Computational framework to analyze agrometeorological, climate and remote sensing data: challenges and perspectives

Luciana A. S. Romani[1,2], **Agma J. M. Traina**[1], **Elaine P. M. de Sousa**[1],
Jurandir Zullo Jr.[3], **Ana M. H. Avila** [3],
Jose Fernando Rodrigues Jr. [4], **Caetano Traina Jr.**[1]

[1]Department of Computer Science - University of São Paulo at São Carlos -Brazil

[2]Embrapa Agriculture Informatics at Campinas - Brazil

[3]Cepagri - State University of Campinas - Brazil

[4]Federal University of São Carlos at Sorocaba - Brazil

{alvim,agma,parros,caetano}@icmc.usp.br

{jurandir,avila}@cpa.unicamp.br, junio@ufscar.br

Abstract. *In the past few years, improvements in the data acquisition technology have decreased the time interval of data gathering. Consequently, institutions have stored huge amounts of data such as climate time series and remote sensing images. Computational models to filter, transform, merge and analyze data from many different areas are complex and challenging. The complexity increases even more when combining several knowledge domains. Examples are research in climatic changes, biofuel production and environmental problems. A possible solution to the problem is the association of several computational techniques. Accordingly, this paper presents a framework to analyze, monitor and visualize climate and remote sensing data by employing methods based on fractal theory, data mining and visualization techniques. Initial experiments showed that the information and knowledge discovered from this framework can be employed to monitor sugar cane crops, helping agricultural entrepreneurs to make decisions in order to become more productive. Sugar cane is the main source to ethanol production in Brazil, and has a strategic importance for the country economy and to guarantee the Brazilian self-sufficiency in this important, renewable source of energy.*

1. Introduction

The Fourth IPCC Assessment Report - ”Climate Change 2007”, indicates a disturbing situation regarding the temperature increase in the planet due to natural and anthropogenic effects [IPCC 2007]. According to IPCC reports, the precipitation should also increase. Researches have been pursued to forecast the changing in the temperature patterns - which is steadily increasing - to define methods to reduce the emission of greenhouse gases and to adapt crops to the new conditions of increased temperature.

An alternative to the reduction of emissions of the greenhouse gases is the replacement of fossil fuels by renewable sources. In Brazil, sugar cane is the main crop used to produce ethanol, a kind of biofuel. Among the several determinant factors of the Brazilian agricultural production, the climate is the main factor affecting the diversity of conditions in all regions.

In fact, according to [Rosseti 2001], about 90% of all losses in the Brazilian agriculture registered until the middle of 90's were directly related with two main climatic factors: dry spells during the reproductive stage and rainfall excess during the harvest period. As an attempt to decrease mainly these two climatic-associated risks for agriculture in Brazil, the Brazilian Ministry of Agriculture bureau started in 1996 an official program of agricultural zoning to define planting calendars for the main crops in country. The calendars have been calculated to achieve less than 20% of risk regarding climate problems. Agrometeorologists have used climate data and agrometeorological methods to prepare these calendars.

In particular, climate data from earth meteorological stations (for instance, pluviometric and temperature) have been used for a long time already. Recently, data from remote sensing images have also been used. The analysis of these data is important to the development of innovative and technologically feasible solutions to assist in crops monitoring and forecasting. Also, the use of information extracted from remote sensing images and climate data can indicate the occurrence of some phenomena that had occurred in the past and have influenced the production. This influence can determine a good (above the average) harvest or not. Moreover, this knowledge can improve the production estimate of agricultural crops.

Satellites have been used to provide the images for monitoring and forecasting, especially the National Oceanic and Atmospheric Administration (NOAA) satellites. A useful sensor over the past few years has been the Advanced Very High Resolution Radiometer (AVHRR) on board of NOAA. This sensor is very important to studies of ecosystems due to the availability of long time series of its image data. Furthermore, other advantages that make AVHRR data attractive for many studies are its global coverage and free data access.

Then, if we combine climate data and images from AVHRR, we have important sources of useful information and meaningful knowledge. However, AVHRR images and climate databases are quite large in many countries. The analysis of this huge amount of data is a challenge. Consider, for example, datasets with climate and remote sensing data of some regions of sugar cane fields. A feature selection algorithm can identify the most relevant attributes of the datasets, that is, the ones that keep the majority of the information regarding a given criterion. It is also helpful to know which attributes from the dataset are correlated to the others, so an attribute (feature) selection algorithm can be designed. Moreover, it is also interesting to know which attributes can better approximate the values of the others. The detection of correlated attributes, their importance and precedence in climate datasets can improve the agricultural monitoring, helping the specialists during the decision making process.

In this scenario, due to the importance of the agriculture production for the Brazilian economy, impact assessment studies for seasonal climate variations and for climate changes have became a major priority. However, the massive data volumes generated and the growing processing complexity bring up several problems and research challenges, which have driven researches in many areas of the the Computer Science.

In this work, we address the main following problems::

- How to organize, select and weight attributes which are the most relevant to be considered in further studies and agrometeorological models?
- How to track multiple variables together in order to find variation patterns in the climate of a region?
- How to discover relevant patterns in climate and remote sensing time series?
- How to visualize a large and varied amount of data to analyze it more efficiently?

In order to address these problems, we proposed a framework to analyze, monitor and visualize climate and remote sensing time series to improve researches on sugar cane crops. The methods combine different techniques such as fractal theory, time series mining and visualization. This work is in progress, and we have proposed a method that combines one supervised algorithm (Omega) to perform discretization on climate and remote sensing data [Ribeiro et al. 2008] with the Apriori algorithm to identify association rules [Agrawal and Srikant 1994]. Other results have also been presented [Romani et al. 2009a, Romani et al. 2009b]. The research is directly related to the challenge number 2, posed by the Brazilian Computer Society: *Computational modeling of complex systems: artificial, natural, socio-cultural, and human-nature interactions* depicted by the Brazilian Computer Society.

The paper is organized as follows. Section 2 presents important concepts concerning this work. Section 3 describes the proposed framework. Section 4 discusses experiments and initial results. Section 5 concludes the paper.

2. Related concepts

This section presents some important concepts and algorithms that we have used to integrate the proposed framework. These concepts involve fractal theory, data mining and visual analytics.

2.1. Fractal theory

Fractal is defined as an object that presents roughly the same characteristics regardless of the scale where it is analyzed. Thus, small scale details are similar to large scale characteristics [Traina Jr. et al. 2005].

Fractal concepts have been applied to several tasks in data mining, such as selectivity estimation [Baioco et al. 2007], clustering [Barbará and Chen 2000], and forecasting [Chakrabarti and Faloutsos 2002], among others. One relevant information provided by the Fractal theory is the estimation of the intrinsic dimension (D) of the dataset [Pagel et al. 2000], which is based on the dataset fractal dimension. Intrinsic dimension is a measure of the amount of information that the dataset represents [Traina Jr. et al. 2005].

For example, consider a set of points distributed along a line. Its intrinsic dimensionality is equal to one. If the set is embedded in a higher dimensional space the intrinsic dimensionality continues equal to one. For instance, in other words, if the line is embedded in a three dimensional space, the set has three linearly correlated attributes and its intrinsic dimensionality is one. The intrinsic dimension of this dataset embedded in a two dimensional space will still be equal to one. Thus, the intrinsic dimension conveys the actual data dimension regardless the dimension of the embedded space.

Faloutsos and Kamel [Faloutsos and Kamel 1994] proposed the use of the intrinsic dimension as a tool to measure the non-uniform behavior of real datasets and to indicate the existence of attribute correlations, linear or not. The number of attributes in a dimensional dataset determines its embedded dimension E. However, if there are correlated attributes, its intrinsic dimension is lower than E, i.e., $D < E$. This conjecture has supported the development of a variety of works whose results confirm that the intrinsic dimension is a meaningful measure to analyze the data distribution [Traina Jr. et al. 2000, Traina Jr. et al. 2005, Sousa et al. 2007b].

The fractal dimension of statistically self-similar datasets, such as real datasets, can be determined by the Correlation Fractal Dimension D_2. An efficient approach to measure the fractal dimension of datasets embedded in E dimensional spaces is the BoxCounting method [Schroeder 1991], which employs D_2 as presented in Equation 1.

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

$$D_2 \equiv \frac{\partial log(\sum_i C_{r,i}^2)}{\partial log(r)} \qquad r \in [r_1, r_2] \tag{1}$$

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 ith cell. An efficient algorithm (linear cost on the number of elements in the dataset) to compute D_2 was proposed in [Traina Jr. et al. 2000]. Thus the D_2 Fractal Dimension can be useful for real datasets, as it gives a suitable approximation of the dataset intrinsic dimension D with a feasible computational cost.

2.2. Fractal Correlation

The intrinsic dimension D, estimated by the Correlation Fractal Dimension D_2, indicates the minimum number of attributes necessary to represent a dataset. D can also be used to discover how many and which attributes may be employed to reduce the data dimensionality. For this purpose, Sousa et all [Sousa et al. 2007b] proposed the FD-ASE algorithm to identify 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 subgroup of attributes to represent the essential characteristics of data.

Correlation means that the value of an attribute can be approximated from some other attributes. Sousa et all [Sousa et al. 2007b] define 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. The second one indicates that an attribute can be only approximated from other attributes, as in fractal correlations. In order to quantify the correlation among attributes, the authors use a threshold ξ that ranges from zero - meaning complete correlation - up to one, when the attributes are independent. It is also defined a subset of attributes B_p as a base of a group p. A correlation group G_p includes the B_p and every attribute correlated to all attributes in B_p, but excludes the attributes not correlated to the full base B_p. Finally, the correlation base CB is defined as a subset of attributes whose partial intrinsic dimension approaches the intrinsic dimension of the whole dataset. That is, CB keeps the most relevant attributes, which should be used to represent the dataset information.

2.3. Fractal Dimension to monitor data streams

A data stream is an ordered sequence of events (or items) $\{e_1, e_2, ..., e_n\}$ in which an event e_j is defined by a set of E measured attributes a_i, such that each $e_j = (a_1, ..., a_E)$.

The first technique aimed at measuring the intrinsic dimension (D) of data streams over time was the SID-meter [Sousa et al. 2007a]. In this work, a data stream is considered a sequence of events $e_1, e_2, ..., e_n$ each of which represented by an array of E measurements. The events occurring within a time interval can be seen as a dimensional dataset of dimension E. Thus, the fractal dimension D_2 can be used to estimate the intrinsic dimension D of a bounded sequence of events.

The SID-meter uses an event-based sliding window divided into n_c sequential periods, named *counting periods*. In each period a predefined number n_i of incoming events is, i.e. n_i events are processed and when a counting period is complete, the events of the oldest one are discarded. Thus, n_c and n_i respectively specify the length of the window and its movement step. Figures 1a) and 1b) illustrate successive sliding windows, divided into four counting periods, through a data stream composed of the attributes a_1, a_2 and a_3.

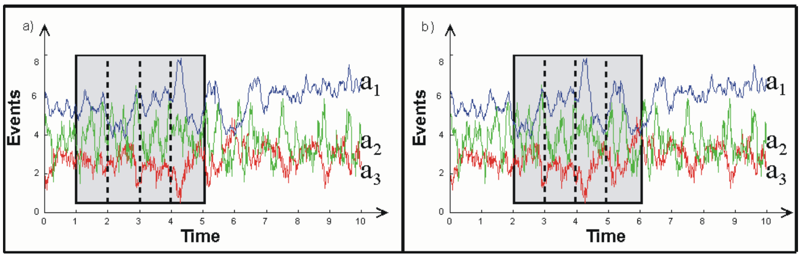


Figure 1. Counting periods of a sliding window

In each movement step, a new value of D is computed considering the events in the current window. Thus, the SID-meter tracks the intrinsic dimension over time, following the behavior changes in the stream.

In our work domain, applications in remote sensing and climate areas have generated continuous sequences of data over long periods of time. These data can be typically stored as time series or be considered data streams as well. Thus, we can take advantage of algorithms and techniques developed to data stream analysis. In particular, SID-meter allows the specialist to follow the data behavior as it evolves over time, what is important for climate datasets.

2.4. Time Series Mining

In the past few years, several methods have been proposed to mine time series. The main time series problems are motif discovery [Chiu et al. 2003], longest common subsequence matching, sequence averaging, segmentation and indexing [Keogh and Kasetty 2002].

Methods of geographic data mining have also been proposed in the literature. Wu et al. [Wu et al. 2008] proposed the GEAM (Geographic Episode Association Pattern Mining) algorithm to find association patterns in abnormal event sequences. Harms and Deogun [Harms and Deogun 2004] developed the MOWCATL algorithm to mine frequent association rules from sequential datasets. They presented an application to drought risk management. Both algorithms work over event sequences with discrete events. For example, an event type is denoted as a tuple in the form $< attribute, level >$, where *attribute* is a variable such as rain or temperature, and *level* is the corresponding level of the variable´s value such as *low*, *normal* or *high*.

Honda and Konishi [Honda and Konishi 2001] proposed a framework for time series image mining. They applied the method to weather satellite cloud images taken by GMS-5. The proposed algorithm extracts features from images and cluster images by the changing in the mass of cloud. Julea et al. [Julea et al. 2006] presented an application of SPADE algorithm [Zaki 2001] to extract frequent evolutions observed on geographical zones represented by pixels. The authors use feature vectors to represent satellite images or symbols associated to a discretization interval representing reflectance values of satellite channels.

2.5. Visual Analytics

Visual Analytics is defined as *the science of analytical reasoning facilitated by interactive visual interfaces* [Thomas and Cook 2005]. The term Visual Analytics is a super area of what has been called, so forth, Visualization. Introduced in 2005, Visual Analytics formally states that visualizing data is not only a matter of graphic design, but a systematic effort that encompasses the following disciplines: analytical reasoning, data representations and transformations, production, presentation, and dissemination, and visual representations and interaction. Indeed, according to Keim [Keim et al. 2006], Visual Analytics is more than just visualization and can rather be seen as an integrated approach combining computer graphics, human factors and data analysis.

The reasons behind the field of Visual Analytics come from its potential in promoting well-founded insights in situations that involve planning, assessment and decision making. The use of Visual Analytics relies on a set of fundamentals; such fundamentals that derive from the fact that visual data representations:

- increase cognitive resources, for example, by expanding human working memory;
- reduce search, such as by condensing data using graphical elements;
- enhance the identification of patterns, especially by organizing things spatially;
- support the perception of inferences that are not obvious when presented in textual form;
- permit the monitoring of multiple events;
- provide a manipulable medium that, in contrast to static diagrams, allows the observation of data from many different perspectives.

The vital factors for the practice of visual analysis are time saving and expanded perception, assisting analysts in tasks as:

- comprehension of situations in the context of time;
- definition of trends and identification of key events that lead to current circumstances;
- monitoring of ongoing phenomena, so that possibilities of interest can be identified and managed;
- determination of relevant indicators;
- supporting strategy;
- communicate knowledge.

In our work we apply a visual analytical tool to explore the climate and remote sensing data. This tool has a set of techniques to visual analytics, allowing the user to move from different techniques, getting the best of each one and acquiring the insight to make the needed decisions over the data.

3. Framework to analyze agrometeorological data

In this work, we propose a framework based on three main layers, illustrated in Figure 2, that to aim at solving the problems listed in Section 1. We have organized NOAA-AVHRR images, which were processed, geometric corrected and stored in database. The climate database was generated from a complete database support by Agritempo (*www.agritempo.gov.br*).

The mining algorithm layer is composed of algorithms based on Fractal Theory and time series mining techniques, as discussed in Section 2. New algorithms to mine patterns in agrometeorological time series are also being developed. These methods are, based on time series mining techniques, presented in Section 2.4. Moreover, we are adapting the MetricSPlat platform [Rodrigues Jr. et al. 2008] to allow better visualization of climate and remote sensing time series in the Visualization Layer. New distance function are also being developed in our research group.

Figure 2 presents a schema divided into layers. Each layer is presented in more detail in the following sections.

3.1. The Storage Layer

The storage layer is basically composed of two databases: remote sensing images and climate data. The remote-sensing-image database stores images from NOAA-AVHRR satellites. At the beginning of this work, we organized this image database, provided by the Center of Meteorological and Climatic Researches applied to Agriculture (CEPAGRI) of the State University of Campinas (UNICAMP). This database has stored images since April 1995 and nowadays, there are more than five terabytes of data.

The climate data came from surface stations installed in the sugar-cane regions and refer to measures of rainfall, maximum and minimum temperature. The data on the climate database was extracted from the Agritempo database. Agritempo (*www.agritempo.gov.br*) is a system developed to organize and store climate and agroclimate data from several institutions of Brazil.

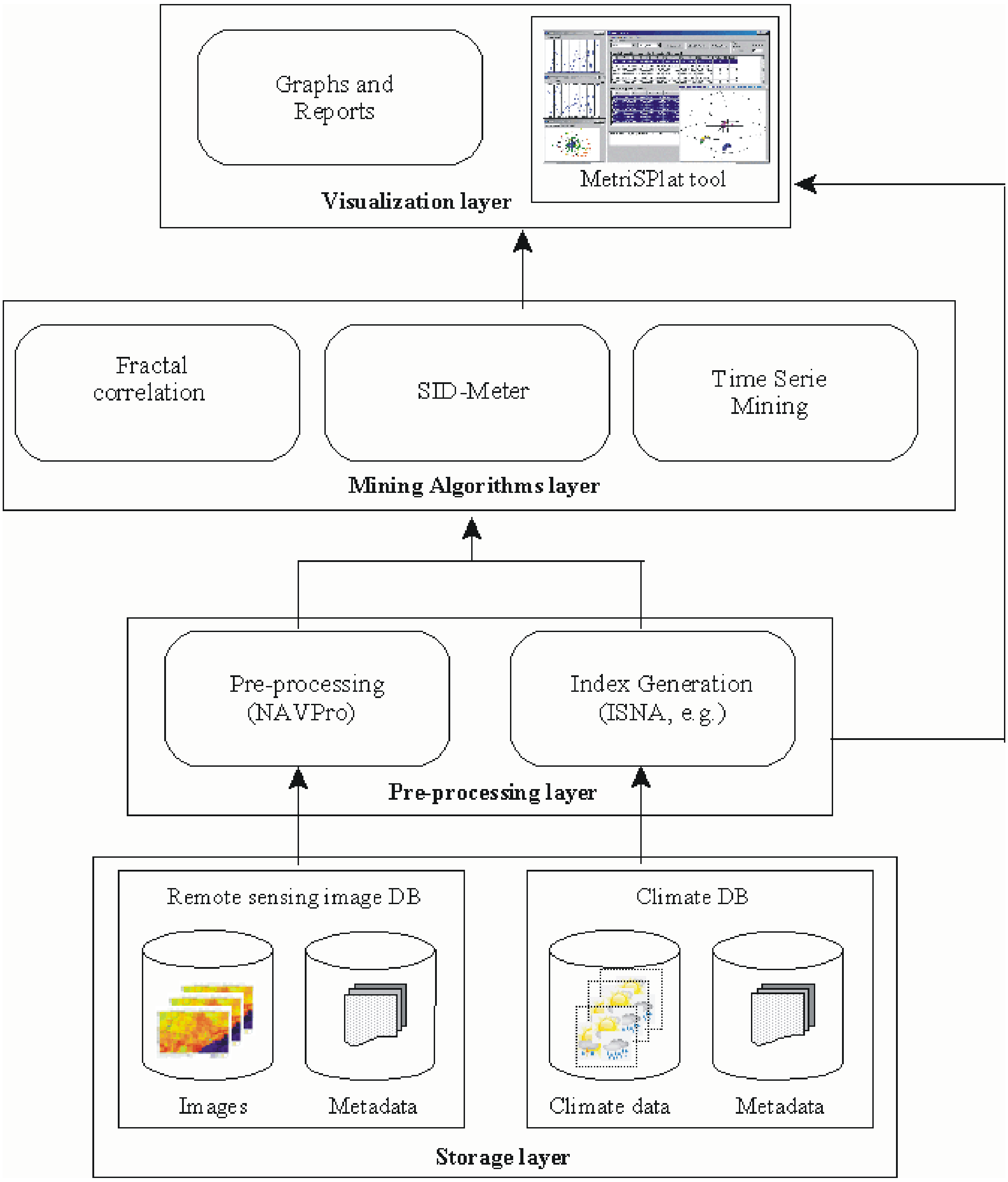


Figure 2. Framework schema

3.2. The Pre-processing Layer

In the pre-processing layer it was necessary to adapt tools to pre-process and prepare data to the analysis phase. AVHRR images have geometric distortions caused by Earth curvature, rotation and satellite clock errors, attitude errors and imprecise orbital [Rosborough 1994]. Then, for land applications, these distortions must be corrected to avoid errors where high geometric precision is required. Thus, AVHRR images must be submitted to a pre-processing system to format conversion from raw images to intermediate format, to make radiometric calibration and geometric correction, to identify pixels classified as cloud and to generate the maximum value of NDVI images.

The NAVPRO system developed by [Esquerdo et al. 2006] was used to pre-process AVHRR images to make geometric corrections. NAVPRO is an automatic set of C-shell scripts that call the subroutines of NAV (NAVigation) [Emery et al. 1989], developed by the Colorado Center for Astrodynamics Research (CCAR), Aerospace Engineering Sciences, with the University of Colorado, Boulder, USA.

The last phase of NAVPRO system is the generation of several products. One of them combines channels 1 (red) and 2 (near infrared) to calculate an important vegetation index - NDVI, proposed by [Rouse et al. 1973]. NDVI index is closely correlated to the leaf area index, green biomass and productivity [Holben et al. 1980]. The effects of shadow, aerosols and water vapor are minimized by the generation of Maximum Value Composite (MVC) NDVI, as described by [Holben 1986]. The MVC is made by using only images from the same satellite.

Algorithms to calculate indexes from climate data have been developed in this work, as we use these indexes to search for correlations. The agroclimatic conditions through the period of analysis are described by WRSI. To describe it, a water balance is made, and maximum and real evapotranspiration are calculated on a period of 10-days, as well as biweekly and monthly periods.

In this works, we used the water balance calculus proposed by [Thornthwaite and Mather 1955]. In the water balance, some variables are calculated such as real, potencial and maximum evapotranspiration. Evapotranspiration is the sum of evaporation and plant transpiration. The ratio between real evapotranspiration and maximum evapotranspiration is indicated by WRSI. This index varies from zero to one and represents a fraction of the amount of water consumed by the plant and the amount of water that would be used by the plant to ensure maximum productivity.

3.3. The Mining Algorithm Layer

Building a framework for climate and remote sensing data analysis is a long term project. Therefore, initially some algorithms were developed, tested, adapted, and improved to be used in the analysis process. The idea is to construct an application program interface (API) that encapsulates different modules, allowing the researchers to explore, analyze and visualize different types of data.

The FD-ASE algorithm (Section 2.2) was tested and some extensions have been added to it. In particular, one of the improvements is related to the incorporation of semantic information that aids to understand the groups of correlated attributes detected by the FD-ASE. For example, let A be a six-dimensional dataset $A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$, where a_i denotes an attribute. When FD-ASE is performed, it returns a group $G1 = \{a_1, a_3, a_5, a_6\}$ with base $B = \{a_1, a_3\}$. This result ($G1$) means that the attributes a_5 and a_6 are correlated to the attributes in the base, namely a_1, a_3. However, during the tests, specialists seemed to be confused about the weight of the correlation. Pearson's correlation, for example, presents a value r that indicates the intensity of linear correlation between two attributes [Pearson 1896]. Then, we are developing some extensions to facilitate this interpretation by the specialists, allowing them to measure the weight of the attributes in each correlation.

In this layer, we also explore the intrinsic dimension as a foundation concept for data stream monitoring in agrometeorological applications, based on SID-Meter (Section 2.3). The appropriateness of the fractal-based approach to monitor data streams is illustrated by employing a statistical approach to compare data in consecutive time periods, pointing out the attributes that are responsible for the trend changes and how they influence them.

Additionally, algorithms to mine patterns on time series are being developed. These algorithm are based on techniques discussed at Section 2.4.

3.4. Visualization Layer

The results generated by the algorithms are presented in graphical and textual ways. Additionally, we have been using the MetricSPlat, a tool that combines visualization techniques and content-based data retrieval methodologies [Rodrigues Jr. et al. 2008]. This tool allows to integrate new modules with minimal effort. MetricSPlat provides a framework with modules to content-based data retrieval and visualization based in parallel coordinates, scatter plots, star coordinates and fastmap. We only define the data domain (images, video, audio, series, and complex data) and the retrieval goals.

The MetricSPlat generates a whole package of ways to explore the data, allowing the user to seamlessly move from one to the other, getting the best of each technique and acquiring the insight to make the needed decisions over the data. For illustration purposes, Figure 3 presents a MetricSPlat snapshot with the visualizations of the Araraquara dataset, which includes climate and remote sensing data of the town Araraquara, such as rainfall, maximum temperature, minimum temperature, NDVI, WRSI, mean temperature, ETP (potential evapotranspiration), ETR (real evapotranspiration) and ETM (maximum evapotranspiration).

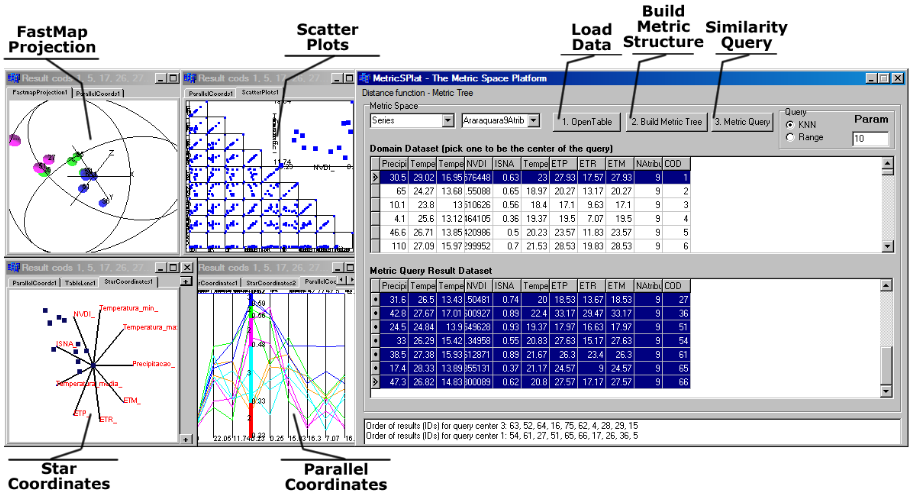
In this example, the main screen, on the right, presents the data table and the metric query result. Figure 3 also presents different ways of data visualization, including parallel coordinates, scatter plots and fastmap visualizations to the metric query result.

4. Experimental Results

In this section we discuss the applicability of the framework we have proposed and preliminary experimental results on real datasets. In particular, we discuss initial studies on techniques of the Mining Algorithms Layer.

4.1. Method to identify groups of correlated attributes

We have tested the FD-ASE algorithm to identify groups of correlated attributes in datasets that includes climatic and NDVI (from remote sensing image) data. Thus, for our experiments, we selected ten sugar-cane producer regions in São Paulo state.



Distance function - Metric Tree

Metric Space: Series, Araraquara9Atrib

Query: ● KNN ○ Range   Param: 10

Domain Dataset (pick one to be the center of the query)

Precipi	Tempe	Tempe	NVDI	ISNA	Tempe	ETP	ETR	ETM	NAtribu	COD
30.5	29.02	16.95	576448	0.63	23	27.93	17.57	27.93	9	1
65	24.27	13.68	.55088	0.65	18.97	20.27	13.17	20.27	9	2
10.1	23.8	13	510626	0.56	18.4	17.1	9.63	17.1	9	3
4.1	25.6	13.12	464105	0.36	19.37	19.5	7.07	19.5	9	4
46.6	26.71	13.85	420986	0.5	20.23	23.57	11.83	23.57	9	5
110	27.09	15.97	299952	0.7	21.53	28.53	19.83	28.53	9	6

Metric Query Result Dataset

Precipi	Tempe	Tempe	NVDI	ISNA	Tempe	ETP	ETR	ETM	NAtribu	COD
31.6	26.5	13.43	.50481	0.74	20	18.53	13.67	18.53	9	27
42.8	27.67	17.01	600927	0.89	22.4	33.17	29.47	33.17	9	36
24.5	24.84	13.9	549628	0.93	19.37	17.97	16.63	17.97	9	51
33	26.29	15.42	.34958	0.55	20.83	27.63	15.17	27.63	9	54
38.5	27.38	15.93	612871	0.89	21.67	26.3	23.4	26.3	9	61
17.4	28.33	13.89	355131	0.37	21.17	24.57	9	24.57	9	65
47.3	26.82	14.83	300089	0.62	20.8	27.57	17.17	27.57	9	66

Order of results (IDs) for query center 3: 63, 52, 64, 16, 75, 62, 4, 28, 29, 15
Order of results (IDs) for query center 1: 54, 61, 27, 51, 65, 66, 17, 26, 36, 5

Figure 3. Snapshot of MetricSPlat visualization tool

A reason for choosing sugar cane satellite images to do the experiments is the strategic importance of this crop to substitute the gasoline usage in some countries, such as Brazil. The attributes of the dataset are a_1 (rainfall), a_2 (maximum temperature), a_3 (minimum temperature), a_4 (NDVI) and a_5 (WRSI).

We first applied the FD-ASE algorithm to the dataset, and evaluated the threshold ϵ values between 0.4 and 0.7. By analyzing the correlations found, we can observe some interesting relationships between regions. It can be noted that groups of correlated attributes (Correlation Group), the relevant attributes in each group (Base) and the set of relevant attributes considering the whole dataset (Correlation Base) are similar for different region. Table 1 presents the Correlation Base generated for each region evaluated.

Table 1. Results of FD-ASE execution

Region	CB
Araraquara	$CB = \{a_4, a_2, a_1, a_5\}$
Araras	$CB = \{a_4, a_2, a_5\}$
Jaboticabal	$CB = \{a_2, a_4\}$
Jardinopolis	$CB = \{a_2, a_4, a_1\}$
Jau	$CB = \{a_2, a_4\}$
Luis Antonio	$CB = \{a_4, a_2, a_1, a_5\}$
Pitangueiras	$CB = \{a_2, a_4, a_1\}$
Pontal	$CB = \{a_2, a_1\}$
Ribeirao Preto	$CB = \{a_4, a_5, a_3\}$
Sertaozinho	$CB = \{a_2, a_4, a_5\}$

Furthermore, by using the method of fractal correlation, we discovered the existence of correlation between NDVI and precipitation, which is not identified when employing the Pearson correlation. It is worth to mention that the Pearson correlation is the technique usually employed by the agrometheorologists to find correlations among data, but 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 between more than two attributes, which is an advantage if compared to other methods from the literature, such as the well-known Pearson correlation.

However, during the test session specialists showed trouble to interpret the output of FD-ASE. Then, we proposed changes to incorporate more semantic meanings. 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 group $G_1 = \{a_1, a_2, a_4, a_5\}$ with base $B_1 = \{a_1, a_2\}$, we propose to present the subgroups of attributes correlated as $SubG_1 = \{a_1, a_2, a_4\}$ and $SubG_2 = \{a_1, a_2, a_5\}$.

Another change being included is an attribute weighting method. The specialists are used to calculate Pearson's correlation that presents a value r to indicate the intensity of linear correlation between two attributes. Then, we are developing the necessary extensions to map the degree of correlation in this interval, making it easier the FD-ASE interpretation by the specialists.

4.2. Method to measure the intrinsic dimension of data streams

We used SID-Meter to monitor long time series and identify changing in data distribution over time. One of the experiments was performed with a climate dataset (*ClimateCps*) from a region in São Paulo state. *ClimateCps* has three attributes, being the value of daily minimum (t_{min}) and maximum (t_{max}) temperatures (° Celsius), and the amount ($rain$) of rain (mm) measured for a period of 114 years in the city of Campinas. To calculate the intrinsic dimension of *ClimateCps* over time we used 3 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 graph of Figure 4 shows the values of the intrinsic dimension over time for the climate data from Campinas.

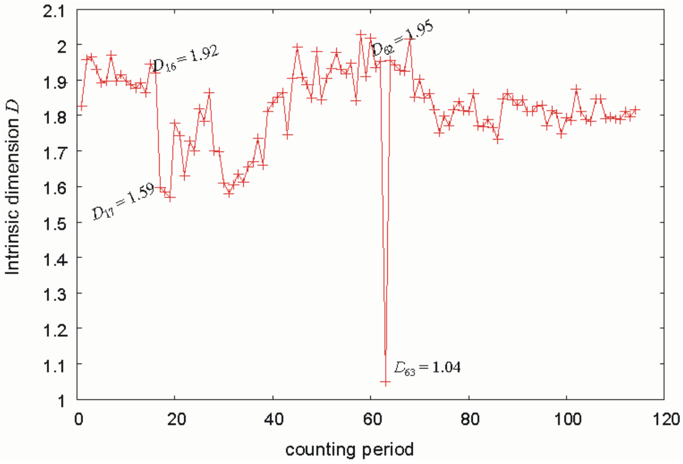


Figure 4. Monitoring process - ClimateCps dataset. It was considered a period of 114 years, where D is updated every year.

As it can be seen in Figure 4, there are three different patterns in the dataset. From the period 17 (1906) until 45 (1933), there is a meaningful variation in the intrinsic dimension. In the next period (45 to 75), the values of the intrinsic dimension are close to 2. However, period $p = 63$ (1951) indicates the highest variation in the intrinsic dimension. The large difference between D_{63} and D_{62} associated to the fact the attribute $rain$ remains stable indicates a meaningful change in climate conditions. Such variation is acknowledged by the specialists.

In the end of the stream, the intrinsic dimension varies less than in the others periods, showing a certain stability in the data distribution. According to the meteorology researchers team, these three patterns indicate a variation in distribution of the rain and an increasing in the minimum temperature in the last years.

We are improving the SID-Meter technique to automatically identify window sizes whenever there are interesting patterns. Therefore the specialists can potentially have a new and powerful tool helping in the process of decision making.

5. Conclusions and Future Perspectives

This work presented a new framework to analyze remote sensing images associated to climate data. We described a set of methods to filter, monitor, mine, analyze and visualize these type of time series. Experiments were performed on datasets of the Sao Paulo state, because it is the main producer of sugar cane in Brazil, and has a strategic relevance to the country economy. Sugar cane is used to produce biofuel and it is the main alternative to replace fossil fuel.

Knowing how the attributes extracted from the data are correlated to each other helps the specialists during the analysis of the data gathered. Furthermore, since the amount of data generated by the satellites 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.

The proposed framework allows analysis, monitoring, mining and visualization of a huge amount of data, which contribute the work of holding the data done by the agrometeorologists. This project is related to the proposed Challenges in Computer Science Research in Brazil, depicted by the Brazilian Computer Society, especially challenge number 2 *The computational modeling of natural complex systems and their interaction between the human beings and the nature*. Our intent with this paper is to present the steps taken until now towards dealing with the motivating problem of analyzing large amounts of agrometeorological, climate and remote sensing data. Our perspective is that the complete framework can provide mechanisms to assist meteorologists in detecting climate changes, keeping up with these changes and forecasting their consequences to the agriculture.

6. Acknowledgments

We thank the Brazilian funding agencies FAPESP and CNPq by financial support, the CIIAGRO/IAC and Agritempo by climate data, and CEPAGRI/Unicamp by the remote sensing images.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *VLDB*, pages 487–499.

Baioco, G. B., Traina, A. J. M., and Traina, Caetano, J. (2007). Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *SSDBM 2007*, pages 6–16, Banff, Canada. ACM Press.

Barbará, D. and Chen, P. (2000). Using the fractal dimension to cluster datasets. In *ACM SIGKDD*, pages 260–264, Boston, MA.

Chakrabarti, D. and Faloutsos, C. (2002). F4: large-scale automated forecasting using fractals. In *CIKM*, volume 1, pages 2–9, McLean, VA - EUA. ACM Press.

Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *SIGKDD*, pages 493–498.

Emery, W. J., Brown, J., and Novak, Z. P. (1989). Avhrr image navigation: summary and review. *Photogrammetric Engineering and Remote Sensing.*, 55(8):1175–1183.

Esquerdo, J. C. D. M., Antunes, J. F. G., Baldwin, D. G., Emery, W. J., and Zullo Jr, J. (2006). An automatic system for avhrr land surface product generation. *IJRS*, 27(18):3925–3942.

Faloutsos, C. and Kamel, I. (1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *ACM PODS*, pages 4–13, Minneapolis, MN.

Harms, S. K. and Deogun, J. S. (2004). Sequential association rule mining with time lags. *JIIS*, 22(1):7–22.

Holben, B. N. (1986). Characteristics of maximum value composite images from temporal avhrr data. *IJRS*, 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*, 46(5):651–656.

Honda, R. and Konishi, O. (2001). Temporal rule discovery for time-series satellite images and integration with rdb. In *PKDD*, pages 204–215, Freiburg, Germany. Springer-Verlag.

IPCC (2007). Intergovernmental panel on climate change. `http://www.ipcc.ch/ipccreports/index.htm`. accessed: March, 2009.

Julea, A., Méger, N., and Trouvé, E. (2006). Sequential patterns extraction in multitemporal satellite images. In *PKDD*, pages 96–99, Berlin, Germany. Springer-Verlag.

Keim, D. A., Mansmann, F., Schneidewind, J., and Ziegler, H. (2006). Challenges in visual data analysis. In *IV '06*, pages 9–16.

Keogh, E. and Kasetty, S. (2002). On the need for time series data mining benchmarks: A survey and empirical demonstration. In *SIGKDD*, pages 102–111.

Pagel, B.-U., Korn, F., and Faloutsos, C. (2000). Deflating the dimensionality curse using multiple fractal dimensions. In *ICDE*, pages 589–598, San Diego, CA. IEEE Computer Society.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii regression, heredity and panmixia. *Philos Trans Royal Soc London Ser A*, 187:253–318.

Ribeiro, M. X., Traina, A. J. M., and Jr., C. T. (2008). A new algorithm for data discretization and feature selection. In *ACM SAC*, pages 953–954, Fortaleza, Ceara, Brazil.

Rodrigues Jr., J. F., Traina Jr., C., and Traina, A. J. M. (2008). Metricssplat - the metric space platform. `http://gbdi.icmc.usp.br/~junio/MetricSPlat/`. accessed: March, 2009.

Romani, L. A. S., Sousa, E. P. M., Ribeiro, M. X., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2009a). Employing fractal dimension to analyze climate and remote sensing data streams. In *SIAM SDM-MDM*, pages 1–12, Sparks Nevada, USA.

Romani, L. A. S., Sousa, E. P. M., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. (2009b). Aplicação de método baseado em fractais para detecção de correlações entre imagens avhrr-noaa e dados agroclimáticos em regiões produtoras de cana-de-açúcar. In *SBSR*, Natal, Brasil.

Rosborough, G. W.; Baldwin, D. G. E. W. J. (1994). Precise avhrr image navigation. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):644–657.

Rosseti, L. (2001). Zoneamento agrícola em aplicações de crédito e securidade rural no brasil: Aspectos atuariais e de política agrícola. *RBAgro*, 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 *Earth Resources TechnologySatellite*, volume 1 of *NASA SP-351*, pages 309–317, Washington, D. C. NASA. Goddart Space Flight Center.

Schroeder, M. (1991). *Fractals, Chaos, Power Laws*. W. H. Freeman, New York, 6 edition.

Sousa, E. P. M. d., Traina, Caetano, J., Traina, A. J. M., and Faloutsos, C. (2007a). Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing Journal*, 25:33–59.

Sousa, E. P. M. d., Traina, Caetano, J., Traina, A. J. M., Wu, L., and Faloutsos, C. (2007b). A fast and effective method to find correlations among attributes in databases. *DMKD*, 14(3):367 – 407.

Thomas, J. and Cook, K. (2005). Illuminating the path: Research and development agenda for visual analytics. In *IEEE-Press*.

Thornthwaite, C. W. and Mather, J. R. (1955). The water balance. *Climatology*, 8(1):104.

Traina Jr., C., Sousa, E. P. M. d., 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*, volume 1, pages 599–630 (Chapter 24). Wiley/IEEE Press.

Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2000). Fast feature selection using fractal dimension. In *SBBD*, pages 158–171, João Pessoa, PB.

Wu, T., Song, G., X., M., X., G., and Jin, X. (2008). Mining geographic episode association patterns of abnormal events in global earth science data. *Science in China*, 51:155–164.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1-2):31–60.