

Finding Spatio-Temporal Patterns in Earth Science Data^{*}

Pang-Ning Tan⁺
Christopher Potter⁺⁺

Michael Steinbach⁺
Steven Klooster⁺⁺⁺

Vipin Kumar⁺
Alicia Torregrosa⁺⁺⁺

⁺ Department of Computer Science and Engineering, Army HPC Research Center
University of Minnesota
{ptan, steinbac, kumar@cs.umn.edu}

⁺⁺ NASA Ames Research Center
{cpotter@mail.arc.nasa.gov}

⁺⁺⁺ California State University, Monterey Bay
{klooster,atorregrosa@gaia.arc.nasa.gov}

Abstract

This paper presents preliminary work in using data mining techniques to find interesting spatio-temporal patterns from Earth Science data. The data consists of time series measurements for various Earth science and climate variables (e.g. soil moisture, temperature, and precipitation), along with additional data from existing ecosystem models (e.g. Net Primary Production). The ecological patterns of interest include associations, clusters, predictive models, and trends. In this paper, we discuss some of the challenges involved in preprocessing and analyzing the data, and also consider techniques for handling some of the spatio-temporal issues. Earth Science data has strong seasonal components that need to be removed prior to pattern analysis, as Earth scientists are primarily interested in patterns that represent deviations from normal seasonal variation such as anomalous climate events (e.g., El Nino) or trends (e.g., global warming). We compare several alternatives (including singular value decomposition (SVD), discrete Fourier transform (DFT), “monthly” Z score, and moving average) with respect to their effectiveness in removing seasonality. We describe the different kinds of association analysis that can be performed on such data. Our current technique for finding associations transforms the time series into transactions and then applies existing algorithms traditionally used for market-basket data. Some of the transformations lead to dense columns in the transaction matrices, causing an exponential growth in the computing requirements. Furthermore, no single interestingness measure accurately reflects the quality of the derived patterns. Indeed, we argue that existing approaches for mining association rules and sequential patterns may not be able to capture all the interesting patterns due to the spatio-temporal nature of this data.

1. Introduction

NASA’s Earth observation satellites are generating increasingly larger amounts of data. This remotely sensed data, combined with additional data from ecosystem models, offers an unprecedented opportunity for predicting and understanding the behavior of the Earth’s ecosystem. However, due to the large amount of data that is available, data mining techniques are needed to facilitate the automatic extraction and analysis of interesting patterns from the Earth Science data. This data consists of a sequence of global snapshots of the Earth (as shown in Figure 1), typically available at monthly intervals, and includes various atmospheric, land and ocean variables such as sea surface temperature (SST), pressure, precipitation and Net Primary Production (NPP). NPP is the net photosynthetic accumulation of carbon by plants. Keeping track of NPP is important because it includes the food source of humans and all other organisms and thus, sudden changes in the NPP of a region can have a direct impact on the regional ecology. An ecosystem model for predicting NPP, called CASA (the Carnegie Ames Stanford Approach [PKB99]), has been used for over a decade to produce a detailed view of terrestrial productivity. Our goal is to find interesting patterns involving events derived from the multi-year output of CASA, and other climate variables.

Mining patterns from Earth Science data is a difficult task due to the spatio-temporal nature of the data. In this paper, we discuss some of the challenges involved in preprocessing and analyzing the data, and also consider techniques for handling some of the spatio-temporal issues. First, we examine the problem of removing seasonal variation from the time series data. This is necessary because patterns derived from these variables are often dominated by the seasonal cycles present in the data. Earth Scientists are often interested in relating ecological events in a specific location to anomalous climate conditions that are occurring in a different part of the world. For

^{*} This work was partially supported by NASA grant # NCC 2 1231 and by Army High Performance Computing Research Center contract number DAAH04-95-C-0008. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by AHPCRC and the Minnesota Supercomputing Institute.

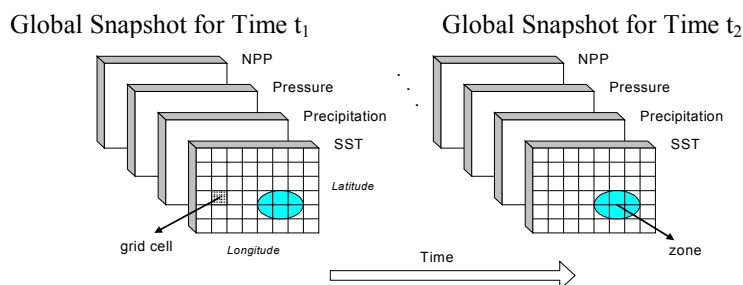


Figure 1: A simplified view of the problem domain.

example, during El-Nino years (i.e. the warming of the ocean surface for specific regions of the Pacific), it has been observed that the eastern part of Australia experiences severe drought conditions. Such anomalous events can become apparent only if the seasonal components of the time series are removed. Another reason for removing seasonal variations is to make the time series stationary, a typical assumption of many statistical time series analysis techniques (e.g., ARIMA). We also investigate the problem of detecting temporal auto-correlation and determining the statistical significance of various descriptive statistics, such as correlation, derived from the data.

Discovering spatio-temporal relationships among ecological events observed at various parts of the earth is critical for understanding how the different elements of the ecosystem interact with each other. A standard approach for finding such patterns is to compute the pair-wise correlation between time series of different geographical locations and then, finding regions that have high correlations (i.e., “similar” time series). This approach is described in more details in a related paper [Ste+01]. An alternative approach is to convert the time series into sequence of events and then apply existing data mining techniques to discover interesting associations in the event sequences. This approach has been studied by the data mining community in the context of *association rules* and *sequential pattern* discovery for market basket analysis [AS94, SA96, JKK99]. For the Earth Science data, we describe various ways to transform the original data into market-basket type transactions, so that existing algorithms can be applied.

The rest of the paper is organized as follows: Section 2 provides a description of the ecology data, while sections 3 and 4 present some of the temporal issues related to mining this kind of data, such as seasonality and temporal autocorrelation. Section 5 shows the results of association pattern discovery, while section 6 concludes with a discussion of future directions.

2. Ecology Data

The data for our analysis contains monthly measurements of various Earth science and climate variables over a period of twelve years, starting in January 1982. These variable values are either observations from different sensors, e.g. precipitation and sea surface temperature (SST), or the result of model predictions, e.g. NPP from the CASA model. In addition, Earth Scientists have developed standard indices (time series) that capture the behavior of various climate variables at a regional and global scale. For example, various El Nino related indices, such as ANOM1+2 and ANOM3.4, have been established to measure sea surface temperature anomalies across different regions of the ocean. Some of the well-known climate indices are shown in Table 1.

Climate Index	Description
SOI	measures the sea level pressure (SLP) anomalies between Darwin and Tahiti
NAO	normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
ANOM3	sea surface temperature anomalies in the region bounded by 90°W-150°W and 5°S- 5°N
ANOM3.4	sea surface temperature anomalies in the region bounded by 120°W-170°W and 5°S-5°N
NP	area-weighted sea level pressure over the region 30N-65N, 160E-140W

Table 1: Description of several well-known climate indices.

3. Dealing with the Seasonality of Data

Yearly patterns such as spring, summer, fall, and winter or rainy season / dry season are important, but well known. Thus, Earth scientists are primarily interested in patterns that represent deviations from the normal seasonal

variation. Examples of such patterns are special events (e.g., El Nino), long-term cycles (e.g., decadal oscillations), or trends (e.g., global warming). Given this focus on deviations from the norm, and the strength of the seasonal patterns in the data, it is necessary to remove them so that other, more interesting patterns can be detected. In the following we consider several transformations for removing seasonal variation: the discrete Fourier transform (DFT), the “monthly” Z score, singular value decomposition (SVD), and the moving average.

We illustrate some of the different possibilities and issues via an example centered around a typical SST (Sea Surface Temperature) time series shown in Figure 2. (This time series was derived from data corresponding to a $\frac{1}{2}^\circ$ by $\frac{1}{2}^\circ$ region of the ocean at 71.5° W, 23° S, just off the Eastern coast of South America.) In what follows, we shall often “standardize” a time series by subtracting its mean and dividing by its standard deviation. We do this to display multiple time series on a single plot without the distorting effects of scale. Also, because our measure of similarity in this domain is Pearson’s correlation coefficient, this sort of normalization seems very appropriate. Figure 3 shows the standardized version of our sample SST time series, which, not surprisingly, looks very similar to the original series in Figure 2.

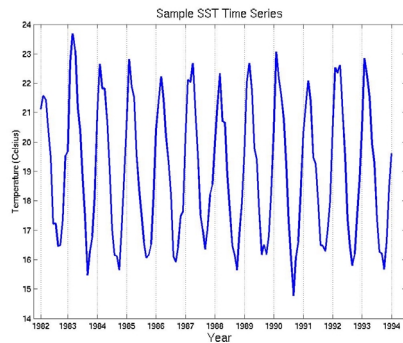


Figure 2: Sample SST time series

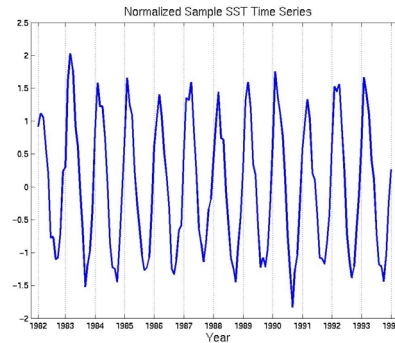


Figure 3: Standardized sample SST time series

Filtering based on the DFT (Discrete Fourier Transform). This approach is based on standard signal processing approaches. By taking the discrete Fourier transform, we can transform the original time series from the time domain to the frequency domain, where it is more readily apparent which frequencies make up the signal. In particular, the power spectrum of a time series can be readily calculated from the transformed series, as shown in Figure 4. (The constant component has been eliminated since otherwise it dominates the plot.) The peaks at 12 and 132 indicate that there is a strong yearly component. (The DFT and hence, the power spectrum, is symmetrical around $N/2$, where N is the length of the time series, and thus, there is just one strong frequency component, not two.) Removing this yearly component and then performing the inverse Fourier transform yields a new time series which should not have any seasonal component. (We also remove the constant component, since we are only interested in variations, not absolute levels.)

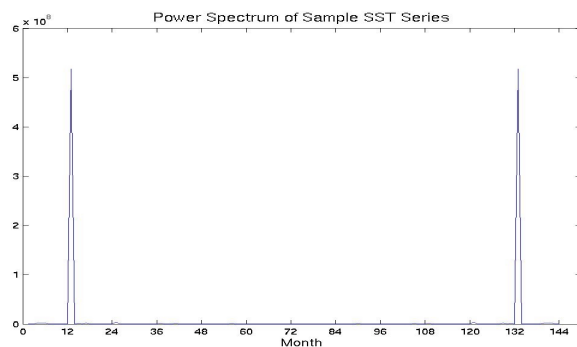


Figure 4: Power Spectrum of sample SST time series (constant component removed).

Monthly Z score. This transformation takes the set of values for a given month, e.g., all Januarys, calculates the mean and standard deviation for that set of monthly values, and then standardizes each value by calculating its Z score, i.e., by subtracting off the mean and dividing by the standard deviation. While this approach seems similar to the first approach, it is actually quite different since it uses the monthly mean and standard deviation instead of the overall mean and standard deviation. Put another way, we express each data value in the time series in terms of its deviation from the mean value for its corresponding month, scaled by the volatility factor for that month. The month-by-month rescaling used in this transformation causes seasonal fluctuations to disappear. Furthermore, scaling by the monthly standard deviation makes the changes more pronounced for those months in which the volatility is low (an issue that will be addressed at the end of this section).

Figure 5 shows the result of applying the monthly Z score and DFT filtering to the sample SST time series. These transforms produce almost identical results, and in fact, the correlation of the two transformed series is 0.98. While there are points in our data set for which the correlation between the monthly Z score and DFT filtered series is only 0.5, for most of our data this equivalence holds.

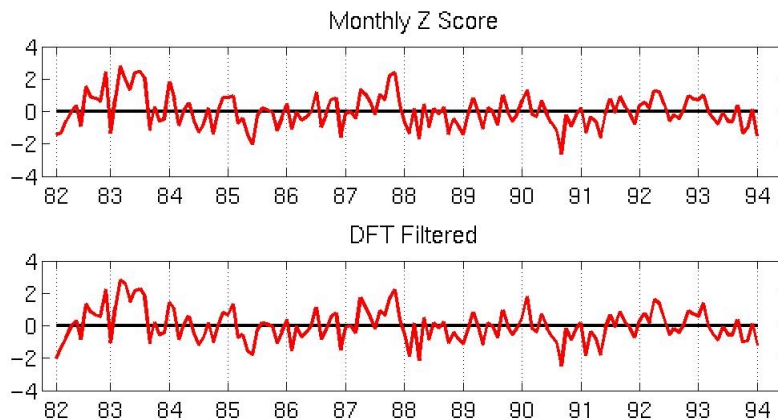


Figure 5: Results of applying monthly Z score and DFT filtering.

Singular value decomposition (SVD). Another approach used in Earth Science study for feature extraction is singular value decomposition [WSB92]. Here we investigate the use of this approach for removing seasonality. We first compute the singular value decomposition of the matrix, M , whose rows consist of the collection of time series that are of interest, i.e., in this case, the matrix rows consist of the sea surface temperature time series for a large number of points on the ocean (~150,000 points). A singular value decomposition expresses an m by n matrix, M , as the sum of simpler rank 1 matrices as follows:

$$M = \sum_{i=1}^n s_i \vec{u}_i \vec{v}_i',$$

where s_i , a scalar, is the i^{th} singular value of M , \vec{u}_i is the i^{th} left singular vector, and \vec{v}_i is the i^{th}

right singular vector. All singular values beyond the first r , where $r = \text{rank}(M)$ are 0 and all left (right) singular vectors are orthogonal to each other and are of unit length.

Thus, a matrix can be approximated by omitting some of the terms of the series that correspond to non-zero singular values. In particular, if a characteristic of the data corresponds to a particular term (singular value), then this characteristic can be removed by eliminating the corresponding term. For example, removing the first term,

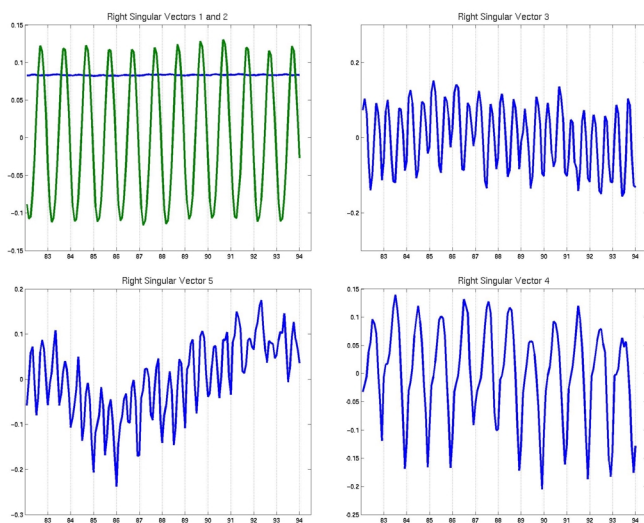


Figure 6: First five right singular values of SST data. (In top left plot, second right singular vector is green.)

which corresponds to the largest singular value, removes a constant component from the data, i.e., after removing the first term the maximum mean value of any times series from is 0.02. (Before there was a wide distribution of mean values, e.g., many time series in the tropics had means in 20's.) Thus, in this case, removing the first term is roughly equivalent to normalizing each time series to have a mean value of 0.

The nature of each term can be analyzed by looking at the associated right singular vector, which, in this case, can be interpreted as a time series. Figure 6 shows the first five right singular vectors for the SST matrix. (Singular values are non-negative and ordered by decreasing magnitude. Since the magnitudes of these singular values often decrease rapidly, it is often sufficient to consider only the first few.) From the first plot we see that the 1st and 2nd right singular

vectors, correspond, respectively, to a constant and a 12-month seasonal component. Right singular vector 4 also corresponds to a 12-month seasonal component, although it is not as regular as that of vector 2. Finally, right singular vectors 3 and 5 seem to correspond to 6-month seasonal cycles.

Figure 7 shows the sample SST time series after the first five singular value components have been removed. For reference it is plotted with the series obtained by using the monthly Z score transformation. The two different approaches produce time series that are relatively close (a correlation of 0.84). However, the SVD approach for removing seasonality is more computationally intensive than the other approaches. Also, the other approaches seem more “direct,” i.e., they can remove seasonality from a single vector, while the SVD approach works on a data set as a whole and only works because seasonality is such a strong characteristic of the entire data set that it manifests itself in the first few terms of the singular value decomposition. However, we plan to investigate the use of SVD to see if it can tell us anything interesting about the underlying Earth science phenomena.

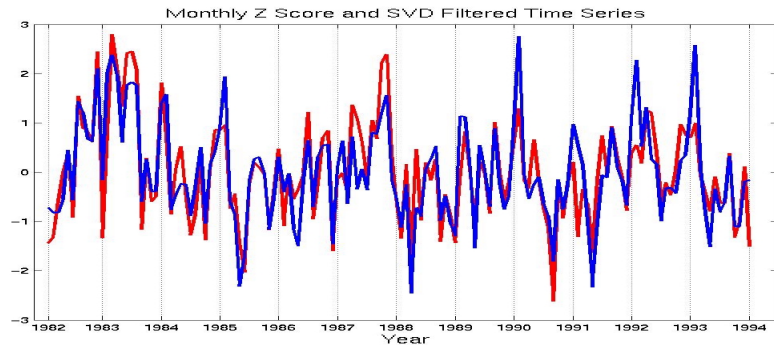


Figure 7: Results of applying monthly Z score and SVD filtering.

Moving average. A 12-month moving average is effective at removing seasonality and it also smoothes the data. To see why a moving average removes certain frequencies, consider that the average of a sine or cosine over the extent of its period is 0. However, it tends to flatten any deviation from the average values by spreading the effects of the deviations to its neighboring points in time. For comparison, Figure 8 shows the monthly Z score and the 12-month moving average transformation of the original SST time series. (The 12-month moving average is 11 months shorter; so for plotting purposes, this missing portion was set to 0.) Figure 8 suggests that if the high frequency fluctuations in the original time series are factored out, then the 12-month moving average of the original time series should be quite similar to the monthly Z score time series.

To illustrate this last point further, we apply a 12-month moving average to the monthly Z score series. This resulting series, along with the 12-month moving average series from Figure 8, are shown in Figure 9. The correlation between the two time series is 0.99. Thus, for our sample times series, using a 12-month moving average

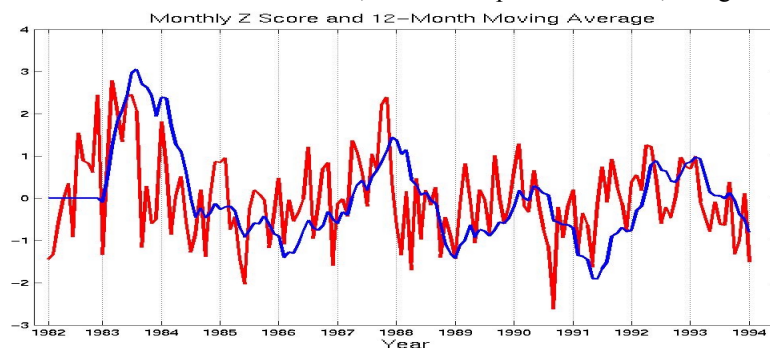


Figure 8: Monthly Z score and 12-month moving average.

to smooth the time series obtained by first applying a monthly Z score results in almost exactly the same time series as obtained by just applying a 12-month moving average to the sample time series. We have noticed for other time series that the correlation between the two approaches is not always quite so high, but this phenomenon seems to hold, in many cases.



Figure 9: Monthly Z score smoothed by 12-month average and 12-month moving average.

To fully understand this phenomenon, consider a time series $\mathbf{x} = \{x_1, x_2, \dots, x_{144}\}$. Let $\mathbf{p} = \{p_1, p_2, \dots, p_{132}\}$ be the 12-month moving average time series for \mathbf{x} and $\mathbf{q} = \{q_1, q_2, \dots, q_{132}\}$ be the 12-month moving average on the Z-score for \mathbf{x} . Note that

$$\Delta p_{12} = p_2 - p_1 = \frac{1}{12} \sum_{i=2}^{13} x_i - \frac{1}{12} \sum_{j=1}^{12} x_j = \frac{x_{13} - x_1}{12}$$

while

$$\Delta q_{12} = q_2 - q_1 = \frac{1}{12} \sum_{i=2}^{13} z_i - \frac{1}{12} \sum_{j=1}^{12} z_j = \frac{z_{13} - z_1}{12} = \frac{x_{13} - x_1}{12\sigma_1}$$

where both x_{13} and x_1 are standardized by the same monthly mean (μ_1) and monthly standard deviation (σ_1). The above analysis suggests that differences between consecutive points in the smoothed Z-score are proportional to the 12-month moving average, scaled by the monthly standard deviation. Thus, the correlation between \mathbf{p} and \mathbf{q} should be high if the volatility of the monthly standard deviations is low. The behavior of the correlation in other cases is still under investigation.

4. Dealing with the Temporal Autocorrelation of Data

Temporal autocorrelation has a direct impact on the significance of statistical correlation computed between two time series. For example, the number of degrees of freedom in a time series is reduced by a factor of k whenever a k -month moving average is applied. One way to evaluate the degree of autocorrelation in a time series data is by computing the autocorrelation coefficient. Given N time-series observations, $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$:

$$\text{Autocorrelation coefficient at lag } k, c(k) = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2}$$

The distribution of $c(k)$ for a completely random time series with large N (> 30) is $\mathcal{N}(0, 1/N)$. Furthermore, a rough estimate of the effective degrees of freedom in a time-series data is given by $-N \log c(1)$. For example, the effective degrees of freedom would be close to N if $c(1)$ is less than $1/e = 0.3679$. A plot of $c(k)$ at various lags k , also known as a correlogram, can be used to aid the interpretation of autocorrelation coefficients. Figure 10 illustrates the effect of applying the monthly Z-score transformation on an NPP time series that resembles Figure 2. The correlogram for the raw NPP exhibits strong periodic oscillations at 12-month intervals. Monthly Z-score reduces significantly the long-term autocorrelation present in the data. Nevertheless, short-term autocorrelations due to temporal locality between adjacent months are still persistent.

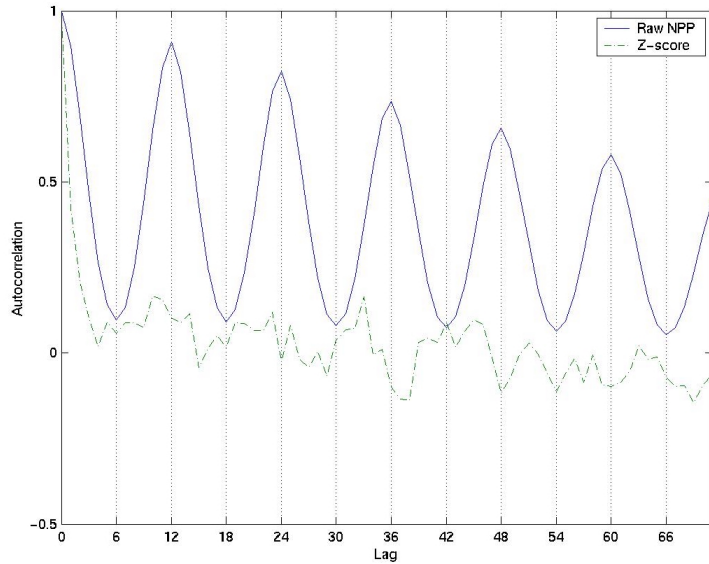


Figure 10: Effect of various transformations on the autocorrelation of an NPP time series data.

One way to reduce the short-term auto-correlation is by aggregating the time series into bins of 3-month intervals. For example, instead of having 12 values for each year, we can create bins by taking the average value of every 3 months (Jan-Feb-Mar, May-Apr-Jun, Jul-Aug-Sep and Oct-Nov-Dec) to obtain 4 bins per year. The effect of binning on temporal autocorrelation is illustrated in Figure 11. The top two figures show the histogram of $c(1)$ for all the NPP time series data with and without binning. These figures suggest that binning can help to reduce the amount of short-term autocorrelation. The bottom two figures illustrate the $c(1)$ histograms when the monthly Z-score transformation is applied. The results indicate that Z-score with binning works very well to reduce the amount of short-term autocorrelation in the time series data.

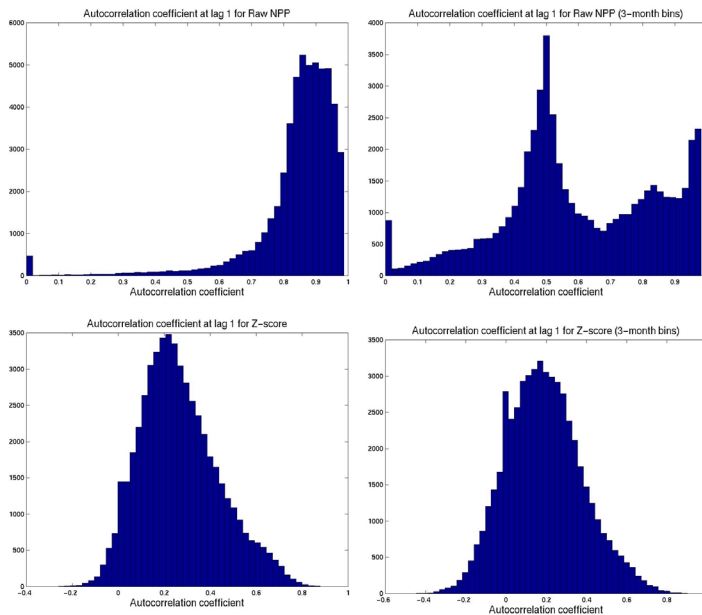


Figure 11: Distribution of autocorrelation coefficients at lag 1, $c(1)$, for raw NPP versus Z-scores (with and without 3-month binning).

5. Association Analysis

The definition and formation of events for our data mining approach are initially based on the domain knowledge of our Earth Science co-investigators. The input data from which the events are formed include NPP, the climate variables and climate indices. For land and ocean variables, we define anomalous events by transforming the variables into their monthly Z scores (to deseasonalize the time series) and then imposing upper and lower thresholds (e.g. ± 2 standard deviations) for these values. For climate indices, we define events based on the 5th and 95th percentiles of their 43-year time series data (from 1958 to 2000).

Ecologists are interested in a variety of spatio-temporal association patterns involving sequences of events abstracted from the measurement values of ecological variables at various spatial locations as shown in Figure 12. The spatio-temporal nature of the Earth science data sets gives rise to four types of patterns:

- **Intra-zone non-sequential patterns** – relationships among events in the same grid cell or zone, ignoring the temporal aspects of the data.
- **Inter-zone non-sequential pattern** – relationships among events happening in different grid cells or zones, ignoring temporal aspects of the data.
- **Intra-zone sequential pattern** – temporal relationships among events occurring within the same grid cell or zone.
- **Inter-zone sequential pattern** – temporal relationships among events occurring at different spatial locations.

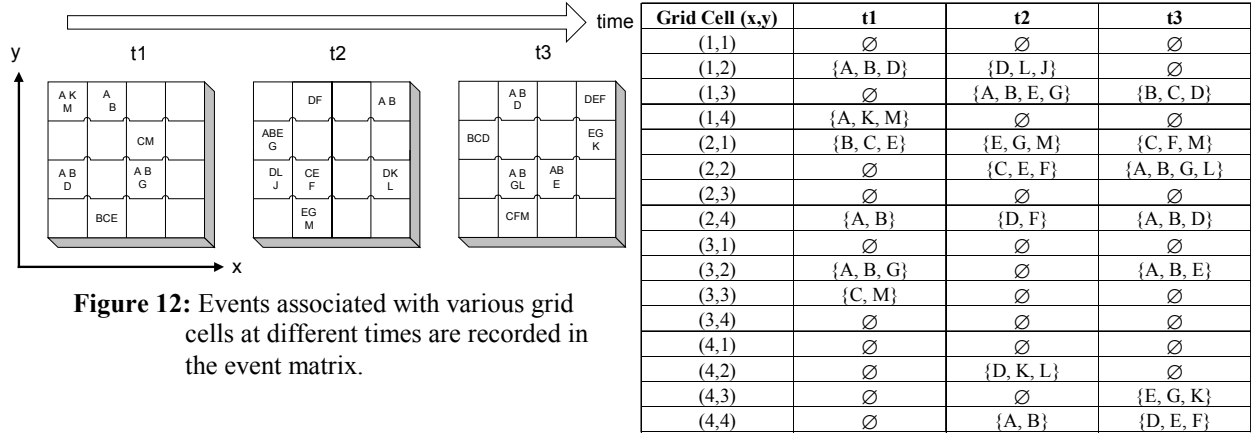


Figure 12: Events associated with various grid cells at different times are recorded in the event matrix.

One way to generate association patterns from the Earth Science data is to transform the spatio-temporal dataset into a set of market-basket type *transactions*. The main advantage of doing this is that we can use many of the existing algorithms to discover the association patterns that exist in the data. We will show how the different kinds of association patterns can be derived and discuss some of their limitations.

(Grid cell, time)	NPP-Lo	NPP-Hi	FPAR-Lo	FPAR-Hi	Temp-Lo	Temp-Hi	Prec-Lo	Prec-Hi	...
((1,1), t ₁)	1	0	0	0	0	0	0	0	...
((1,2), t ₁)	0	0	0	1	1	0	1	0	...
...
((1,1), t ₂)	0	1	1	0	0	0	1	0	...
((1,2), t ₂)	1	0	1	0	0	0	0	0	...

Table 2: Transforming the spatio-temporal data into market-basket type transactions.

Intra-zone Non-sequential Association Patterns

In the simplest case, we can look for intra-zone non-sequential associations among events occurring at the same spatial location, irrespective of the time of occurrence. The abstracted event matrix of Figure 12 can be transformed into a transaction format as shown in Table 2. This representation allows us to apply existing association rule mining algorithms, such as Apriori [AS94] and FP-tree [HPY00], to extract the intra-zone non-sequential patterns described in Section 3. The interestingness of an association rule $A \rightarrow B$ can be evaluated according to various objective interestingness measures:

1. Confidence = $\frac{P(A, B)}{P(A)}$
2. Correlation = $\frac{P(A, B) - P(A)P(B)}{\sqrt{P(A)(1 - P(A))P(B)(1 - P(B))}}$
3. Maximum Entropy = $\frac{H(A) + H(B) - H(A, B)}{\min(H(A), H(B))}$

where $H(A) = -P(A)\log P(A) - P(\neg A)\log P(\neg A)$, $H(B) = -P(B)\log P(B) - P(\neg B)\log P(\neg B)$, and
 $H(A,B) = -P(A,B)\log P(A,B) - P(\neg A,B)\log P(\neg A, \neg B)$
 $-P(A, \neg B)\log P(A, \neg B) - P(\neg A, \neg B)\log P(\neg A, \neg B)$

4. lift = $\frac{P(B | A)}{P(B)}$
5. Interest-support, IS = $\frac{P(A, B)}{\sqrt{P(A)P(B)}}$
6. Interest = $\frac{P(A_1, A_2, \dots, A_k, B_1, B_2, \dots, B_j)}{P(A_1)P(A_2)\dots P(A_k)P(B_1)P(B_2)\dots P(B_j)}$

Table 3 illustrates four of the highest ranked rules ordered according to the different objective interest measures. Some of the measures behave very similarly to each other in terms of the rankings they produce (e.g. correlation and IS; entropy and lift). We also found that most of the high-confidence rules have very low correlations, and that most of the high-correlation rules have very low confidence. For example, the highest confidence rule in Table 3 (PET-Hi Prec-Hi FPAR-Hi Temp-Hi \rightarrow NPP-Hi) has a correlation of 0.0216 whereas the most highly correlated rule (NPP-Lo \rightarrow FPAR-Lo) has a confidence of 58.7%. This is because the confidence measure does not take into account the support of the items in the rule consequent, a fact that was pointed out by Brin, et al. in [BSM97]. On the other hand, interest measures such as correlation, lift, interest factor and IS are symmetric (e.g. correlation of the rules $A \rightarrow B$ and $B \rightarrow A$ are the same). Thus, they are more appropriate to rank frequent itemsets instead of association rules.

Rules ordered by Confidence	Rules ordered by Correlation	Rules ordered by Entropy
1. PET-Hi Prec-Hi FPAR-Hi Temp-Hi \rightarrow NPP-Hi (Conf = 100%)	1. FPAR-Lo \rightarrow NPP-Lo (Corr = 0.4327)	1. Prec-Hi FPAR-Hi Solar-Lo Temp-Lo \rightarrow NPP-Hi PET-Lo (ent = 0.4320)
2. PET-Hi Temp-Lo \rightarrow Solar-Hi (Conf = 99.4%)	2. FPAR-Hi \rightarrow NPP-Hi (Corr = 0.4013)	2. PET-Hi Prec-Lo FPAR-Lo \rightarrow NPP-Lo Temp-Hi (ent = 0.3207)
3. PET-Hi Prec-Hi FPAR-Hi \rightarrow NPP-Hi (Conf = 98.6%)	3. Solar-Lo \rightarrow PET-Lo (Corr = 0.2752)	3. NPP-Hi Solar-Lo Temp-Lo \rightarrow PET-Lo FPAR-Hi (ent = 0.2899)
4. NPP-Lo PET-Lo Temp-Hi \rightarrow Solar-Lo (Conf = 98.0%)	4. PET-Lo FPAR-Lo \rightarrow NPP-Lo (Corr = 0.1966)	4. FPAR-Hi Solar-Hi Temp-Hi \rightarrow NPP-Hi PET-Hi (ent = 0.2870)

Rules ordered by Lift	Rules ordered by IS	Rules ordered by Interest factor
1. Prec-Hi FPAR-Hi Solar-Lo Temp-Lo \rightarrow NPP-Hi (Lift = 366.2)	1. NPP-Lo \rightarrow FPAR-Lo (IS = 0.4667)	1. FPAR-Hi \rightarrow NPP-Hi (I = 0.0362)
2. NPP-Hi Solar-Lo Temp-Lo \rightarrow PET-Lo FPAR-Hi (Lift = 308.7)	2. FPAR-Hi \rightarrow NPP-Hi (IS = 0.4611)	2. PET-Lo \rightarrow Solar-Lo (I = 0.0309)
3. NPP-Hi Prec-Hi Solar-Lo Temp-Lo \rightarrow PET-Lo FPAR-Hi (Lift = 293.4)	3. PET-Lo \rightarrow Solar-Lo (IS = 0.3674)	3. NPP-Lo \rightarrow FPAR-Lo (I = 0.2899)
4. PET-Hi Prec-Lo FPAR-Lo \rightarrow NPP-Lo Temp-Hi (Lift = 284.6)	4. PET-Lo FPAR-Lo \rightarrow NPP-Lo (IS = 0.2060)	4. PET-Lo Prec-Hi \rightarrow Solar-Lo (I = 0.0073)

Table 3: Intra-zone Non-sequential Association Patterns.

Our overall results show that using a single interest measure may not be sufficient to capture all the interesting patterns. Furthermore, we need the help of domain experts to interpret many of the discovered patterns. We found that visualization is an important tool to assist the domain experts in evaluating the interestingness of these patterns. Figure 13 shows the regions that are covered by one of the highly correlated pattern, FPAR-Hi \rightarrow NPP-Hi. FPAR (Fractional Intercepted Photosynthetically Active Radiation) is an attribute derived from NDVI (the Normalized Difference Vegetation Index), a greenness index based on satellite measurements. Anomalously high FPAR means that the vegetation has generated more “light-harvesting” photosynthetic capability than average, which allows for higher than normal NPP. Regions that show this pattern correspond mainly to semi-arid annual grasslands, a type of vegetation, which is able to more quickly take advantage of periodically high precipitation (and possibly solar radiation) than forests. The FPAR-Hi events could be related to unusual precipitation conditions, but more study is needed to verify this hypothesis. Another interesting pattern relates NPP-Lo events to PET-Lo and FPAR-Lo (as shown in Figure 14). PET (Potential EvapoTranspiration) measures the potential loss of water to the atmosphere by evaporation, and by the transpiration of water through plants. This pattern occurs frequently in the regions of evergreen forests. Our tentative hypothesis is that these regions represent forests that have temporarily lost their

photosynthetic capability (FPAR-Lo) due to a large fire or other disruptive events, but this needs to be verified by consulting historical records, which are not easily obtained through conventional sources.

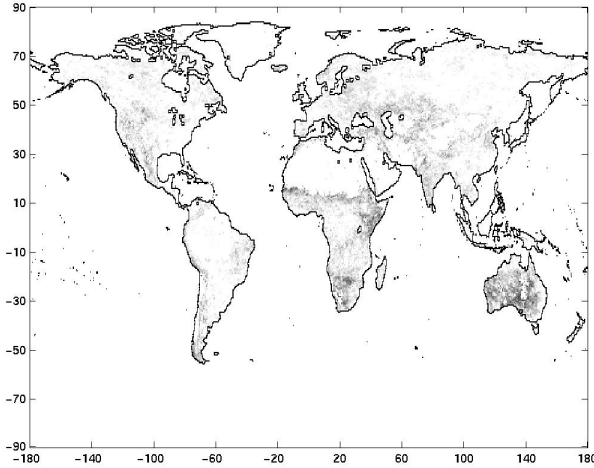


Figure 13: Regions that show the intra-zone non-sequential association rule $\{FPAR-Hi\} \rightarrow \{NPP-Hi\}$. The dark region corresponds to areas that have high support for the rule.

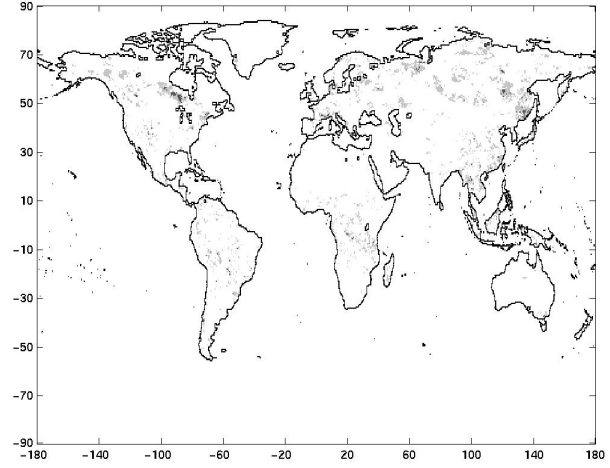


Figure 14: Regions that show the intra-zone non-sequential association rule $\{FPAR-Lo, PET-Lo\} \rightarrow \{NPP-Lo\}$.

Intra-zone Sequential Association Patterns

If temporal information is incorporated, we can derive intra-zone sequential associations among these events using existing sequential pattern discovery algorithms such as [SA96] and [MTV97]. The input data for these algorithms are the event sequences shown in Figure 12. We have used the GSP algorithm, which was initially proposed by Agrawal et al. [SA96], for finding frequent sequential patterns in market-basket data. In the GSP approach, a sequence is represented as an ordered list of itemsets, $s = \langle s_1, s_2, \dots, s_n \rangle$. Each element s_i of the sequence is subjected to three timing constraints: window-size (i.e. maximum time interval among all items in the element), min-gap (i.e. minimum time difference between successive elements) and max-gap (maximum time difference between successive elements). In our experiments, we have chosen the window-size to be 0 (i.e. all events in the same element must occur in the same month), min-gap to be 0 and max-gap to be 3 months. We have generated sequential patterns for two separate regions, Peru and Australia. We have chosen these two regions primarily because they are located close to the regions in the Pacific where the El-Nino related events occur. Table 4 shows the intra-zone sequential patterns for both regions (note that we only consider patterns with NPP events as one of items in the last element of the sequence). We observe that the sequential patterns for Australia often contain the NPP-Hi event whereas the sequential patterns for Peru often contain the NPP-Lo events. Furthermore, the confidence of the patterns for Australia is higher than those for Peru.

Intra-zone sequential pattern (for Australia)	Confidence	Rank	Correlation
(Solar-Lo) \rightarrow (NPP-Hi) \rightarrow (Temp-Hi) \rightarrow (NPP-Hi)	80.0%	1	-0.1834
(Prec-Hi) \rightarrow (NPP-Hi) \rightarrow (NPP-Hi) \rightarrow (NPP-Hi Prec-Hi) \rightarrow (NPP-Hi)	78.8%	2	-0.1958
(PET-Lo) \rightarrow (NPP-Hi) \rightarrow (Temp-Hi) \rightarrow (NPP-Hi)	69.2%	3	-0.2941
(Prec-Hi) \rightarrow (NPP-Hi)	67.6%	7	-0.3088

Intra-zone sequential pattern (for Peru)	Confidence	Rank	Correlation
(PET-Hi) \rightarrow (PET-Hi) \rightarrow (NPP-Lo) \rightarrow (NPP-Lo)	61.7%	1	-0.3324
(NPP-Lo) \rightarrow (NPP-Lo)	49.4%	2	-0.4628
(Prec-Lo) \rightarrow (NPP-Lo)	37.3%	3	-0.5956
(Temp-Hi) \rightarrow (Prec-Hi) \rightarrow (Prec-Hi) \rightarrow (NPP-Lo)	34.0%	6	-0.6025

Table 4: Intra-zone sequential patterns for two regions: Australia and Peru. These patterns are obtained using the GSP algorithm with mingap=0, maxgap=3 and window size=0.

In the GSP approach, each data sequence contributes at most once to the support count of a pattern. As a result, the support and confidence measures depend only on the number of spatial locations for which the pattern is observed. This type of counting strategy may not be appropriate because it does not take into account the number of times the pattern occurs at each spatial location. We are currently investigating the possibility of using other support counting schemes as suggested in [JKK99]. Another problem is that statistical correlation may no longer be an appropriate measure because the support of the last element (NPP-Hi or NPP-Lo) is often significantly larger than the support of the entire sequence. As a result, long sequences often have correlation values that are either negative or close to zero.

Inter-zone Sequential and Non-Sequential Association Patterns

There are various ways to incorporate the spatial components of the data into the association pattern discovery problem. Koperski, et al. [KH95] have extracted spatial association rules by using geographical landmarks to represent interesting spatial locations and spatial predicates to represent relationships between the spatial objects. Shekhar, et al. [SH01] have derived spatial co-location rules based on the frequent co-occurrences of events within the same spatial window. In this paper, we use an approach that is similar to [Kh95] except we replace the geographical landmarks with events abstracted from climate indices. Each climate index represents the behavior of a particular climate variable over certain regions of interest (e.g. regions associated with the El-Nino phenomena). Thus, instead of finding spatial association patterns that may exist among any spatial locations, we have restricted the analysis to a few regions of interest. This reduces the number of spatial features dramatically. Furthermore, since we are dealing with regions instead of individual grid cells, we no longer have the problem of insufficient support. In the following work, we have used events derived from the climate indices given in Table 1. Our goal is to find interesting associations between NPP and other climate events defined at a particular land point to interesting ocean events abstracted from the climate indices. Table 5 shows the input data to for the inter-zone non-sequential pattern algorithms. Tables 6 and 7 illustrate some of the highly ranked inter-zone non-sequential and sequential patterns derived from this data. Some of the patterns we found suggest teleconnections between ocean basins, as seen with the sequential associations between ocean indices in the Pacific basin (e.g., NINO12-Hi) and Atlantic basin (e.g., NAO-Hi). This parallels the recent climatological research results that have identified the tropical oceans as drivers for the atmospheric heating which is altering the spatial structure of the NAO [HHX01].

(Grid cell, time)	NPP-Lo	NPP-Hi	FPAR-Lo	FPAR-Hi	...	SOI-Hi	SOI-Lo	AO-Hi	...
((1,1), t ₁)	1	0	0	0	...	0	1	0	...
((1,2), t ₁)	0	0	0	1	...	0	1	0	...
...
((1,1), t ₂)	0	1	1	0	...	0	1	1	...
((1,2), t ₂)	1	0	1	0	...	0	1	1	...

Table 5: Transaction data for mining inter-zone non-sequential association patterns.

Rules ordered by Confidence	Rules ordered by Correlation
1. PET-Hi Prec-Lo FPAR-Lo Temp-Hi AO-Hi NINO12-Hi NINO3-Hi NINO4-Hi NINO34-Hi SOI-Lo WP-Hi → NPP-Lo (Conf = 100%)	1. FPAR-Hi → NPP-Hi (Corr = 0.4013)
2. FPAR-Hi Solar-Hi Temp-Hi AO-Lo NAO-Lo NINO3-Hi NINO34-Hi PDO-Hi QBO-Hi → NPP-Hi (Conf = 100%)	2. FPAR-Hi Solar-Lo → NPP-Lo (Corr = 0.1992)
3. Prec-Lo FPAR-Lo Temp-Lo NAO-Hi SOI-Lo → NPP-Lo (Conf = 100%)	3. FPAR-Hi PDO-Hi → NPP-Hi (Corr = 0.1975)

Table 6: Inter-zone Non-sequential Association Patterns.

Inter-zone sequential pattern (for Australia)	Confidence	Correlation
(Temp-Lo) → (NINO12-Hi) → (NAO-Hi) → (NPP-Hi) → (NPP-Hi) → (NPP-Hi)	95.7%	-0.0213
(AO-Hi) → (Solar-Lo) → (SOI-Lo) → (QBO-Hi) → (Prec-Hi) → (NPP-Hi)	95.1%	-0.0278
(SOI-Lo) → (NAO-Hi) → (Solar-Lo) → (Prec-Hi) → (NAO-Hi) → (NPP-Hi)	95.0%	-0.0290

Inter-zone sequential pattern (for Peru)	Confidence	Correlation
(WP-Hi) → (Solar-Hi) → (NINO34-Lo) → (Temp-Hi) → (NPP-Lo)	89.7%	-0.0225
(Solar-Hi) → (NINO34-Lo) → (NINO34-Lo) → (Temp-Hi) → (NPP-Lo)	89.7%	-0.0225
(NAO-Lo) → (NPP-Lo) → (NINO34-Lo) → (NPP-Lo)	87.1%	-0.0519

Table 7: Inter-zone sequential patterns for two regions: Australia and Peru. These patterns are obtained using the GSP algorithm with mingap=0, maxgap=3 months and window size=0.

One potential problem with this approach is that one may end up generating too many associations among events abstracted from the climate indices. For example, suppose events SOI-Lo and AO-Hi co-occur together only once at time t_k . However, the support of these two events will be as high as the total number of grid cells in the data set because the values of SOI-Hi and AO-Hi are both equal to 1 for all grid cells at time t_k . This increases the execution time of the standard association pattern discovery algorithms dramatically and reduces the significance of the support and confidence measures. Techniques for handling these problems are currently under investigation.

6. Conclusion

Our initial approach for finding association patterns transformed the data so that standard techniques could be applied. These techniques have uncovered some interesting ecosystem patterns for Earth scientists to investigate. However, some of these approaches lead to dense transaction matrices, and consequently, require significant computational time. Also, the standard measures of interestingness do not consistently identify interesting associations in this domain. For future work, we will investigate other methods for counting support, such as the ones suggested in [JKK99].

We have explored several techniques for deseasonalizing Earth Science time-series data, and our results show that several of these techniques are effective. However, there are still issues related to autocorrelation and its effect on the significance of the correlation between two time series. Although binning and removing seasonality reduce the level of autocorrelation significantly, additional investigation is needed to explore different binning techniques and to quantify the effects of any remaining autocorrelation on the significance of observed correlations. Finally, trends (the long-term change in the mean value of the time series) are another important source of variation in time series data and we plan to include trend detection in our future work.

References

- [AS94] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," In *Proc. of the 20th VLDB Conference* (1994).
- [BSM97] S. Brin, C. Silverstein, R. Motwani, "Beyond Market Baskets: Generalizing Association Rules to Correlations", *Data Mining and Knowledge Discovery*, 2: 39-68 (1998).
- [HHX01] Martin P. Hoerling, James W. Hurrell, and Taiyi Xu, "Tropical Origins for Recent North Atlantic Climate Change" *Science* 292: 90-92 (2001).
- [HPY00] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", In *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00)* (2000).
- [JKK99] M. V. Joshi, G. Karypis, and V. Kumar, "A Universal Formulation of Sequential Patterns", Technical Report # 99-021, University of Minnesota (1999).
- [KH95] K. Koperski and J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases", In *Proc. 4th Int'l Symp. on Large Spatial Databases (SSD95)*, 47-66 (1995).
- [LCM+97] Z. Li., J. Cihlar, L. Moreau, F. Huang, and B. Lee, "Monitoring fire activities in the boreal ecosystem," *Journal Geophys. Res.*, 102(29): 611-629 (1997).
- [MTV97] H. Mannila, H. Toivonen, A.I. Verkamo, "Discovery of Frequent Episodes in Event Sequences", *Data Mining and Knowledge Discovery*, 1(3): 259-289 (1997).
- [NVA+99] D. C. Nepstad, A. Verissimo, A. Alencar, C. Nobre, E. Lima, P. Lefebvre, P. Schlesinger, C. Potter, P. Moutinho, E. Mendoza, M. Cochrane, and V. Brooks. "Large-scale impoverishment of Amazonian forests by logging and fire. ", *Nature*, 398: 505-508 (1999).
- [PKB99] C. S. Potter, S. A. Klooster, and V. Brooks, "Inter-annual variability in terrestrial net primary production: Exploration of trends and controls on regional to global scales," *Ecosystems*, 2(1): 36-48 (1999).
- [SA96] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements," In *Proc. of the Fifth International Conference on Extending Database Technology* (1996).
- [Ste+01] M. Steinbach, P. N. Tan, V. Kumar, C. Potter, S. Klooster, A. Torregrosa, "Clustering Earth Science Data: Goals, Issues and Results", In *Proc. of the Fourth KDD Workshop on Mining Scientific Datasets* (2001).
- [SH01] Shashi Shekhar and Yan Huang, "Discovering Spatial Co-location Patterns: a Summary of Results," In *Proc. of 7th International Symposium on Spatial and Temporal Databases (SSTD01)* (2001).
- [WSB92] John M. Wallace, Catherine Smith, and Christopher S. Bretherton, "Singular Value Decomposition of Wintertime Seas Surface Temperature and 500-mb Height Anomalies," *Journal of Climate*, 561-576 (2001).