Mining Scientific Data: Discovery of Patterns in the Global Climate System [*]

Vipin Kumar[+] Michael Steinbach[+] Pang-Ning Tan[+]
Steven Klooster[+++] Christopher Potter[++] Alicia Torregrosa[+++]

[+] Department of Computer Science and Engineering, Army HPC Research Center
University of Minnesota
{ kumar, steinbac, ptan@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 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 first discuss some of the challenges involved in preprocessing and analyzing the data. 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. After preprocessing, we apply clustering and different kinds of association analysis to the data to discover spatio-temporal relationships among ecological variables at various parts of the Earth. Our current technique for finding associations extracts sets of events from the time series data and then applies existing algorithms traditionally used for market-basket data. We use K-means clustering to divide the land and ocean areas of the earth into disjoint regions in an automatic, but meaningful, way that enables the direct or indirect discovery of interesting patterns.

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

[*] 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.

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 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 investigated the problem of detecting temporal auto-correlation and determining the statistical significance of various descriptive statistics, but those results are reported in [Tan+01].

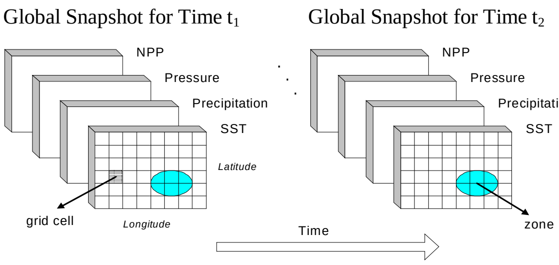

Figure 1: A simplified view of the problem domain.

Our goal is to discover spatio-temporal relationships among ecological variables observed at various parts of the Earth. This 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). An effective way to do this is to use clustering to divide areas of the land and ocean into disjoint regions in an automatic, but meaningful way. This enables us to more easily identify regions of the earth whose constituent points have similar short-term and long-term climate characteristics. Given relatively uniform clusters we can then identify how various ecosystem phenomena, such as El Nino, influence the climate and NPP of different regions.

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]. We describe the various types of spatio-temporal association that can be extracted from this data.

The rest of the paper is organized as follows: Section 2 provides a description of the ecology data, while section 3 presents some of the temporal issues related to the removal of seasonality. Section 4 discusses association pattern discovery, while sections 5 and 6 introduce our clustering approach and show the results. Section 7 concludes with a summary and 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 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 ½° by ½° 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.

Filtering based on the DFT (Discrete Fourier Transform). This approach is based on standard signal processing techniques. 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.)

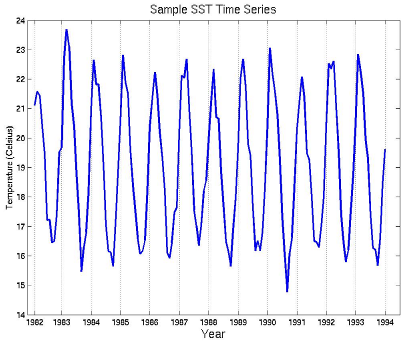

Figure 2: Sample SST time series

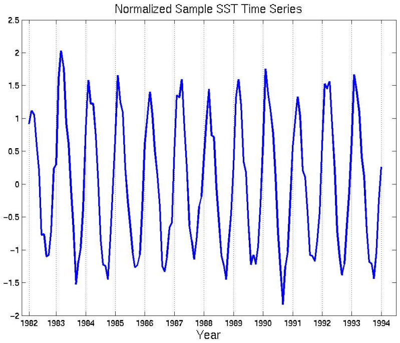

Figure 3: Standardized sample SST time series

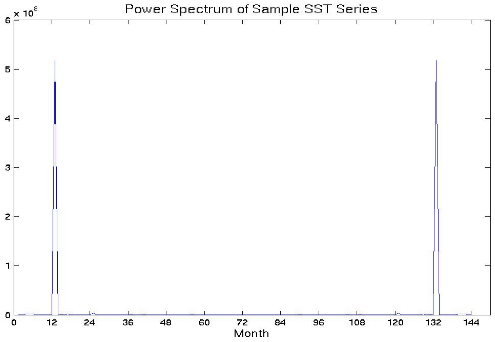

Figure 4: Power Spectrum of sample SST time series (constant component removed).

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.

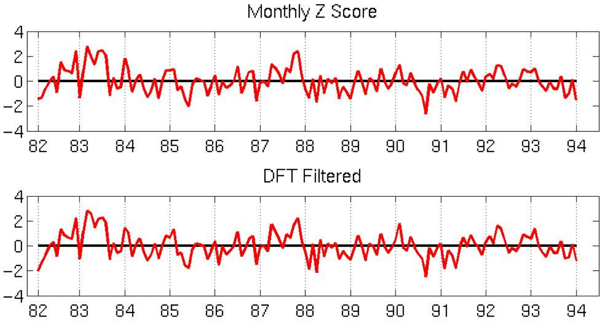

Figure 5: Results of applying monthly Z score and DFT filtering.

Singular value decomposition (**SVD**). Another approach used in Earth Science study for feature extraction is singular value decomposition. 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* = 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, 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.

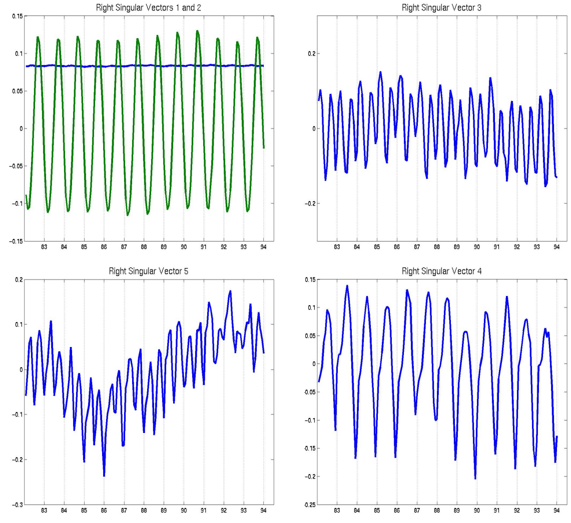

Figure 6: First five right singular values of SST data. (In top left plot, second right singular vector is green.)

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.

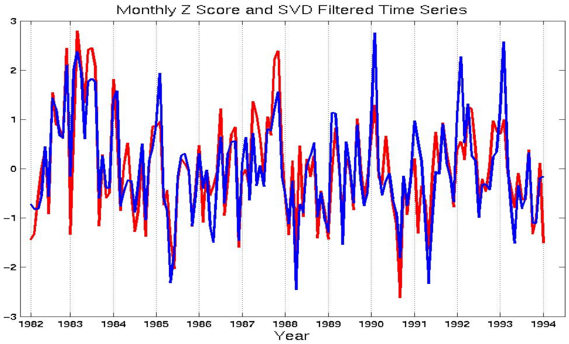

Figure 7: Results of applying monthly Z score and SVD filtering.

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 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.

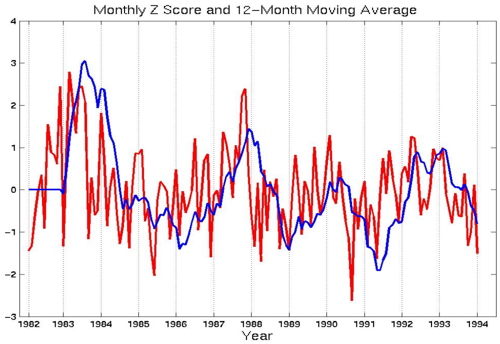

Figure 8: Monthly Z score and 12-month moving average.

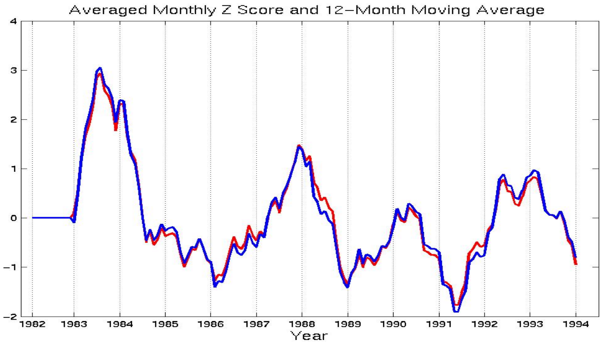

Figure 9: Monthly Z score smoothed by 12-month average and 12-month moving average.

To fully understand this phenomenon, consider a time series ***x*** = { x_1, x_2, …, x_{144}}. Let ***p*** = {p_1, p_2, …, p_{132}} be the 12-month moving average time series for **x** and ***q*** = { q_1, q_2, …, q_{132}} be the 12-month moving average on the Z-score for **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 ***p*** and ***q*** should be high if the

(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 traditional, market-basket transactions.

volatility of the monthly standard deviations is low. The behavior of the correlation in other cases is still under investigation.

4. 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. The spatio-temporal nature of the Earth science data sets gives rise to four types of association 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.

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. For brevity, we will only discuss the results of mining intra-zone association patterns, i.e., strong associations among events that are occurring at the same spatial location, irrespective of their time of occurrences.

Spatio-temporal events can be transformed into a transaction format as shown in Table 2. This representation allows us to apply any existing association rule algorithm, such as Apriori [AS94] or FP-tree [HPY00], to extract the intra-zone non-sequential patterns. An association rule is an implication expression of the form A ➔ B, where A and B are sets of events.

The interestingness of an association rule A ➔ B can be evaluated objectively, using various objective measures such as confidence, correlation, lift and interest, or subjectively, with the help of domain experts. Our work uses both objective and subjective interestingness criteria, to filter out patterns that occur infrequently or are statistically insignificant, and to find novel or unexpected patterns. Visualization is an important tool to assist the domain experts in evaluating the subjective interestingness of these patterns. For example, Figure 10 shows the regions that are covered by one of the highly correlated pattern, FPAR-Hi ➔ 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. This pattern occurs at least once in 52.2% of all the land data points. However only 5.0% of all the land data points have support counts greater than 4 for this pattern. The region where the pattern has high support is shown in Figure 10. Regions that show this pattern correspond mainly to shrublands (Figure 11), a type of vegetation, which is able to more quickly take advantage of periodically high precipitation (and possibly solar radiation) than forests. This led the domain experts of our team to believe that the FPAR-Hi events could be related to unusual precipitation conditions, but more study is needed to verify this hypothesis.

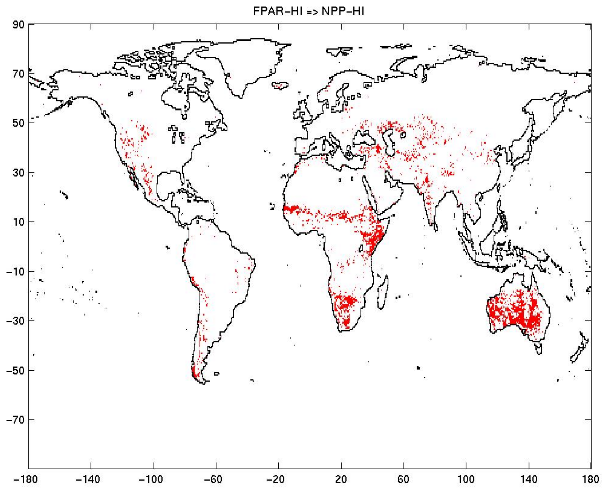

Figure 10: Regions that show the intra-zone non-sequential association rule {FPAR-Hi} ➔{NPP-Hi}. The red region corresponds to areas that have high support for the rule.

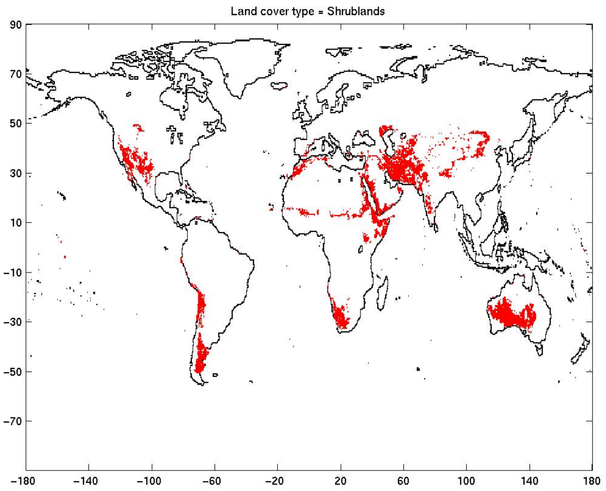

Figure 11: Shrubland regions.

5. A K-means Based Clustering Approach

Clustering, often better known as spatial zone formation in this context, segments oceans and land into smaller pieces that are relatively homogeneous in some sense. While these zones can be specified directly by researchers, clustering provides a general data mining approach for automatically creating zones. Thus, our basic approach is to treat the zone creation problem as a cluster analysis problem [DJ88, KR90]. Cluster analysis groups objects (grid cells) so that the objects in a group are similar to one another and different from the objects in other groups. The clusters produced may be nested (hierarchical) or un-nested (partitional), overlapping or non-overlapping.

For our initial clustering approach, we chose the widely used K-means clustering algorithm [DJ88], which is simple and efficient. As our results will show, it was effective for our use of clustering during exploratory data analysis.

The K-means algorithm discovers K (non-overlapping) clusters by finding K centroids ("central" points) and then assigning each point to the cluster associated with its nearest centroid. (Note that a cluster centroid is typically the mean or median of the points in its cluster and "nearness" is defined by a distance or similarity function.) Ideally the centroids are chosen to minimize the total "error," where the error for each point is given by a function that measures the discrepancy between a point and its cluster centroid, e.g., the squared distance. Note that a measure of cluster "goodness" is the error contributed by that cluster. For squared error and Euclidean distance, it can be shown [And73] that a gradient descent approach to minimizing the squared error yields the following basic K-means algorithm. (Note that the previous discussion still holds if we use similarities instead of distances, but our optimization problem becomes a maximization problem.)

Basic K-means Algorithm for finding *K* clusters.

1. Select *K* points as the initial centroids.

2. Assign all points to the closest centroid.

3. Recompute the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids don't change (or change very little).

K-means has a number of variations, depending on the method for selecting the initial centroids, the choice for the measure of similarity, and the way that the centroid is computed. For this work, we followed the common practice of using the mean as the centroid and selecting the initial centroids randomly. For our similarity measure, we chose Pearson's correlation coefficient, which is defined as follows: The correlation coefficient r of two data vectors, x and y is given by

$$\boldsymbol{r} = \frac{\sum_{\boldsymbol{i}}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})(\boldsymbol{y_i}-\bar{\boldsymbol{y}})}{\sqrt{\sum_{\boldsymbol{i}}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^2\sum_{\boldsymbol{i}}(\boldsymbol{y_i}-\bar{\boldsymbol{y}})^2}}$$

, where x_i (y_i) is the value of the i^{th} attribute of $\boldsymbol{x}$ ($\boldsymbol{y}$), and $\bar{\boldsymbol{x}}$ ($\bar{\boldsymbol{y}}$) is the average value of all attributes of $\boldsymbol{x}$ ($\boldsymbol{y}$). Correlation has a value between –1 (perfect negative linear correlation) and 1 (perfect positive linear correlation), with a value of 0 indicating no linear correlation.

Since we are using correlation instead of Euclidean distance, there is a question of whether K-means will still "work." However, if the data is standardized by subtracting off the mean and dividing by the standard deviation, then a bit of algebraic manipulation will show that the correlation and the Euclidean distance are monotonically related, as shown in following equation

$$r(x^*, y^*) = 1 - \frac{d^2(x^*, y^*)}{2n},$$

where x^* and y^* are the standardized vectors of dimension n, and r and d are the correlation and Euclidean distance functions, respectively. Thus, the traditional K-means algorithm will "work" when used with correlation. Furthermore, the measure of cluster goodness that corresponds (at least monotonically) to the traditional squared distance is the sum of the similarity of each point in a cluster to the cluster centroid.

We make a brief comment about our reasons for using correlation. First, correlation is insensitive to changes in scale, and since we want to compare time series of different variable types, e.g., NPP and SST, we need this property. Also, correlation has been well studied by statisticians and thus, confidence intervals and tests for non-zero correlation are readily available. Finally, correlation is widely used as a measure of similarity between time series.

6. Results

In this section we show the use of clustering for detecting different sorts of ecosystem patterns. To do this we employ two kinds of diagrams. The first diagram shows which points on the globe belong to specific clusters by associating each cluster with a particular color. The second type of diagram plots the cluster centroids. Since the cluster centroids are time series, this type of a plot can show various types of temporal patterns. For example, for a cluster consisting of land points, each of which is characterized by a series of monthly NPP values, the centroid of a cluster provides a "summary" description of NPP for the points in that cluster.

Finding Seasonal Patterns and Anomalous Regions. Figure 12 shows the result of finding two clusters for NPP and (separately) finding two clusters for SST. (Note that the seasonal component has not been removed from this data.) The four clusters approximate the northern and southern hemispheres, for land and ocean. The plots of the land and sea centroids show strong yearly cycles. Interestingly, while the northern and southern hemisphere land clusters are mostly contiguous, some areas in the northern hemisphere, e.g., part of southern California, correspond to the "southern hemisphere" cluster and vice-versa. These regions correspond to climates, e.g., a Mediterranean climate, whose plant growth patterns are reversed from those typically observed in the hemisphere in which they reside. The existence of these anomalous climate regions is well known, but clustering allows them to be easily detected.

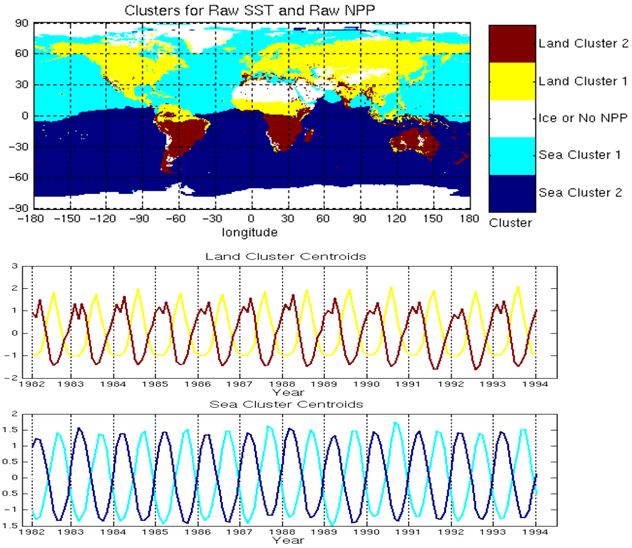

Figure 12. Two Ocean (SST) and Land (NPP) Clusters.

Identifying Connections between Land and Ocean Clusters. Another use of clustering is to investigate the relationship of various land and sea areas. In particular, by finding land and sea clusters that are highly correlated, we can identify potential teleconnection patterns, i.e., recurring and persistent climate patterns that span vast geographical areas. This works as follows. A large number of clusters are found for the land (NPP) and the sea (SST), say 100 for each. Then the correlations between various sea and land centroids are calculated, and the land and sea clusters with the highest correlations are plotted. Figure 13 shows such a diagram for sea cluster 39 (which is a region of ocean near the Philipines) and land cluster 87 (which consists of parts of Eastern Brazil, Southern Africa, and a bit of Australia). The NPP centroid of land cluster 87 is correlated with the SST centroid of sea cluster 39 at a level of 0.47. (For this analysis we removed seasonal variation by using the monthly Z score.) Figure 14 shows a plot of the centroid of sea cluster 39 (black) versus the cluster centroid of land clusters 87 (red). To better display the overall relationships between the centroids, Figure

15 shows the same centroids after they have been smoothed using a 12-month moving average.

Although this approach has the potential to detect new, previously unknown relationships, the teleconnection shown here is known to Earth scientists. In particular, sea cluster 39 is highly correlated (0.66), with SOI, which is a climate index related to El Niño, and it is known that parts of Southern Africa and Australia experience droughts related to El Nino.

Finally note that our work on clustering is described in more detail in [Ste+01].

7. Conclusions

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 removing seasonality and binning reduce the level of autocorrelation significantly [Tan+01], 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.

Our initial approach for finding intra-zone, non-sequential 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, for inter-zone patterns, these approaches lead to dense transaction matrices, and consequently, require significant computational time. Also, the standard measures of "what is interesting" do not consistently identify interesting associations in this domain. For future work, we will investigate other methods for determining which patterns are interesting.

Our clustering results indicate that clustering can play a useful role in the discovery of interesting ecosystem patterns. The patterns revealed by the clusters and their associated (centroids) time series are sometimes well known, e.g., the yearly seasonal variation of Figure 12. However, we have also started to investigate how clustering might be used to discover previously unknown relationships between regions of the land and sea. In particular, we have looked at which regions of the land are most highly correlated to the centroids of ocean regions. So far the ecologists on our team have found the results interesting and have recognized some familiar patterns. One challenge is to find techniques to automatically select interesting patterns and eliminate spurious ones.

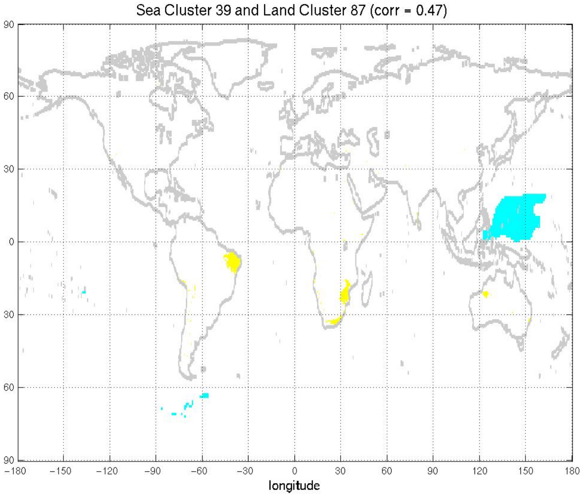

Figure 13: One Sea Cluster and Highly Correlated Land Cluster.

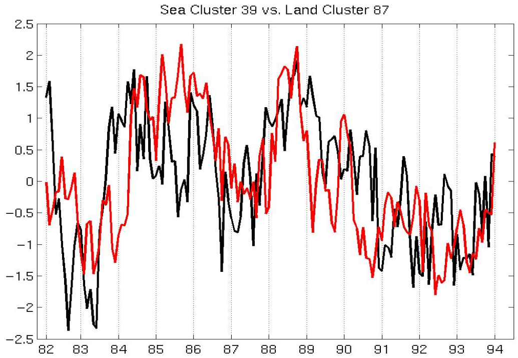

Figure 14: Comparison of Cluster Centroids.

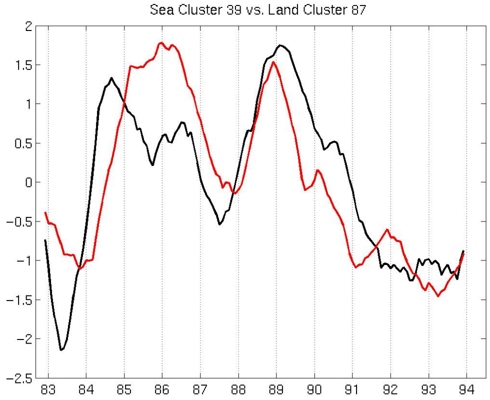

Figure 15: Comparison of Smoothed Cluster Centroids.

In clustering, there are a number of opportunities for future research. For instance, we could try other similarity measures, e.g., Euclidean distance or the cosine measure. We could also try the other clustering approaches, e.g., bisecting K-means [SKK00]. Along somewhat different lines, we may want to look at clusters that vary over time or we may want to try to define clusters in terms of events. (However, for some transformations of the data, e.g., the monthly Z score, we are in some sense already looking at events, i.e., deviations from the norm.) Also, our current clustering approach only looks at the time series for one variable for each point. This is a potential limitation in terms of the goodness of the clusters and their suitability for predicting the behavior of one region (cluster) based on the time varying behavior of another region.

Other limitations in our current work, both in clustering and in association analysis, result from the fact that often, only extreme events are correlated. For example, the El Nino indices have values for each month of each year, but the effects of El Nino on other regions often occur only when the index has an extreme value, i.e., when an El Nino effect is actually occurring. Although there may be a number of possible ways to address these problems and make the analysis more effective, it seems likely that some patterns will best be detected by other data mining techniques that are naturally more event-based, e.g., association rules or co-location rules [SH01].

REFERENCES

[And73] Michael R. Anderberg, *Cluster Analysis for Applications*, Academic Press (1973).

[AS94] R.Agrawal, and R.Srikant, “Fast Algorithms for Mining Association Rules,” In *Proc. of the 20th VLDB Conference* (1994).

[DJ88] R. C. Dubes and A. K. Jain, *Algorithms for Clustering Data*, Prentice Hall (1988).

[IND1] http://www.cgd.ucar.edu/cas/catalog/climind/

[IND2] http://www.cdc.noaa.gov/USclimate/ Correlation/ help.html

[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).

[KR90] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley and Sons (1990).

[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).

[RH96] C. F. Ropelewski and M. S. Halpert, "Quantifying Southern Oscillation - precipitation relationships", *J. Climate*, 9,1043-1059 (1996).

[SKK00] Michael Steinbach, George Karypis, and Vipin Kumar, “A Comparison of Document Clustering Techniques,” *Text Mining Workshop, KDD 2000*. Boston, MA (2000).

[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).

[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).

[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).

[Tan+01] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster, Christopher Potter, Alicia Torregrosa, “Finding Spatio-Termporal Patterns in Earth Science Data: Goals, Issues and Results**,**” KDD Temporal Data Mining Workshop, KDD2001 (2001).