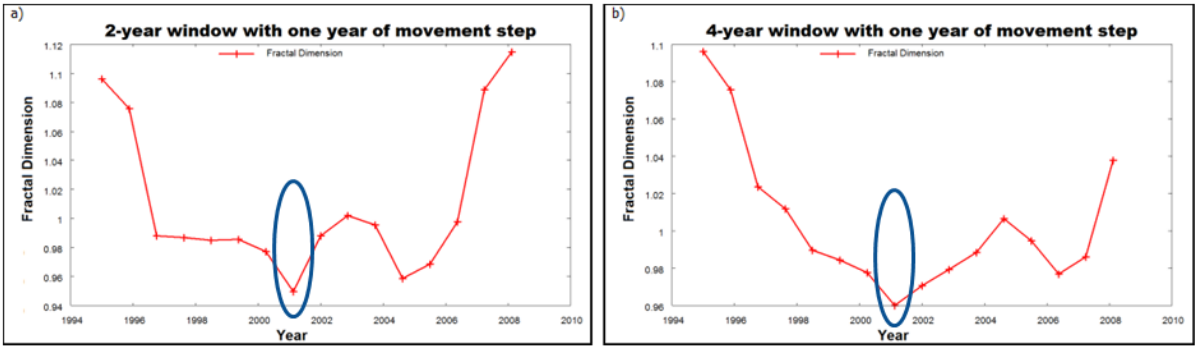



Figure B.4: Variation of $D_2$ for the São Paulo state. Patterns found show the occurrence of La Niña (blue): (a) two-year window with one year of movement step; (b) four-year window with one year of moment step.

Niña. In this period (2000/2001), there was a warning in São Paulo with the possibility of electricity rationing in South Central, probably due to a rainfall decrease in the central region of Brazil. In the Southeast, the years 2000 and 2001 were characterized by a pronounced dry season. In 2000, the months of April, May and June were very dry and the rainy season was delayed.

Both graphs in Figures B.4(a) and (b) show a trend change between 2004 and 2009. Our analysis indicates that the year 2007 shows a wide variation in the values of both mean and standard deviation for the variables temperature and rainfall. In 2007, the region had a hot and rainy Summer, and a wetter and slightly warmer Fall. According to the specialists, in El Niño years there is an increase in the occurrences of heavy rains between October and February, with a break in January. However, in the case of total monthly precipitation, the signal is less pronounced, occurring in some regions and not in others.

In general, we observed that, for both regions we analyzed, El Niño years are related to a decrease in the correlation between temperature and rainfall (higher $D_2$). On the other hand, La Niña occurrences are related to an increase in this correlation (lower $D_2$). According to meteorologists, in years of La Niña rainfalls are caused by Convective Complexes of Meso-scale, which are heavy rains, severe thunderstorms, and it can cause a discontinuity in the relationship between variables.

Our analysis process also includes a clustering task, which groups the graphs generated by the extended *SID-meter*. The main goal, in these experiments, is to find similar patterns that are revealed even if data is analyzed under different time granularities, such as the relevant $D_2$ variations identified in Figures B.3(a) and B.3(b). Furthermore, the specialist also receives as output the graphs selected as centroids of the clusters. Thus, she/he can determine the suitable window sizes to be used to highlight the climatic phenomena that are being evaluated.

For example, Figure B.5 shows the centroids of five clusters generated to group the graphs of São Paulo. With the centroid information experts can have a general picture of the set of graphs within each cluster and define which clusters could be used to further analyses. This approach can make the analysis task easier, as the number of possible combinations of window configurations (counting periods and movement step) can be considerably high.

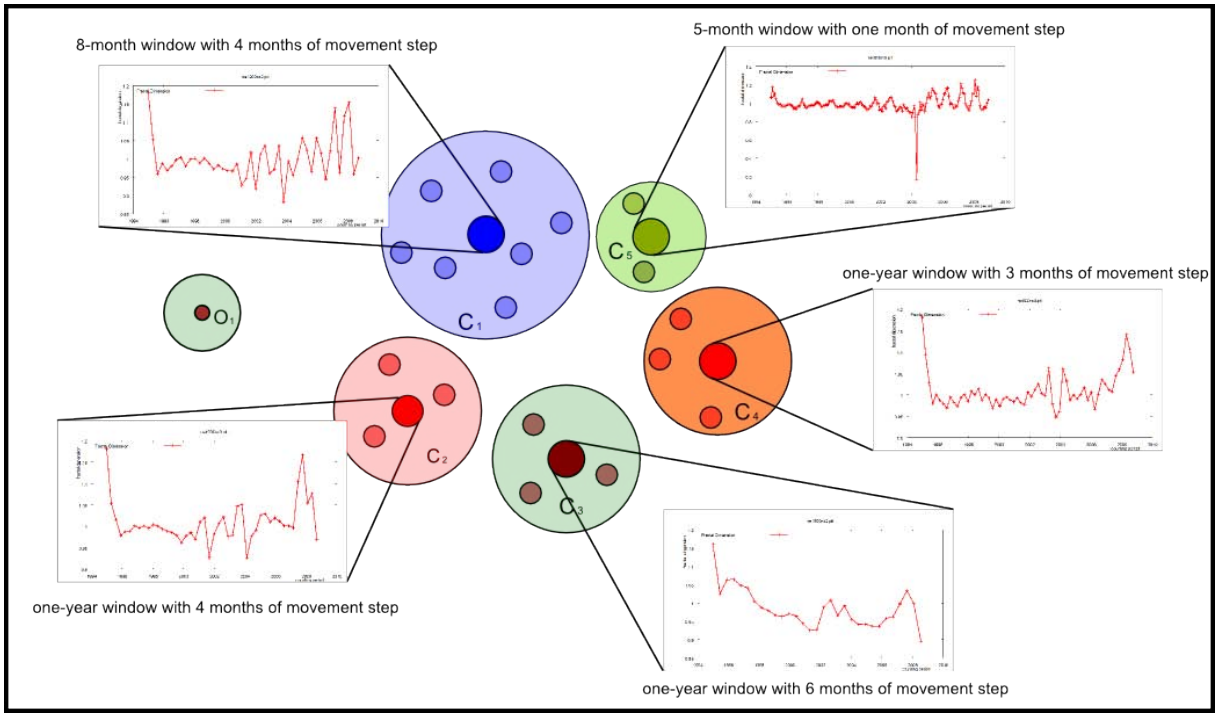



Figure B.5: Clusters of $D_2$ graphs for the state of São Paulo (C - clusters and O - outliers).

## B.4 Future Work

The time series analysis of precipitation and temperature together through the fractal theory shows that there is a rhythmic pattern in time between the variables. When a trend change occurs, the approach presented in this work is able to identify the change and the variables responsible for the pattern variation. Thus, it was possible to correlate those variations with changes in global weather patterns, such as El Niño-Southern Oscillation.

These results indicate that time series studies with more than one variable at a time may allow the identification of patterns in more details than using just one variable. Furthermore, our approach can aid to understand the interdependence among meteorological variables. Additionally, analysis performed directly on the data of weather stations allows identifying the individual contribution of different stations, detecting extreme events in each region.

Future work includes trend change forecasting based on the analysis of general climate data behavior associated with semantic information provided by domain specialists.