Mining Relevant and Extreme Patterns on Climate Time Series with CLIPSMiner

Luciana A. S. Romani[1,2], Ana Maria H. Ávila[3], Jurandir Zullo Jr.[3],
Caetano Traina Jr.[1], Agma J. M. Traina[1]

[1] Department of Computer Science, University of São Paulo, São Carlos - SP, Brazil
{alvim, caetano, agma}@icmc.usp.br
[2] Embrapa Agriculture Informatics, Campinas - SP, Brazil
[3] Cepagri - Unicamp, Campinas - SP, Brazil
{avila, jurandir}@cpa.unicamp.br

Abstract. One of the most important challenges for the researchers in the 21st Century is related to global heating and climate change that can have as consequence the intensification of natural hazards. Another problem of changes in the Earth's climate is its impact in the agriculture production. In this scenario, application of statistical models as well as development of new methods become very important to aid in the analyses of climate from ground-based stations and outputs of forecasting models. Additionally, remote sensing images have been used to improve the monitoring of crop yields. In this context we propose a new technique to identify extreme values in climate time series and to correlate climate and remote sensing data in order to improve agricultural monitoring. Accordingly, this paper presents a new unsupervised algorithm, called CLIPSMiner (CLImate PatternS Miner) that works on multiple time series of continuous data, identifying relevant patterns or extreme ones according to a relevance factor, which can be tuned by the user. Results show that CLIPSMiner detects, as expected, patterns that are known in climatology, indicating the correctness and feasibility of the proposed algorithm. Moreover, patterns detected using the highest relevance factor is coincident with extreme phenomena. Furthermore, series correlations detected by the algorithm show a relation between agroclimatic and vegetation indices, which confirms the agrometeorologists' expectations.

1. INTRODUCTION

In the last decades, researches in Climatology have 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, the results of several analyses have showed that some extreme weather events have changed in frequency, duration and intensity over the last years [Meehl and Tebaldi 2004; Vincent et al. 2005; Groisman et al. 2005; Goswami et al. 2006; Alexander et al. 2006; Ganguly and Steinhaeuser 2008]. Consequently, increased temperatures and regional changes in precipitation patterns can have adverse effects on natural and human systems.

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 indices derived from daily temperature and precipitation data in order to organize and

---

allow comparisons among works around the World.

Extreme precipitation events such as strong daily rainfall and many days with rainfall above the daily average can cause flooding, 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.

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. Developing countries such as Brazil usually do not have an ideal conception of monitoring network. The use of remote sensing data is an alternative to the more conventional methods, because the remote sensors have an excellent spatial and temporal coverage. These remote sensors also enable the achievement of continuous information about the country, with spatial resolution of few kilometers and temporal order of minutes. However, measurements obtained from remote sensors are indirect and, therefore, it is necessary to develop models that relate features available in the satellite spectral channels to parameters associated with this information. In this scenario, several satellites are being used to assist in monitoring and forecasting.

Climate data from ground-based stations, remote sensors, weather radars or sensor network have been increased, yielding terabytes of data every week. In addition, climate change models have been processed for different scenarios generating huge amounts of data. Consequently, experts have much more effort to analyze and 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 deeply endeavor.

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 drought (days with rainfall bellow 10 mm or without rainfall), day with extreme rainfall, periods of drought in the Winter associated to high temperatures and other phenomena. In most cases, these events are similar to positive and negative peaks. Most of these studies are done using time series that can be correlated with other data sources such as remote sensors aboard satellites. Another way of studying the climate is to correlate time series from ground-based meteorological stations with values of sea surface temperature.

Knowledge discovery in database is an important approach that can be used to answer several agrometeorologists' questions. Thus, analyzing large amounts of climate data from real measures or model outputs associated to remote sensing data and geographical information is an important reason to develop new data mining algorithms. In this context, we consider the problem of finding relevant patterns and extreme phenomena from heterogeneous time series. In general, 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 character sequence without lost information about data range.

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. In this work, 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?

In order to address these problems, we propose a new unsupervised algorithm (CLIPSMiner) to discover relevant and extreme patterns in heterogeneous climate and remote sensing time series. This new method works on multiple time series of continuous data, identifying all 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 (years, for example). Thus, meteorologists can examine and compare patterns in each period, considering its tendency, i.e., whether there was an increase/decrease on the number of patterns and if the maximum and minimum values in each pattern varies between periods.

This paper is organized as follows. Section 2 presents the main related work. Section 3 formalizes the necessary concepts related to the problem of pattern mining and presents CLIPSMiner, an efficient algorithm for mining patterns from climate time series. Section 4 presents and analyzes experimental results obtained by executing CLIPSMiner over synthetic data as well as over real data. Finally, section 5 summarizes the conclusions and discusses future research.

2. RELATED WORK

Ganguly in [Ganguly and Steinhaeuser 2008] presented a work that maps climate requirements to solutions available in temporal, spatial and spatio-temporal data mining. They introduce climate challenges to the data mining community and suggest that relatively simple data mining methods can result in scientific insights with social impacts, such as analyses of heat waves [Meehl and Tebaldi 2004] or detection of extreme rain events [Goswami et al. 2006]. These studies use well-known statistical techniques, which may not be as efficient as the volume of data increases, because it demands several scans over the data.

Understanding spatio-temporal variability of precipitation extremes is an important task since precipitation extremes are related to natural disasters, such as flooding. Extreme Value Theory (EVT) is a statistical area that deals with extreme deviations from the median of probability distributions. Wu in [Wu and Chawla 2007] have analyzed precipitation extremes over El Niño years using spatial autocorrelation of Generalized Pareto distribution parameters from EVT. Spatio-temporal variability of weekly precipitation extremes in South America also has been investigated [Khan et al. 2007; Kuhn et al. 2007]. These papers analyze extreme phenomena in rainfall series using mature statistical techniques. However, such methods do not automatically detect the context and period in which extreme events occur, as the new proposed method in this paper does.

Methods of data mining are more suitable to find patterns in massive databases and have been proposed in the literature for discovering sequential patterns [Zaki 2001; Wang and Han 2004; Cao et al. 2005; Huang et al. 2006], mining association rules from them [Agrawal et al. 1993; Das et al. 1998] and clustering [Steinbach et al. 2003]. Mannila in [Mannila et al. 1997] also proposed a method to episodical sequential data mining that uses all frequent episodes within one sequence.

There are several methods that use constraints to focus on the mining process to find relevant items. Zaki in [Zaki 2000] proposed the use of temporal constraints in transactional sequences. Harms in [Harms et al. 2001] defined methods that combine constraints and closure principles with a sliding window approach. Their objective was to find frequent closed episodes in multiple event sequence. They developed the MOWCATL algorithm [Harms and Deogun 2004] to mine frequent association rules from sequential datasets. Wu [Wu et al. 2008] proposed the GEAM (*Geographic Episode Association Pattern Mining*) algorithm to find association patterns in abnormal event sequences.

Lin in [Lin et al. 2007] have applied data mining techniques to a global land precipitation dataset and a Sea Surface Temperature (SST) dataset. They have discovered well-known teleconnection patterns and previously unknown ones, such as an abnormally low sea surface temperature (SST) in the Eastern Pacific that coincides with abnormally high precipitation in Shanxi.

This paper proposes a new way to analyze heterogeneous time series in order to discover relevant or extreme patterns in complete time series or their parts divided into time windows. Moreover, we propose a method to verify correlations between time series.

3. PROBLEM FORMALIZATION AND THE CLIPSMINER ALGORITHM

The CLIPSMiner algorithm detects sequential patterns on climate and remote sensing time series that represent meaningful phenomena, such as periods with a small variation in data distribution and negative or positive peaks. The size of these peaks may represent extreme phenomena, such as meaningful drops in temperature or heavy rains in a short period of time. The CLIPSMiner algorithm works on multiple and continuous time series.

3.1   Problem definition

The following definitions are fundamental to better understand the proposed method. Table I summarizes symbols and mnemonics employed in this paper.

Definition 3.1. A *Time series S* is defined as a sequence of pairs (a_i, t_i) with $i = 1, \ldots, n$, i.e. $S = [(a_1, t_1), \ldots, (a_i, t_i), \ldots, (a_n, t_n)]$, such that $(t_1 < \ldots < t_i < \ldots < t_n)$, where each a_i is a data value and each t_i is a time value in which a_i occurs.

Each pair (a, t) is called an *event e*. A set of events E contains n events of type (a_i, t_i) for $i = 1, \ldots, n$. Each a_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 and R, the values t_i of both must be measured in the same time unit.

Definition 3.2. 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 $e_i = (a_i, t_i)$ 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 $d_i = (a_{i+1} - a_i)$ and a given δ parameter. The extracted event sequences comprise a period of events having the tendency to rise or fall, when plotted as a graph.

The value of δ is usually very small, tending to zero $(\delta \to 0)$. Therefore, we define three exclusive types of event sequences.

Definition 3.3. 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 3.4. The *descending event sequence* 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 3.5. The *stable event sequence* 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 in a given location. Thus, we define three types of patterns used to quantize a time series S.

Definition 3.6. *Valley pattern (V)* is defined as the concatenation of a descending event sequence and an ascending event sequence, i.e. $V \Rightarrow S_{ed}S_{ea}$.

Definition 3.7. *Plateau pattern (P)* is defined as a stable event sequence, i.e. $P \Rightarrow S_{es}$.

Definition 3.8. *Mountain pattern (M)* is defined as the concatenation of an ascending event sequence and a descending event sequence, i.e. $M \Rightarrow S_{ea}S_{ed}$.

Figure 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. Figure 1(b) presents an interval in a time series and highlights a pattern P. For WRSI (Water Requirement Satisfaction Index) time series, a pattern P can occur when a_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. Finally, Figure 1(c) presents an interval in a time series and highlights a pattern M. In a real dataset, pattern M occurs when there is a significant variation in the amplitude, such as a very heavy rain, for example.

Two thresholds (ρ and λ) are defined to identify only the relevant patterns, acting as filters. The threshold ρ 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 λ is the plateau length and is defined to identify relevant P patterns. Both thresholds get a default value, what can be tuned by the user.

Table I.   Table of symbols and mnemonics

| Symbols	Definition
S, R	Time series
a_i	Data value
t_i	Time value
e_i	Events of type (a_i, t_i)
d_i	Difference of events
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
δ	Minimum variation between two consecutive events
ρ	Relevance Factor
λ	Plateau Length
τ	Time delay
E	Set of events
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
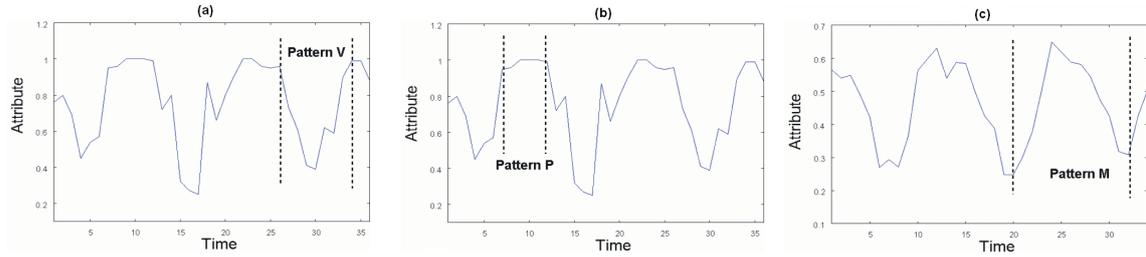
Fig. 1. Examples of patterns detected by CLIPSMiner are presented in graph format where the y axis represents attribute value and time is measured in x axis. (a) Pattern of type V is similar to negative peaks. (b) Pattern of type P implies an interval in the time series with low variation. (c) Pattern of type M is equivalent to a positive peak.

Definition 3.9. *Amplitude (y)* is defined as the difference between the maximum and the minimum values of the time series, $y = a_{max} - a_{min}$.

Definition 3.10. *Relevance factor (ρ)* 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 3.11. *Plateau length (λ)* defines the length of an stable event sequence (S_{es}).

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. In this work, the default value for the parameter was defined empirically and corresponds to $\rho = y * 40/100$. The default value of λ is 4 to allow discovering plateaus composed of four consecutive events, which was also defined empirically.

Definition 3.12. *Time delay* τ 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.

3.2 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. Figure 2 presents a schema of the algorithm steps using a time series of 16 values as example. In the first step, it generates an array containing the differences between the previous and current values of the series. Thus, all event sequences are identified (2nd step) and the patterns M, V and P are found in the 3rd step . According to parameters, some patterns are pruned and only relevant patterns are presented as result. Additionally, CLIPSMiner can divide time series in parts of a time period (usually in years) defined by the user. Thus, it is able to discover different patterns in each part of the series, observing trends and changes in different time periods.

Algorithm 1 summarizes the CLIPSMiner algorithm. For each time period (p) of time series, CLIPSMiner first calculates an array composed by the differences between previous and current values, i.e. $d_i = a_{i+1} - a_i$ (lines 2 to 4). Thus, by analyzing the array of d_i, it can discover 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
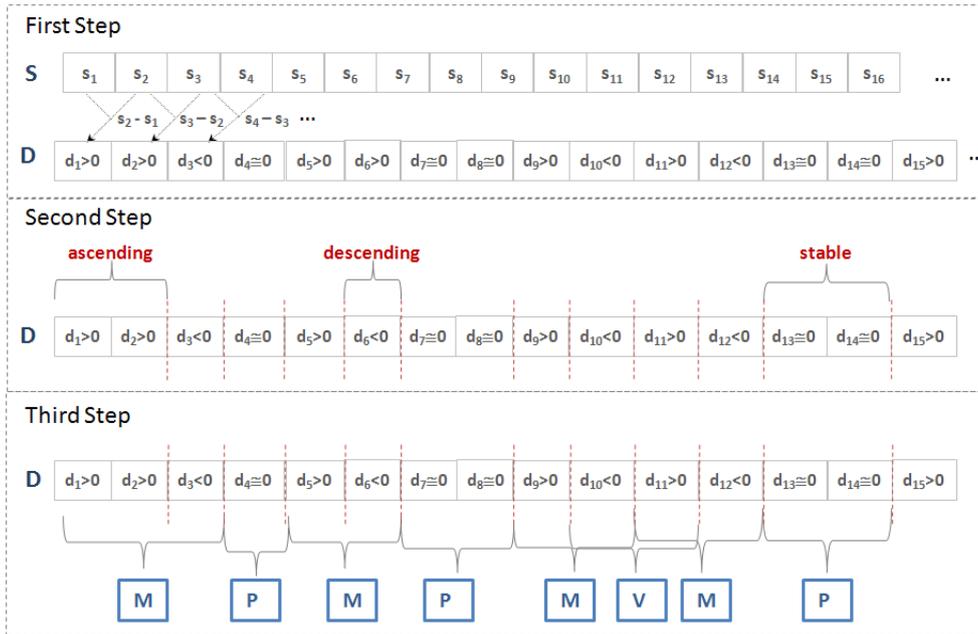
Fig. 2. Representation of the three steps executed by algorithm CLIPSMiner. (1) Calculation of differences between previous and current values of time series, (2) Identification of ascending, descending and stable event sequences and (3) Detection of patterns M, V and P.

according (lines 5 to 9). In the next step, CLIPSMiner prunes event sequences S_{ea} and S_{ed} smaller than ρ, and S_{es} smaller than λ (lines 10 and 11).

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). CLIPSMiner stores the mined patterns in an array for each time series S. The format of the patterns is an event interval, such as $[a_{init}, a_{mid}, a_{end}]$, where mid is an intermediate value, and the time interval $[t_{init}, t_{end}]$ where the event e occurs (line 17).

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 τ is the mean value found.

3.3 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 $O(n)$.

4. EXPERIMENTAL RESULTS

In this section, we discuss representative experiments performed on synthetic and real datasets. 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.

Algorithm 1 CLIPSMiner Algorithm

Input: Time series S; thresholds δ, ρ, λ and p
Output: Patterns V, M, P and time delay τ
    for each time period p of time series S **do**
2:    **for all** a_i **do**
        calculate array of differences $d_i = a_{i+1} - a_i$
4:    **end for**
    for all d_i values **do**
6:        Find S_{ea} = Set of ascending event sequences
        Find S_{ed} = Set of descending event sequences
8:        Find S_{es} = Set of stable event sequences
    end for
10:    Prune S_{ea} and S_{ed} when $\sum d_i < \rho$
    Prune S_{es} when $\sum d_i < \lambda$
12:    **for all** S_e not pruned **do**
        V = concatenation of $S_{ed}S_{ea}$
14:        M = concatenation of $S_{ea}S_{ed}$
        $P = S_{es}$
16:    **end for**
    Set of all patterns as $[a_{init}, a_i, a_{end}](t_{init}, t_{end})$
18:    **for all** Patterns V, M, P **do**
        write e_{init}, e_i, e_{end}
20:    **end for**
    for each pair of patterns array **do**
22:        calculate time delay between patterns in different array
        write time delay τ
24:    **end for**
    end for

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 climatology researchers to analyze climate data: percentile and cross-correlation. Percentile is a measurement of the relative position of one value regarding all other values. The pth 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.

The calculation of time lag (τ) made by our algorithm was compared with results presented by the cross-correlation method. This technique identifies the correlation between two time series, identifying how much one series must be shifted along the x-axis to make it identical to other one.

A qualitative analysis is also presented and results are analyzed for different values of the relevance factor ρ and length of plateau λ patterns.

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 100,000 events and 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, respectively. The Cps dataset is composed by 41,700 events and three attributes: a daily rain value ($rain$), maximum (t_{max}) and minimum (t_{min}) temperatures measured over a period of 118 years at Campinas in Brazil.

The dataset FiveRegions is composed of values extracted from NDVI (Normalized Difference Vegetation Index) images and WRSI values. The NDVI index is calculated through the ratio between channels 1 and 2 of the Advanced Very High Resolution Radiometer (AVHRR) sensor on board the National Oceanic and Atmospheric Administration (NOAA) satellites. The NDVI index is closely correlated with leaf area, green biomass and productivity indexes. This dataset has seven years of sugar cane harvest for five regions. It is composed of monthly values of NDVI, which was obtained using the MVC (Maximum Value Composition) technique. This technique makes composition of the NDVI maximum value, obtained from a series of multi-temporal georeferenced images. This method is applied to minimize the atmospheric effects in NOAA-AVHRR images.

The agroclimatic conditions through the period of analysis are described by WRSI, which is the ratio between the real evapotranspiration and maximum evapotranspiration. Evapotranspiration is the sum of evaporation and plant transpiration. WRSI index varies from zero to one and represents a fraction of water consumed by plants from the total amount of water that would be used by plants to ensure maximum productivity.

Multitemporal remote sensing images have been widely used in research works in recent years. In these datasets, each tuple corresponds to one NDVI measurement per month, for a period of 7 years. There are months in the year, such as January, February and March, where there is not good images to analyze due to clouds coverage in Brazil. Then, this database has 500 events.

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 3 shows the performance of the CLIPSMiner algorithm considering two aspects: number of events found and variation of the relevance factor. The graph of Figure 3(a) shows the execution time by 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 3(b). Similarly, the number of patterns decreases as the relevance factor increases, as expected (Figure 3(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 ρ and λ factors increase.

When parameters were set to $\rho = y * 70\%$ and $\lambda = 8$, the detected patterns of type P, M and V corresponds to relevant patterns with a meaningful variation in the amplitude. For example, Figure 4(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 4a. 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. Usually, 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.
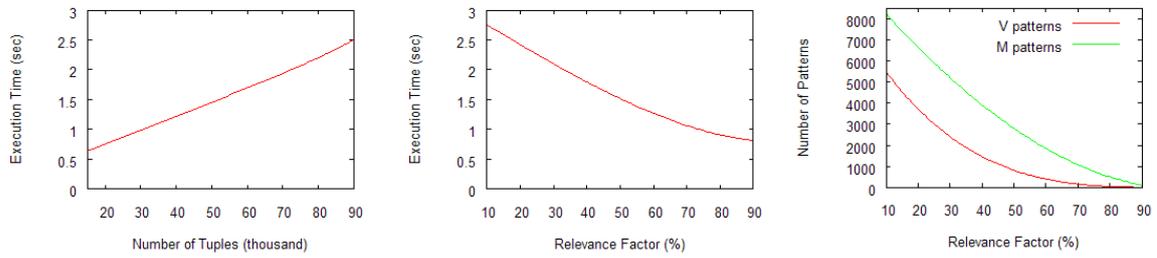
Fig. 3. Performance of CLIPSMiner algorithm considering the number of patterns found and variation of the relevance factor: (a) execution time by number of tuples (b) execution time by relevance factor (c) number of patterns by relevance factor.
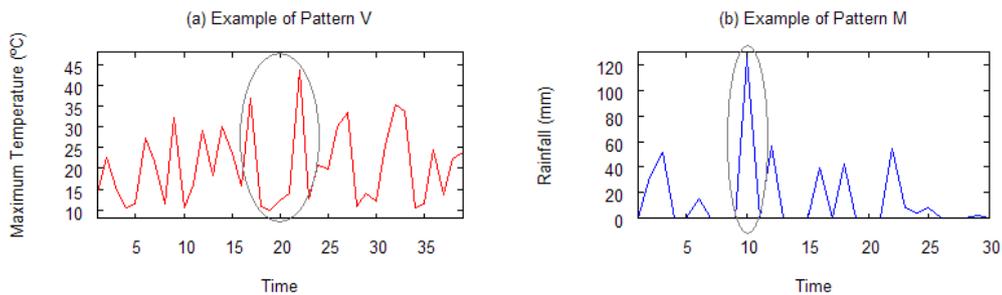

Fig. 4. Example of extreme patterns: (a) V pattern similar to a negative peak (period 17 to 22) and (b) M pattern similar to a positive peak (9 to 11).

When we increase the value of the relevance factor (ρ) to 95% of amplitude, CLIPSMiner becomes more restrictive and only very extreme patterns are detected. Figure 4(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 time series. In real databases, this pattern is similar to an increase in the amount of rainfall, for example.

Many days without rain especially in 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 λ (plateau lenght) value. Increasing the λ value allows to find the most extreme phenomena. In the time series a_3, for example, when we increased λ, 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 the analysts needs is a big difference of the algorithm CLIPSMiner, when comparing 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 value of extreme daily rain for every 50 year-period, and know if occured changes in the time series trend. Any period of time could be set.

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 find extremes in a efficient way, but does not provide information on the period where the extreme event occured, neither about preceding and subsequent events, what CLIPSMiner does.

The CLIPSMiner algorithm outputs show that in the majority of the extreme patterns occurred 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 one 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 τ 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.

4.4 Results on Real Data - The Cps dataset

The same process described for the Synth dataset was employed. Two experiments were executed, respectively, assigning the default and maximum values for parameters ρ and λ, in order to find the relevant and the extreme patterns. For both experiments, the δ value was set to 0.9. Figures 5(a) and (b) show the number of patterns discovered for the three time series of the Cps database 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 $\lambda = 20$
—$rain$: $\rho = y * 70\%$ and $\lambda = 30$

The parameters are different because the range 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 5, we can observe a meaningful decrease of patterns when ρ and λ values increase, i.e. they become more restrictive. Patterns of type V in time series t_{max} drops from 11 to 3. This pattern represents variations in maximum temperature in different periods of time. Table II shows the V patterns found in time series t_{max} using $\rho = y * 55\%$.

As it can be seen in Table II, 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 comes 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.

Table III presents 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 flooding. Researchers are interested 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.

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

Table II. 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]
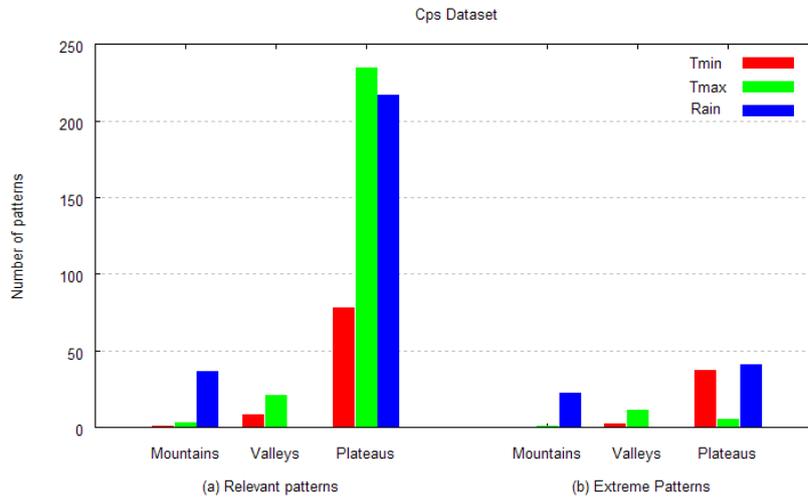
Fig. 5. Results for Cps dataset: y-axis represents the number of patterns and the type of patterns are represented in x-axis. (a) number of relevant patterns and (b) quantity of extreme patterns.

Table III. M patterns found for $rain$ in Cps dataset

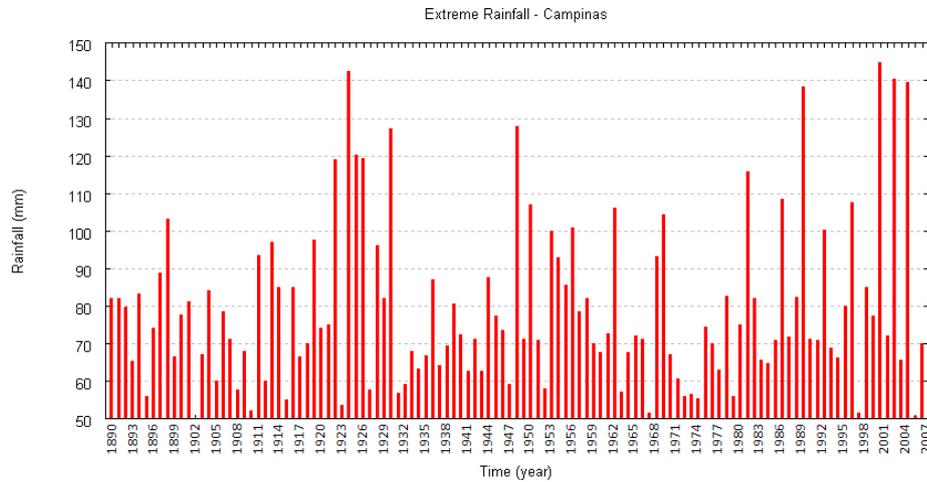| # patterns	$rain$ values	date
1:	[0.0; 103.0; 0.0]	[24/01/1899-27/01/1899]
2:	[0.0; 119.0; 0.0]	[11/13/1923-11/15/1923]
3:	[0.0; 142.4; 0.0]	[12/23/1925-12/26/1925]
4:	[0.0; 127.7; 1.6]	[12/23/1949-12/25/1949]
5:	[0.0; 107.0; 0.0]	[11/23/1951-11/27/1951]
6:	[0.0; 115.7; 0.0]	[01/01/1982-01/04/1982]
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]

Fig. 6. Graph with extreme rainfall per year in the Campinas region with 17 values above 100 mm.

illustrated in Figure 6. This fact confirms the hypothesis of the researchers that the distribution of rainfall has increased in a fast pace, in the last decades.

Many P patterns were found in all time series, specially for $rain$. This dataset is composed by 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 δ and λ, CLIPSMiner discovered prolonged droughts, that is a pattern studied by experts, because of their consequences for agriculture. Mainly, if 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 statistical techniques for detection of extended dry periods.

Setting the parameter p, CLIPSMiner divides the time series into p pieces and discover patterns in accordance with the relevance factor ρ for each period p_i. Thus, it is possible to analyze the trend of the series in each period separately.

In this experiment, we set p to 10 years and ρ 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 $130mm$, as it can be seen in Figure 7.

To compare this result with the output generated by the percentile method, similarly we defined periods of 10 years. We set Percentile algorithm to 99th to find extremes. This method detected only a value that represents extremes in each period of 10 years. 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 not 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, what CLIPSMiner does.

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 set to be monthly. Figures 8(a) and (b) summarizes the patterns detected when the parameters were set to be smaller ($\rho = y * 10\%$ and $\lambda = 3$) and more sensitive ($\rho = y * 70\%$ and

(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]

Fig. 7. Extreme rain values for the beginning and the end of the time series: (a)rainfall reached values of 80 mm approximately in the beginning of the time series (1901 to 1905), (b) rainfall values increased to 130 mm in the end of the time series (1989 to 2005).
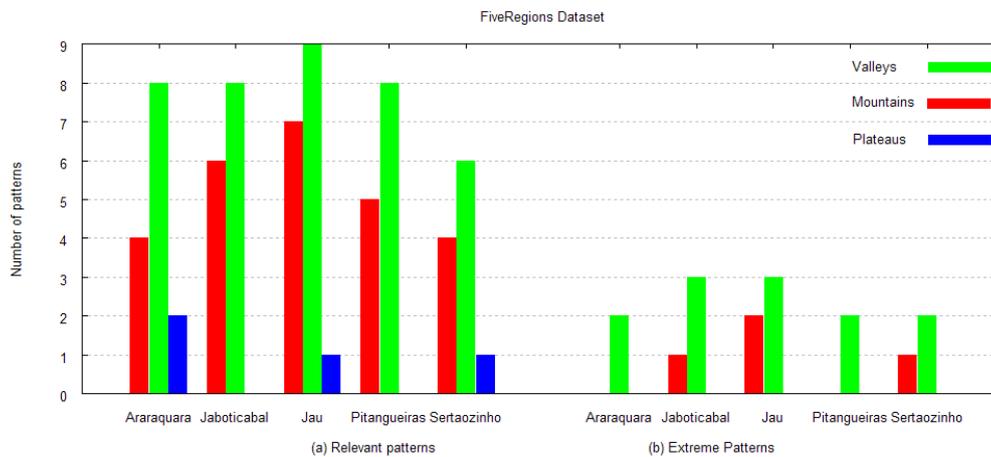

Fig. 8. Results for FiveRegion dataset: y-axis represents the number of patterns and the type of patterns are represented in x-axis. (a) number of relevant patterns and (b) quantity of extreme patterns.

$\lambda = 5$), respectively.

CLIPSMiner found few M patterns using default values for parameters in NDVI time series. The patterns detected were $[0.247385; 0.648107; 0.307657]$ in $[09/2003 - 10/2004]$ for Jaboticabal, $[0.296928; 0.615282; 0.237196]$ in $[10/2004 - 10/2005]$ and $[0.237196; 0.618585; 0.264748]$ in $[10/2005 - 10/2006]$ for Jau and $[0.264471; 0.611832; 0.269969]$ in $[10/2002 - 09/2003]$ for Sertaozinho. These M patterns are related to periods when the green biomass reaches its highest values, before the sugar cane harvest that begins in May. 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 series. This phenomenon occurs when the maximum soil water content is reached after a long period of rainfall.

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 9. CLIPSMiner found τ 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.
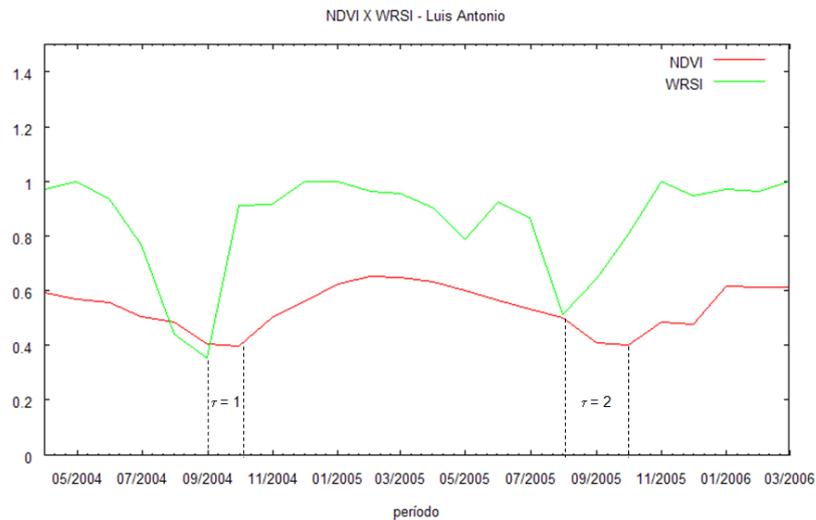
Fig. 9.   WRSI and NDVI time series from Luis Antonio with example of two different time lags ($\tau = 1$ and $\tau = 2$).

5. CONCLUSION

This work presented CLIPSMiner, a new unsupervised algorithm to find relevant and extreme patterns in climate and remote sensing time series, as well as correlations between time series, showing the relationship between the series and when one affects the other. Therefore, 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 certain 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 correlation between series, as well as provide a simple way to quantize 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 the 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 to model outputs in order to assist in climate change research.

6. ACKNOWLEDGMENTS

REFERENCES

Agrawal, R., Faloutsos, C., and Swami, A. Efficient similarity search in sequence databases. In *Proceedings of the International Conference on Foundations of Data Organization and Algorithms*. Chicago, USA, pp. 69–84, 1993.

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. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research* vol. 111, pp. 1–22, 2006.

Cao, H., Mamoulis, N., and Cheung, D. W. Mining frequent spatio-temporal sequential patterns. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. Houston, USA, pp. 8 pp., 2005.

Das, G., Lin, K., Mannila, H., Renganathan, G., and Smyth, P. Rule discovery from time series. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. New York, USA, pp. 16–22, 1998.

Ganguly, A. and Steinhaeuser, K. Data mining for climate change and impacts. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, IEEE (Ed.). IEEE, Pisa, Italy, pp. 385–394, 2008.

Goswami, B., Venugopal, V., Sengupta, D., Madhusoodanan, M., and Xavier, P. Increasing trend of extreme rain events over india in a warming environment. *Science* 314 (5804): 1442–1445, 2006.

Groisman, P., Knight, R., Easterling, D., Karl, T., Hegerl, G., and Razuvaev, V. N. Trends in intense precipitation in the climate record. *Journal of Climate* vol. 18, pp. 1326–1350, 2005.

Harms, S. K., Deogun, J., Saquer, J., and Tadesse, T. Discovering representative episodal association rules from event sequences using frequent closed episode sets and event constraints. In *Proceedings of the International Conference on Data Mining*. San Jose, USA, pp. 603–606, 2001.

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

Huang, Y., Zhang, L., and Zhang, P. Finding sequential patterns from a massive number of spatio-temporal events. In *Proceedings of the SIAM International Conference on Data Mining*. Maryland, USA, pp. 633–637, 2006.

Khan, S., Kuhn, G., Ganguly, A. R., Erickson III, D. J., and Ostrouchov, G. Spatio-temporal variability of daily and weekly precipitation extremes in south america. *Water Resources Research* vol. 43, pp. 1–25, 2007.

Kuhn, G., Khan, S., Ganguly, A. R., and Branstetter, M. L. Geospatial̆temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in south america. *Advances in Water Resources* vol. 30, pp. 2401–2423, 2007.

Lin, F., Jin, X., Hu, C., Gao, X., Xie, K., and Lei, X. Discovery of teleconnections using data mining technologies in global climate datasets. *Data Science Journal* vol. 6, pp. 749–755, 2007.

Mannila, H., Toivonem, H., and Verkamo, A. I. Discovery of frequent episodes in event sequences. *DMKD* vol. 1, pp. 259–289, 1997.

Meehl, G. and Tebaldi, C. More intense, more frequent, and longer lasting heat waves in the 21st century. *Science* 305 (5686): 994–997, 2004.

Steinbach, M., Tan, P., Kumar, V., Klooster, S., and Potter, C. Discovery of climate indices using clustering. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*. Washington, USA, pp. 446–455, 2003.

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. Observed trends in indices of daily temperature extremes in south america 1960-2000. *Journal of Climate* vol. 18, pp. 5011–5023, 2005.

Wang, J. and Han, J. Bide: efficient mining of frequent closed sequences. In *Proceedings of the International Conference on Data Engineering*. Boston, USA, pp. 79–90, 2004.

Wu, E. and Chawla, S. Spatio-temporal analysis of the relationship between south american precipitation extremes and the el niño southern oscillation. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, IEEE (Ed.). IEEE, Omaha, USA, pp. 685–692, 2007.

Wu, T., Song, G., Ma, X., Xie, K., Gao, X., and Jin, X. Mining geographic episode association patterns of abnormal events in global earth science data. *Science in China* vol. 51, pp. 155–164, 2008.

Zaki, M. Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the International Conference on Information and Knowledge Discovery*. Washington, USA, pp. 422–429, 2000.

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