DETECTING REGIMES IN TEMPERATURE TIME SERIES

PATRICK J. CLEMINS
Dept. of Electrical and Computer Eng.
Marquette University
Milwaukee, WI

RICHARD J. POVINELLI
Dept. of Electrical and Computer Eng.
Marquette University
Milwaukee, WI

ABSTRACT

In the field of climate prediction, regimes are used to model long-term cyclic trends. Although air pressure regimes have been discovered, there has been little exploration into the possibility of temperature regimes. This paper develops an approach to finding regimes in a temperature time series. First, the time series is reconstructed in a phase space. Then, a clustering algorithm is used to search the phase space for clusters. Finally, the number of transitions between these clusters is recorded. A low ratio between the number of transitions into the cluster and the number of points in the cluster indicates that a regime structure is present. Results are given for various temperature time series.

INTRODUCTION

The Earth's climate is a complex system with an undetermined number of variables. Many long-term prediction models have been proposed, but most are based on the assumption that the earth's climate is a linear system. The sentiment in the field of meteorology seems to be that linear models are accurate enough for prediction even though evidence has been uncovered that seems to indicate some nonlinear trends in the Earth's climate (Palmer, 1993). However, there has not been much research on these nonlinear climatic trends, and these trends may provide insight into climate prediction. If regimes could be found in the earth's climate, they could be used to help predict future climatic trends.

One technique used to expose patterns in a nonlinear time series is to reconstruct the time series in a phase space (Povinelli, 1999). A phase space is constructed by creating a vector space $[s(t)\ s^{(1)}(t)\ s^{(2)}(t)\ \ldots\ s^{(n)}(t)\ ]$ where n+1 is the embedding dimension and $s^{(x)}(t)$ is a time delayed s(t) or a time series of a related system parameter. Appropriate time delays can be determined by various statistical methods such as autocorrelation or auto mutual information (Abarbanel, 1996, Kantz, 1997).

An example of a non-linear system is the Lorenz system (Figure 1). The Lorenz system was developed to model atmospheric flow. The system is defined by three differential equations.

$$\begin{aligned} dX &= -aX + aY \\ dY &= -XZ + rX - Y \\ dZ &= XY - bZ \end{aligned} \tag{1}$$

Two regimes, or states, are easily discernable in the Lorenz system. Once the system is in one regime, it tends to stay in that regime. The area between the two regimes

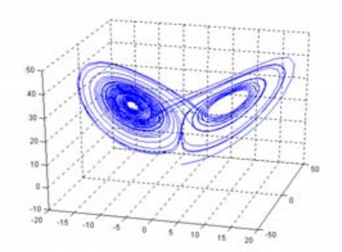


Figure 1 – Lorenz System

is called the unstable region, because once the system enters this region it has a higher likelihood of switching regimes. For the purpose of this paper, a regime is defined as a cluster in phase space where the trajectory of the original time series tends to stay for a period of time. To determine whether a regime exists, the transitions out of each cluster are compared to the number of intra-cluster transitions. If the number of intra-cluster transitions is large compared to the number of transitions out of the cluster, then a regime is present.

Research has shown that regime structures are present in certain parts of the Earth's climate (Palmer, 1999). Some parts of the globe follow the traditional linear model, while other areas, such as the northern hemisphere during the winter, tend to follow a regime structure. The El Nino cycle, which has two quasi-equilibrium states, also exhibits non-linear regime structure. Here, we propose a method for detecting temperature regimes using temperature time series phase spaces.

To detect these regimes and count transitions between them, the phase space must be clustered. The general idea of clustering is to group similar objects together. In the context of this experiment, an object is a point in phase space. Euclidean distance is used to quantify two data point's similarity.

In this paper, the shape, size and number of the clusters is unknown. Density-based and hierarchical clustering techniques such as OPTICS (Ankerst, 1999), and Chameleon (Karypis, 1999) have been shown to work well on such data sets. However, these algorithms are not simple to implement and are unnecessarily complex for this research. Therefore, ideas from both of these algorithms were used to create a simple clustering algorithm that can be used to computationally find clusters given the appropriate parameters.

METHOD

To show that a nonlinear structure is present in the earth's climate, a phase space is constructed using temperature time series. If the temperature time series are nonrandom, a phase space plot should show clusters of points, or the points forming a line as the time series is plotted. However, if the time series are random, the phase space would simply be a random scattering of data points.

A clustering algorithm is formulated from K-means (step 3) and hierarchical (step 4) clustering techniques. A visual summary of the algorithm is in figure 2. The algorithm is as follows:

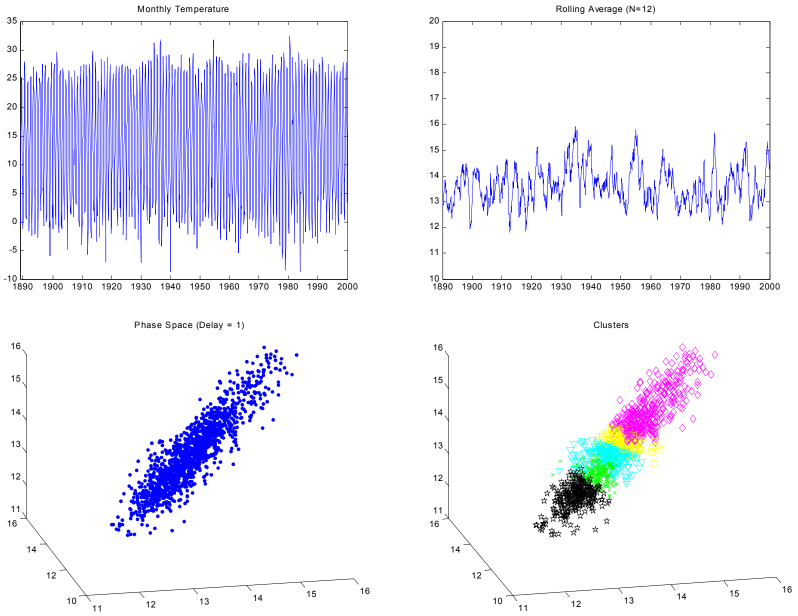


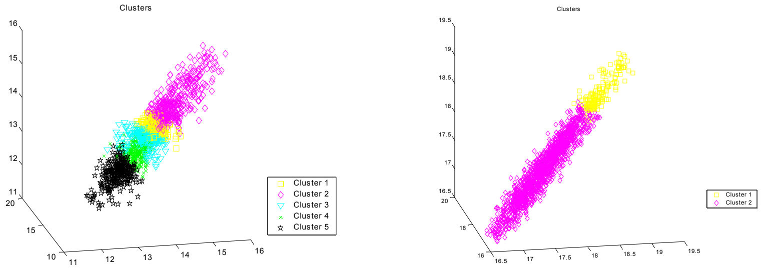
Figure 2 – Methodology

Upper right – Original time series (Wichita, KS); Upper left – Rolling average using 12 months; Lower right – Phase space; Lower left – Discovered clusters (5 clusters found)

1) Preprocess data.
   a) Rolling average, yearly average, etc.
2) Construct phase space.
3) Do an initial K-means binary split clustering of the phase space.
   a) Number of clusters is user-defined (.1*N is a good choice, where N is number of data points)
   b) Remove empty clusters in each iteration (random initialization may give a cluster center in an empty part of the phase space)
4) For each pair of clusters i and j, merge them if there are any two points between the clusters that have a Euclidean distance less than the linear density of cluster i times a scaling factor (values between 3 and 4 worked well). The linear density is defined by:

$$(D_{XY}) / N_i \tag{2}$$

   where D_{XY} is the maximum Euclidean distance between any two points X and Y, and N_i is the number of points in cluster i. This is equivalent to saying merge the clusters if they are close with respect to their densities.
5) Repeat 4 until no more clusters can be merged.
6) Postprocessing of clusters
   a) If a cluster only transitions to one other cluster, merge the clusters



	1	2	3	4	5
1	51.7%	25.5%	20.0%	2.8%	0.0%
2	8.2%	90.6%	1.3%	0.0%	0.0%
3	17.2%	0.5%	57.9%	23.5%	0.9%
4	0.0%	0.0%	16.5%	70.3%	13.3%
5	0.0%	0.0%	4.6%	9.6%	85.7%

	1	2
1	89.2%	10.8%
2	1.2%	98.8%

Figure 3 – Results

Left – Wichita, Kansas; Right – Sydney, Australia

b) Merge small clusters with appropriate bigger clusters (i.e. merge with the one they transition to the most)

After the clusters are identified, the number of transitions between each cluster is counted so that the number of times the temperature series exits each regime can be determined.

DATASETS

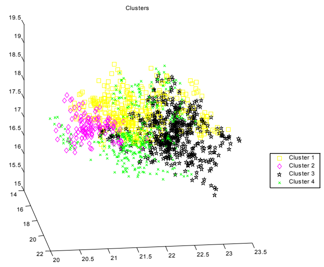
Two kinds of datasets are analyzed in this paper, modern and paleoclimatic. The modern temperature data consists of average monthly temperature readings for about the last 150 years. These datasets were extracted from the Global Historical Climatology Network (GHCN) database that is available online (2000). The GHCN database contains readings from thousands of weather stations around the globe. The biggest disadvantage with this type of dataset is that it is rather short. Some climatic trends (such as an ice age) last for thousands of years and take hundreds of years to develop. Therefore, this type of dataset may not contain enough points for long-term regime structures to be identified. This dataset does have the advantage that all measurements are accurate.

Paleoclimatic datasets contain estimated temperatures over the last few hundred years. These estimates are calculated from ice core samples, tree rings, and other geologic temperature indicators. The biggest disadvantage with these datasets is that the data points are estimates. The advantage with these data sets is that there are enough data points for longer regimes to become apparent.

RESULTS

The regime structure of many different time series was explored. An examination of the GHCN data from Wichita, KN will be discussed first. This time series has average monthly temperatures from 1989 to 1999. After taking a moving average (N=12), a three-dimensional phase space was constructed using time delays of 1 and 2. These time delays were chosen because upon inspection of most of the temperature time series, the

	1	2	3	4
1	50.9%	28.2%	1.2%	19.6%
2	27.2%	71.8%	0.0%	1.0%
3	1.3%	0.0%	70.7%	28.0%
4	17.4%	8.7%	14.5%	59.4%



	1	2	3	4
1	30.4%	10.7%	27.7%	31.3%
2	15.3%	34.7%	31.9%	18.1%
3	19.4%	14.2%	41.9%	24.5%
4	16.9%	6.1%	16.8%	59.8%

	1	2	3	4
1	75.4%	1.9%	15.7%	6.9%
2	6.0%	78.9%	2.3%	12.8%
3	10.7%	0.6%	81.3%	7.4%
4	8.3%	5.1%	8.6%	78.0%

Figure 4 – Results

Left – Mann Reconstruction (3D at top, 8D at bottom); Right – Multiple Sites

autocorrelation function decreased linearly with increasing τ. This phase space was then clustered using 256 initial clusters and a combination threshold of 3. The algorithm detected five clusters. The transition matrix and final clustering are shown in figure 3.

The transition matrix is formed by counting the number of transitions between each cluster. The rows represent the clusters the time series is transitioning from and the columns represent the clusters the time series is transitioning to. The percentage of transitions from cluster i to cluster j is given by the value in row i, column j. From the transition matrix for the Wichita time series, it appears that a regime structure is present because once the time series enters a phase space cluster, especially cluster 2, there is a high probability that it will stay in that cluster.

The next GHCN temperature station analyzed is Sydney, Australia. This time series consists of average monthly temperatures from 1859 to 1991. A moving average (N=12) was applied to the time series and time delays of 1 and 2 were chosen for the phase space embedding. The number of initial clusters was set at 256 and a combination threshold of 3.5 was used. The clusters found by the algorithm and transition matrix are in figure 3. This time series, although it has fewer clusters than the Wichita time series, also shows a very prominent regime structure with a very low inter-cluster transition probability.

The next dataset presented is a paleoclimatic reconstruction created by Mann et al (1998). A three-dimensional phase space was constructed using time delays of 1 and 2. The transition matrix is shown in figure 4. The initial number of clusters was set at 64 and a combination threshold of 3 was used.

An eight-dimensional phase space was also considered for this time series. The eight-dimensional phase space was constructed using time delays of 1 through 7. The phase space was clustered using 128 initial clusters and a combination threshold of 3.5. The transition matrix after the clustering is in figure 4.

The Mann reconstructed time series does not show as prominent a regime structure as some of the other time series reviewed here. This might be because a moving average was not applied to the time series before analysis to remove some of the high frequency components of the time series. It is also interesting to note that the algorithm did a much better job of finding prominent regimes in the three-dimensional phase space than in the eight dimensional phase space.

The final experiment performed was using time series from different locations to construct a phase space. Average monthly temperatures from Corpus Christi, TX, El Paso, TX, and Fresno, CA were used to construct a three-dimensional phase space. Temperature data was available from all three sites for the years 1888 – 1998. A moving average (N=12) was calculated for all three time series before the phase space was constructed. The initial number of clusters was set at 256 and the combination threshold was set to 4. The clusters determined by the algorithm and transition matrix are shown in figure 4. The clusters found for this phase space are rather complex and bend around each other. There appears to be a strong regime structure at work in these time series as well because of the low percentages of inter-cluster transitions. This was the expected result for this trial because all three cities are located within the effect of the El Nino cycle.

CONCLUSION

Our analysis shows that regimes can be detected in most temperature time series. However, the strength of the regimes varies greatly. These results are expected however since previous research has shown that only certain parts of the earth have a regime-dominated climate. Regimes were found in various phase space constructions with different types of data preprocessing and different phase space construction techniques.

One thing to explore in the future is to analyze the duration of each visit to a cluster to see if these durations are consistent in anyway. Also, the clustering algorithm could be improved to include some kind of transition minimization instead of relying mostly on the density characteristics of the phase space.

REFERENCES

Abarbanel, Henry D., 1996, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, New York.

Ankerst, Mihael, Breunig, Markus M., Kriegel, Hans-Peter, Sander, Jörg, 1999, "OPTICS: Ordering Points to Identify the Clustering Structure", *Proc. ACM SIGMOD'99 Int. Conf. On Management of Data*, Philadelphia, PA.

Kantz, Holger, 1997, *Nonlinear Time Series Analysis*, Cambridge University Press, New York, New York.

Karypis, George, Han, Eui-Hong, Kumar, Vipin, 1999, "Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling", *Computer,* Vol. 32, No. 8, pp. 68-75.

National Climatic Data Center, 2000, *Global Historical Climatology Network Version 2 Dataset*, ftp://www.ncdc.noaa.gov/pub/data/ghcn/v2/.

Mann, Michael E., Bradley, Raymond S., Hughes, Malcolm K., 1998, "Global Six Century Temperature Patterns", *IGBP PAGES/World Data Center-A for Paleoclimatology Data Contribution Series # 1998-016*, NOAA/NGDC Paleoclimatology Program, Boulder, CO.

Palmer, T. N., 1993, "A nonlinear dynamical perspective on climate change", *Weather*, Vol. 48, pp. 314-326.

Palmer, T. N., 1999, "A nonlinear dynamical perspective on climate prediction", *Journal of Climate*, Vol. 12, pp. 575-591.

Povinelli, Richard J., 1999, "Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events", Ph.D. Dissertation, Marquette University, Milwaukee, WI.