Clustering Earth Science Data: Goals, Issues and Results*

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

[+] Department of Computer Science and Engineering, Army HPC Research Center
University of Minnesota
{steinbac, ptan, 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 reports on recent work applying data mining to the task of finding interesting patterns in earth science data derived from global observing satellites, terrestrial observations, and ecosystem models. Patterns are "interesting" if ecosystem scientists can use them to better understand and predict changes in the global carbon cycle and climate system. The initial goal of the work reported here (which is only part of the overall project) is to use 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. Finding "meaningful" clusters requires an approach that is aware of various issues related to the spatial and temporal nature of earth science data: the "proper" measure of similarity between time series, removing seasonality from the data to allow detection of non-seasonal patterns, and the presence of spatial and temporal autocorrelation (i.e., measured values that are close in time and space tend to be highly correlated, or similar). While we have techniques to handle some of these spatio-temporal issues (e.g., removing seasonality) and some issues are not a problem (e.g., spatial autocorrelation actually helps our clustering), other issues require more study (e.g., temporal autocorrelation and its effect on time series similarity). Nonetheless, by using the K-means as our clustering algorithm and taking linear correlation as our measure of similarity between time series, we have been able to find some interesting ecosystem patterns, including some that are well known to earth scientists and some that require further investigation.

Keywords

K-means clustering, time series, earth science data, scientific data mining

1. INTRODUCTION

The project team to which we belong is a group of computer and ecosystem scientists focusing on the development of algorithms and tools to help ecologists discover changes in the global carbon cycle and climate system. These techniques will aid ecologists in their efforts to better understand global scale changes in biosphere processes and patterns, and the effects of widespread human activities, such as deforestation, biomass burning, industrialization, and urbanization. Ecologists who work at the regional and global scale have identified Net Primary Production (NPP) as a key variable for understanding the global carbon cycle and the ecological dynamics of the Earth. NPP is the net assimilation of atmospheric carbon dioxide (CO_2) into organic matter by plants. Terrestrial NPP is driven by solar radiation and can be constrained by precipitation and temperature. Keeping track of NPP is important because it includes the food source of humans and all other animals 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, CASA (the Carnegie Ames Stanford Approach [PKB99]), has been used for over a decade to produce a detailed view of terrestrial productivity.

Our project uses the multi-year output of CASA, as well as other climate variables, such as long term sea level pressure, sea surface temperature (SST) anomalies, etc., to discover interesting patterns relating changes in NPP to land surface climatology and global

* 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.

climate. Predicting NPP based on, for example, sea surface temperature, would be of great benefit given the near real-time availability of SST data and the ability of climate forecasting to anticipate SST El Nino/La Nina events. For a number of years, ecosystem scientists on our team have used traditional statistical tools for spatio-temporal data analyses relating NPP and other climate variables. Data mining [KH99] can complement these statistical tools in many ways, e.g., some of the steps of hypothesis generation and evaluation can be automated, facilitated and improved.

In this paper we report on a portion of the work involved in this project. In particular, the initial goal of the work reported here is to use clustering to divide areas of the land and ocean into disjoint regions in an automatic, but meaningful way that enables us to 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.

There are significant issues related to the spatial and temporal nature of earth science data: the "proper" measure of similarity between time series, the seasonality of the data, and the presence of spatial and temporal autocorrelation (i.e., measured values that are close in time and space tend to be highly correlated, or similar). Although sophisticated approaches to time series similarity are available, e.g., dynamic time warping, we chose standard linear correlation as our similarity measure since it works well with our clustering algorithm (K-means) and lends itself to statistical tests. Since earth science data has a very cyclical (e.g., seasonal) nature, and since earth scientists are mostly interested in non-seasonal patterns, we typically used a couple of preprocessing techniques (moving average and monthly Z-score) to remove seasonality from the data before clustering. However, these seasonality removal techniques affect the degree of temporal autocorrelation of the data (both positively and negatively), and hence, affect the "significance" of the observed correlations. On the other hand, the high degree of spatial autocorrelation of the earth science data we are analyzing actually is beneficial, allowing our K-means clustering algorithm to produce clusters consist mostly of a relatively small number of geographically contiguous regions.

The basic outline of this paper is as follows. Section 2 provides a description of the earth science data. Section 3 describes our clustering technique, which is based on K-means. Section 4 discusses related clustering work and Section 5 considers the issue of how to preprocess the data to remove seasonality patterns. Section 6 describes our initial results in applying clustering to earth science data, while section 7 is a short conclusion and an indication of future directions.

2. Earth Science Data

The earth science data for our analysis consists of global snapshots of measurement values for a number of variables (e.g., NPP, temperature, pressure and precipitation) collected for all land surfaces or water (see Figure 1). 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, and are typically available at monthly intervals that span a range of 10 to 50 years. The attribute data within a global snapshot is represented using spatial frameworks, i.e., a partitioning of the Earth's surface into a set of mutually disjoint regions which collectively cover the entire surface of Earth. For the analysis presented here, we focus on attributes measured on latitude-longitude spherical grids of different resolutions, e.g., NPP, which is available at a resolution of 0.5° x 0.5°, and sea surface temperature, which is available for a 1° x 1° grid.

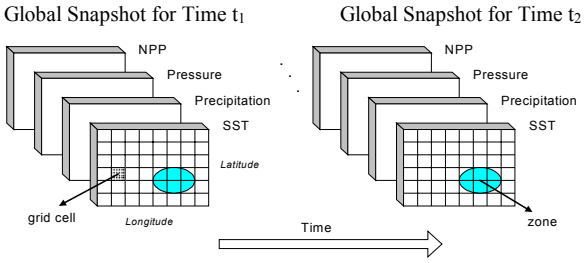

Figure 1: A simplified view of the problem domain.

Using variables derived from sensor observations, earth scientists have developed standard climate indices. These indices are useful because 1) they can distill climate variability at a regional or global scale into a single time series, 2) they are related to well-known climate phenomena such as El Nino, and 3) they are well-accepted by earth scientists. For example, various El Nino related indices, such as ANOM1+2 and ANOM4, have been established to measure sea surface temperature anomalies across different regions of the Pacific Ocean. (El Nino is the anomalous warming of the eastern tropical region of the Pacific, and has been linked to various climate phenomena such as droughts in Australia and heavy rainfall along the western coast of South America.) Some of the well-known climate indices are shown in Table 1 [IND1, IND2]. Figure 2 shows the time series for the ANOM1+2 index. Note that the peak in 1982 and 1983 corresponds to a severe El Nino event.

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
ANOM 1+2	Sea surface temperature anomalies in the region bounded by 80°W-90°W and 0°-10°S
ANOM 4	Sea surface temperature anomalies in the region bounded by 150°W-160°W and 5°S-5°N
NP	Area-weighted sea level pressure over the region 30N-65N, 160E-140W

Table 1: Description of well-known climate indices.

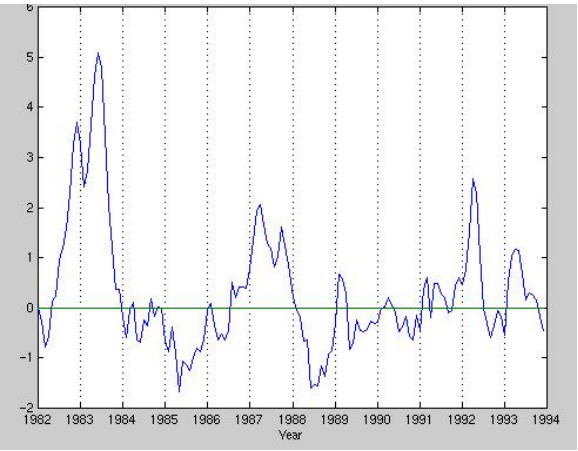

Figure 2: ANOM 1+2 time series.

3. 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, $\boldsymbol{x}$ and $\boldsymbol{y}$ 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.

4. Related Work

In this section we discuss other techniques that have recently been used to cluster earth science data. The goal is to indicate possible alternatives to K-means, and to further illustrate some of issues involved in clustering earth science data.

In [SID99], a mixture model approach is used to identify the cluster structure in atmospheric pressure data. (Mixture models assume that the data is generated probabilistically from a mixture of Gaussian distributions and use the data to estimate the parameters of these distributions.) This approach is related to K-means [Mit97], but has two advantages. First, it assigns a "membership" probability to each data point and each cluster. These probabilities provide a measure of the uncertainty in cluster membership. Second, it is sometimes possible to estimate the most appropriate choice for K [SID99]. (It is also possible to estimate the best K for K-means by plotting the overall error or similarity for different values of K and looking for the knee in the plot.)

Another possible approach to clustering, particularly in spatially oriented domains, is to use "region growing." Starting with individual points as clusters, each cluster is grouped with the most similar, physically adjacent cluster, until there is only one cluster. (Sometimes various criteria are applied to prevent clusters from being merged if the resulting cluster is too "poor.") This approach can be viewed as a form hierarchical clustering which has the constraint that clusters can only be merged if the resulting cluster is contiguous, i.e., not split into disconnected sets of points [Mur95].

However, it is sometimes desirable to have clusters that are "piecewise contiguous," i.e., consist of points which are similar, but not all in one contiguous region. An example such an approach is presented in [Til98] and was applied to the problem of land use classification based spectral image data. The technique, Recursive Hierarchical Image Segmentation, consists of alternating steps in which similar, adjacent, regions are merged (a region growing step) and similar, non-adjacent regions are merged (a spectral clustering step). For land use classification, this allows the grouping of points, which may represent the same type of land cover, but which are in disconnected regions. (The K-means approach that we use will automatically produce piecewise contiguous regions.)

Perhaps the work that is most closely related to ours is [Viv00}, which introduces ACTS (Automatic Classification of Time Series), a clustering method for remote sensing time series. (The data considered is NDVI, the Normalized Difference Vegetation Index, or greenness index [NASA].) The goal of this work was to use clustering as an initial step for deriving continental-scale to global-scale vegetation maps. After the removal of components with a period of one year or less, clustering was also used to group points that had similar patterns of inter-annual variation in NDVI. However, there was no investigation of the relationships between different regions of the land and the ocean.

While there has been considerable research into hierarchical clustering and spatial clustering [HKT01], many issues still remain. Some of the new issues of zone formation are zonal formation over time, the multi-scale nature of the data, and constrained zone formation.

5. Dealing with the Seasonality of Data

Another important task in our research work is the removal of seasonal variation from the time-series data. Mostly, earth scientists are interested in non-seasonal patterns, instead of the yearly patterns of (Spring, Summer, Fall, Winter) or (Rainy Season, Dry Season). It is not that these patterns are unimportant, but rather that they are well known, and the events of interest are deviations from the normal seasonal patterns that represent long term cycles, e.g., decadal oscillations, or trends, e.g., global warming. Given such a focus, and the strength of the seasonal patterns in the data, it is necessary to remove them to see other patterns.

There are several ways to do this and Figure 3 shows the results of applying two different types of transformations (filtering) to a particular time series of values. In particular, we focus on a sample time series for sea surface temperature. (This time series was derived from data corresponding to a ½° by ½° region

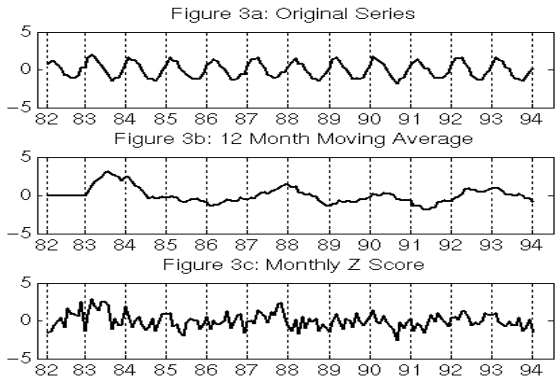

Figure 3: Effects of data pre-processing to remove seasonal variation.

of the ocean at 71.5° W, 23° S, just off the Eastern coast of South America.) This original time series, which clearly has a strong seasonal pattern, is shown by Figure 3a.

While we briefly show the effects of two different types of transformations, these issues and other time series specific issues are discussed in more detail in a related paper [Tan+01]. (Among other issues, that paper discusses the removal of seasonality based the use of DFT (Discrete Fourier Transform and SVD (singular value decomposition.) To allow all the time series to be displayed on a similar scale, all time series were standardized by subtracting off the mean and dividing by the standard deviation.

Moving average. A 12-month moving average is effective in removing seasonality and also smoothes the data significantly. However, as discussed in [Tan+01], a moving average increases the magnitudes of the observed correlations, and at the same time, makes these higher correlations less meaningful. Figure 3b shows the 12-month averaged time series.

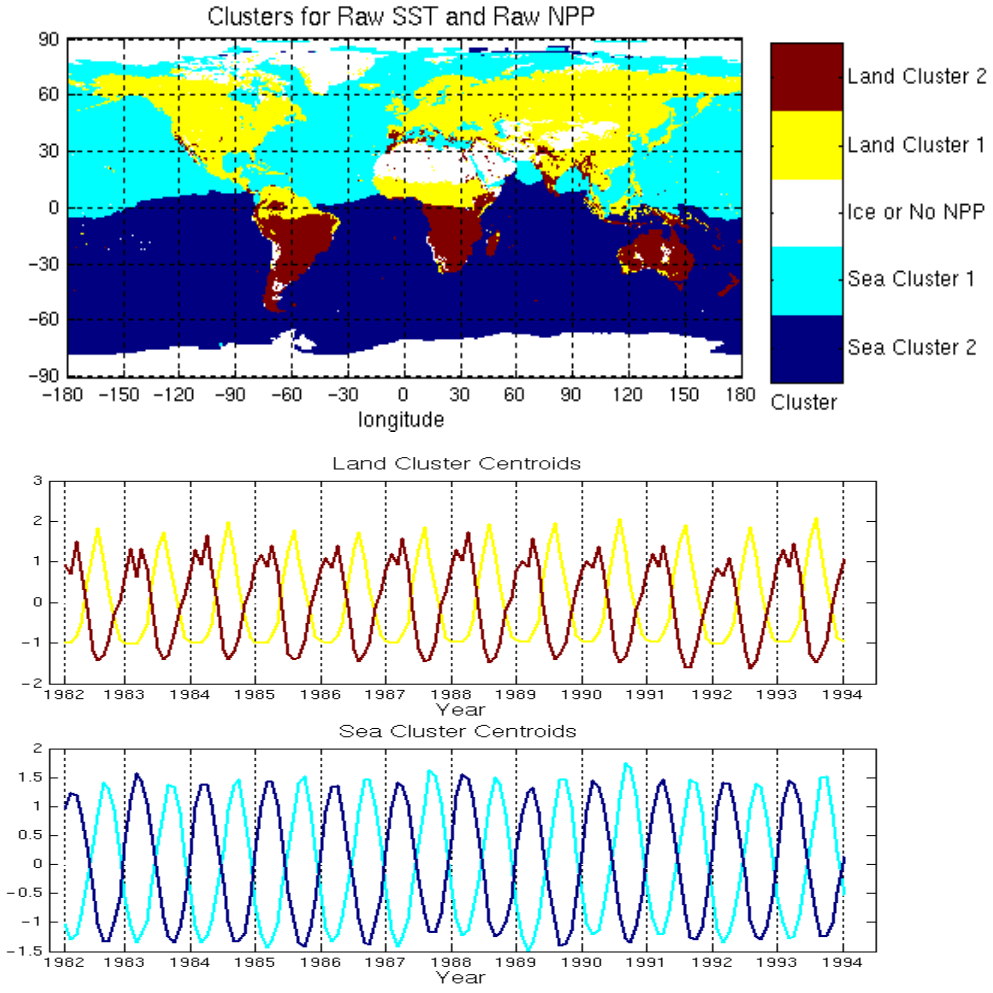

Figure 4. Two Ocean (SST) and Land (NPP) Clusters.

Monthly Z score. This transformation takes the set of values for a given month, calculates the mean and standard deviation of that set of values, and then "standardizes" the data by calculating the Z-score of each value, i.e., by subtracting off the corresponding monthly mean and dividing by the monthly standard deviation. This is slightly different from the usual statistical (Z score) standardization of subtracting the mean and dividing by the standard deviation, since each data point is standardized by using the mean and standard deviation of the values for its month, not the overall mean and standard deviation. Since it removes seasonality (but does not smooth), the monthly Z score transformation reduces autocorrelation [Tan+01]. The result of applying a monthly Z score filter is shown in Figure 3c.

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 4 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.

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

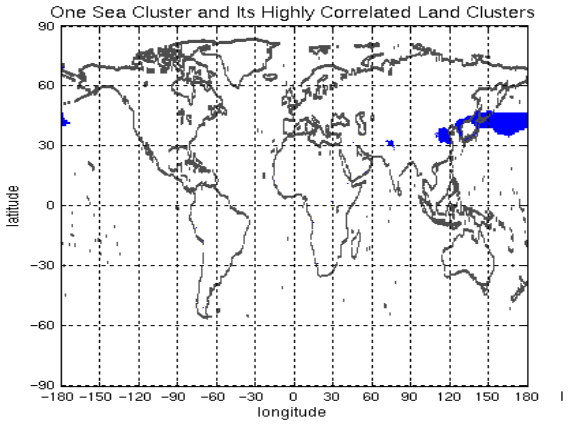

Figure 5: One Sea Cluster and Highly Correlated Land Clusters.

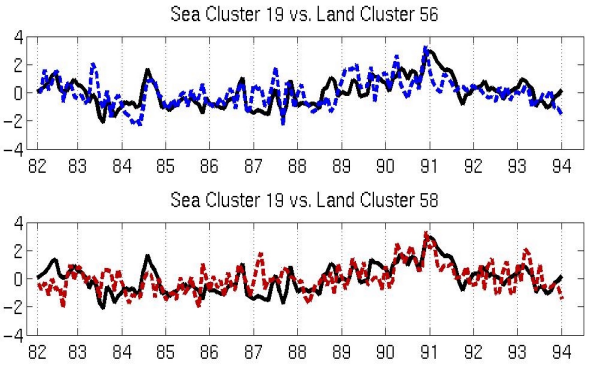

Figure 6: Comparison of Cluster Centroids.

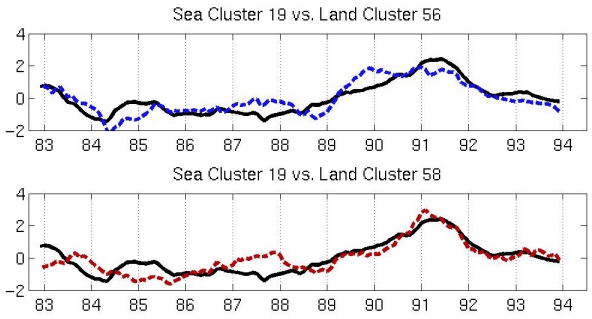

Figure 7: Comparison of Smoothed Cluster Centroids.

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 5 shows such a diagram for sea cluster 19 (which is a region of ocean off the coast of Japan) and land clusters 56 (which consists of parts of Japan and

Korea, and a region near Pakistan-northwestern India) and 58 (which consists of part of China near the coast). The NPP centroids of land clusters 56 and 58 are correlated with the SST centroid of sea cluster 19 at a level of 0.56 and 0.50, respectively. (For this analysis we removed seasonal variation by using the monthly Z score.) Figures 6 shows a plot of the centroid of sea cluster 19 versus the cluster centroids of land clusters 56 and 58. To better display the overall relationships between the centroids, Figure 7 shows the same centroids after they have been smoothed using a 12-month moving average.

Unlike the pattern that we found in the previous section, the teleconnection pattern displayed in Figure 5 between the sea region (sea cluster 19) and the land regions (land clusters 56 and 58) is not well known to ecosystem scientists. While further investigation by ecosystem scientists is needed to determine whether these relationships are meaningful or not, these clustering results have at least provided the basis for an initial hypothesis. In particular, it would be interesting to see whether the teleconnection between sea cluster 19 and the region near Pakistan-northwestern India can be verified, since these regions are far apart.

Sea cluster 19 is highly correlated (-0.77), with one of the ocean indices, PDO, which is a long-lived El Niño-like pattern of Pacific climate variability. The new hypothesis suggested by this apparent teleconnection is that ENSO (El Nino Southern Oscillation) influences NPP in the Pakistan-India region through variations in seasonal rainfall patterns. This type of El Nino association with rainfall has been noted before for the Indian subcontinent. As the mean sea level pressure difference between the south central Pacific (e.g. Tahiti) and the Indian Ocean weakens, the trade winds can relax, monsoons become weaker, and there can be strong drought in India and Australia. This relationship was noted as far back as 1904 by Sir Gilbert Walker, a British mathematician serving the British Colonial Service. However, the monsoonal teleconnection pattern to ENSO events has not been consistently strong in recent times, (see [KRC99]), which means that more work is required on our part to better understand the patterns shown in Fig 5.

Finding Correlations between Land Clusters and (Ocean) Climate Indices. We also investigated the land-ocean connection by using climate indices that are based on the SST or pressure differences, either between two points on the ocean or over an area of the ocean (see Table 1). For example, some of the indices relate to the El Nino effect. These indices are also time series and thus, we can find the clusters on the land and sea that display a strong correlation to a particular index. Figure 8 shows the

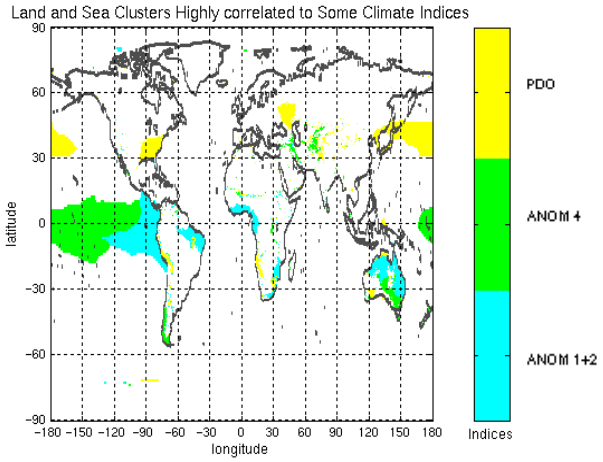

Figure 8: Clusters that are Highly Correlated with Climate Indices

land and sea clusters that correlate highly (positive or negative correlation of 0.5 or above) to three different climate indices: PDO (Pacific Decadal oscillation) and two El Nino indices, ANOM 4 and ANOM 1+2 [IND1, IND2]. For this analysis we removed seasonal variation by using the monthly Z score. The ocean regions that are highly correlated with the two El Nino indices are related to the regions used to define the two indices.

To illustrate the potential for clustering to find interesting teleconnections between land and ocean regions, note that there is a land cluster near Zimbabwe, in southern Africa, which is highly correlated to the ANOM 1+2 index. A connection between southern African rainfall and the El Nino phenomenon has been observed. For instance, Ropelewski and Halpert [RH96] have shown a positive correlation between the southern Oscillation Index (SOI) (another El Nino related climate index) and southern African rainfall. More specifically, the droughts which have occurred in southern Africa since the end of the 1960s are associated with warmer temperatures in the eastern and central tropical Pacific, in the tropical Indian Ocean, and in the equatorial Atlantic. The spatial structure of these anomalies may be associated with El Nino/La Nina events.

7. Conclusions

A key conclusion of this paper is 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 4. 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 this effort, we have focused on climate indices, which are time series of temperature or

pressure that correlate well with certain regions of the ocean from which they are derived. In particular, we have looked at which regions of the land are most highly correlated to these centroids. 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.

To produce meaningful clusters it is necessary to take into account the spatio-termporal nature of the data. Seasonality must be removed by using appropriate pre-processing steps if non-seasonal patterns are to be detected, and there are significant issues concerning what levels of correlation between time series indicate significant connections. However, on positive side, it is likely that the simple K-means clustering approach we are using works as well as it does because of the high level of spatial auto-correlation in the data. Otherwise, the clusters produced by K-means might consist of a large number of widely separated small regions. The use of clusters that are only piecewise contiguous has not been a problem so far, although much of the evaluation proceeds via visualization and people are good at noticing interesting patterns and ignoring noise. The chief insights come when the clusters consist mostly of large, coherent areas, although, in such cases, the exceptions to the rules can also be interesting as with the case of Figure 4 and southern California.

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 mentioned in Section 4 or variants of K-means, e.g., bisecting K-means [SKK00]. Along somewhat different line, 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 approach result from the fact that often, only extreme events that 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 clustering 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. Nonetheless, we are hopeful that our clustering approach, and any improvements that we make to it, will continue to produce interesting and useful results.

REFERENCES

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

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

[HKT01] J. Han, M. Kamber and K. H. Tung, "Spatial Clustering Methods in Data Mining: A Survey", Harvey J. Miller and Jiawei Han (eds.) (2001), *Geographic Data Mining and Knowledge Discovery,* Taylor and Francis, forthcoming (expected 2001).

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

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

[KH99] M. Kamber, and J. Han, *Data Mining: Concepts & Techniques*, Morgan Kaufmann (1999).

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

[KRC99] K. K. Kumar, B. Rajagopalan, and M. A. Cane, "On the weakening relationship between the Indian monsoon and ENSO," *Science*, 284, 2156-2159 (1999).

[Mit97] Tom Mitchell, *Machine Learning*, McGraw Hill (1997).

[Mur95] F. Murtagh, "Contiguity-constrained hierarchical clustering," In I.J. Cox, P. Hansen and B. Julesz, eds., *Partitioning Data Sets*, DIMACS, AMS, 143-152 (1995).

[NASA] http://earthobservatory.nasa.gov/Library/

[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).

[SIG99] Padhraic Smyth, K. Ide, and M. Ghil, "Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering," *Journal of Atmospheric Science*, 56, 3704-3723 (1999).

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

[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," Submitted to KDD Temporal Data Mining Workshop, KDD2001 (2001).

[Til98] J. C. Tilton, "Image Segmentation by Region Growing and Spectral Clustering with a Natural Convergence Criterion," *Proc. of the 1998 International Geoscience and Remote Sensing Symposium (IGARSS '98)*, Seattle, WA (1998).

[Viv00] N. Vivoy, "Automatic Classification of Time Series (ACTS): a new clustering method for remote sensing time series," *International Journal of Remote Sensing* (2000)