Statistical Analysis and Visualization of Time-Varying Multivariate Climate Data Sets

Jeffrey Sukharev Chaoli Wang Kwan-Liu Ma

University of California, Davis

Abstract

We present a statistical approach to study time-varying, multivariate climate data sets. Aided by domain expertise from the NOAA scientists, we have developed a solution for correlation analysis of multivariate spatial-temporal climate data sets.

1 Introduction

We present a statistical approach to study the climate data sets provided by the scientists from the National Oceanic and Atmospheric Administration (NOAA). Accurate climate forecast for an extended period of time into the future has become a very important problem with far-ranging applications. Leveraging the statistical information of data, we have designed a visualization system that provides the scientists with insight into their data. Part of the excitement of discovering the implications of climate change is the fact that it involves so many different processes linked together into an intricate web. The challenge of discovering which processes matter the most involves not only knowing how a given perturbation may disrupt the climate but also how different time scales affect the analysis of change. Deciding what changes matter the most, and when and where they occur, requires evaluation of how they are related to one another [1].

2 Problem Statement

We held a number of exchanges with the NOAA scientists. It is our understanding that the scientists are severely limited by the visualization tools that are currently in existence. Throughout this paper, we use the El Niño phenomenon as an example to explain the challenge that the scientists face in their data analysis and visualization. The El Niño phenomenon can only be found when the analysis takes into consideration the correlation among multiple variables over multiple time steps. Due to the complexity of the data in the spatial, temporal, and variable domain, the scientists are much more interested in visualization tools that allow them to explore the multifaceted nature of their data.

3 Our Approach

We present a visualization system customized to address the needs raised by the scientists. Our system features an integrated user interface that allows the users to examine their data in 3D (volume rendering) and 2D (slices) spatial views, and variable views simultaneously. The system also includes a correlation analysis function which produces clusters that partition the data into separate regions based on the similarity of temporal behaviors. Our approach includes the use of parallel coordinate for the variable view, and a suite of techniques (i.e., time-activity curves, principal component analysis, and graph partitioning) for temporal activity study.

3.1 Parallel Coordinate

Dealing with multivariate data requires an intuitive way to visualize correlations between multiple variables. We have chosen to use parallel coordinates because it is a widely-used solution for displaying and detecting relationships among multiple variables. The parallel coordinate view has been integrated into our system and linked with other views of the data.

3.2 Time-Activity Curves

To study the temporal aspect of the climate data, we utilized the time-activity curve (TAC). The basic idea of TACs is to treat each voxel in the volume as a temporal function, and the source of this temporal behavior varies with a particular modality [2]. If we could compare TACs of different voxels and classify them into clusters based on their similarity, we are able to gain insights as to how different regions of the data change over time. This knowledge would add to the understanding of the climate data, which is not readily available to the scientists.

The TAC collection of selected voxels (e.g., the use can select a 2D region from a data slice) is represented in a high-dimensional space (when the number of time steps is large). Therefore, dimension reduction is needed for further analysis. One way for dimension reduction is to derive a single variable from each TAC. For instance, we can use the variance of the TAC curve for every individual voxel. Another way to reduce dimension is to use the principal component analysis method.

3.3 Principal Component Analysis

Principal component analysis (PCA) is a powerful tool for deriving the dominant patterns in a statistical field (e.g., a random vector, usually indexed by location in space). PCA can be used to display the data as a linear projection from the original data space to a subspace that best captures the variances of data. In order to faithfully represent the data in the low-dimensional space, it is preferable that 90% of the data variances are mapped on the first two principal components. Our test results on the climate data set show that it is ideally suitable for PCA dimension reduction, as the first principal component already describes 80-90% of data variances.

3.4 Graph Partitioning

After dimension reduction, a suitable clustering method is needed so that we can take distances between the data points in the reduced dimensions and generate clusters accordingly. In this paper, we use the normalized cut algorithm from image segmentation literature by treating data after dimension reduction (using either PCA or derived statistics) as images. We have decided to use the normalized cut due to its ability to find perceptually significant groups first before detecting smaller, less significant groups. The normalized cut takes three parameters as input: the image itself, the desired number of clusters, and the distances between image data points. We calculate these distances using two metrics: Euclidian and Manhattan.

The normalized cut, introduced by Shi and Malik [4], is graph partitioning method that breaks a graph into segments. The algorithm represents the input image as a fully connected graph where every pixel has a link to every other pixel. It was designed to overcome outliers. Instead of looking at the value of total edge weight connecting the two partitions A and B ($A \bigcup B = Q$), the method computes the cut cost as a fraction of the total edge connections to all nodes:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,Q)} + \frac{cut(A,B)}{assoc(B,Q)} \tag{1}$$

where

$$assoc(A,Q) = \sum_{a,q} w(a,q), \quad assoc(B,Q) = \sum_{b,q} w(b,q) \tag{2}$$

$$a \in A, \quad b \in B, \quad q \in Q$$

Assuming the size of the input image is $n \times m$, the product of TAC or TAC variances, can be represented as 1D vector Q of size $N = n \times m$. We compute the weight matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$, where $\mathbf{W}(i,j)$ represents relationship between points i and j in Q. Given the weight matrix $\mathbf{W}$ and the number of clusters c, we compute the degree matrix $\mathbf{D} = Diag(\mathbf{W}_{1N})$, where $\mathbf{W}_{1N} \in \mathbb{R}^N$ and each element is the sum of the corresponding rows in $\mathbf{W}$.

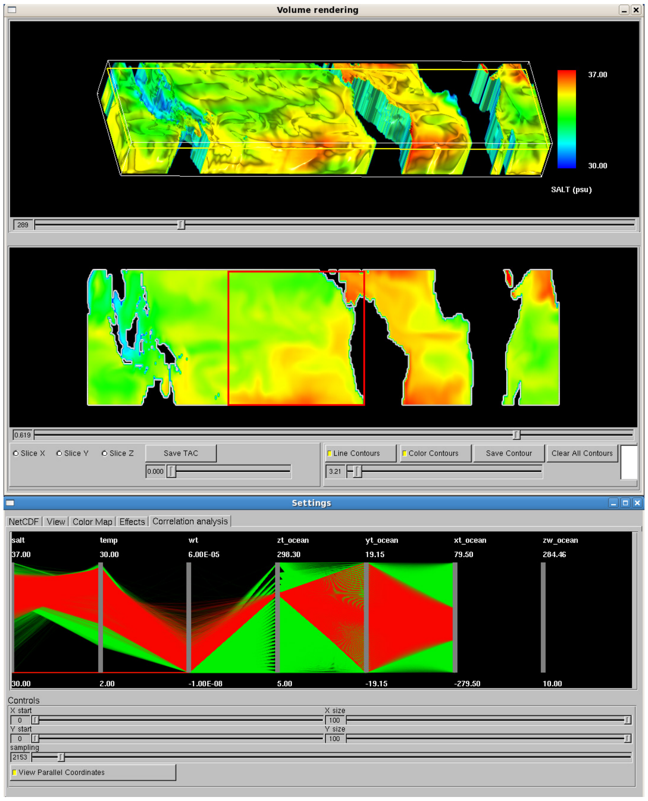
We then find the optimal eigensolution $\mathbf{Z}^*$ by solving the leading c eigenvectors using the standard eigensolver:

$$\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-\frac{1}{2}}\mathbf{v} = \lambda\mathbf{v} \tag{3}$$

$$\mathbf{Z}^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}_{[c]} \tag{4}$$

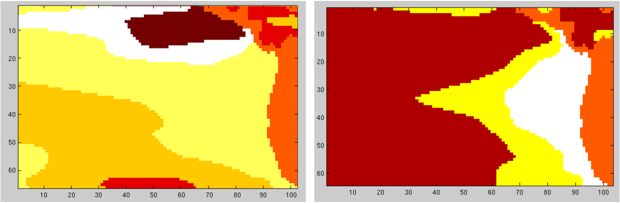
where $\mathbf{v}$ is the eigenvector and λ is the eigenvalue. The clustering results can be displayed in the principal component space, or they can be displayed directly on the regions selected.

4 Results



X start 0
Y start 0
X size 100
Y size 100
sampling 2153

Figure 1: Our user interface consists of a volume renderer (top), a slice viewer (middle), and a variable viewer using parallel coordinate (bottom). The salinity variable is used in the rendering.



(a) salinity (b) temperature

Figure 2: The clustering results corresponding to the 2D region selected in the slice viewer in Figure 1 over 72 time steps (a six-year span). Six and four clusters denoted by different colors are shown in (a) and (b) respectively.

Figure 1 shows a screen shot of our user interface. The volume renderer provides a 3D view of the data. The user can select a slice of the data and the result is displayed in the slice viewer. In Figure 1, we also show the selected region on the slice and their corresponding voxel values highlighted in red in the variable view. Figure 2 shows the clustering results with two different variables: salinity and temperature. The TACs include 72 time steps (a six-year span) and the derived clusters help the scientists better understand their data. For example, the cluster boundaries shown in Figure 2 (b) indicate a change of temperature activity towards the west from the orange cluster (corresponding to a part of the South America), which is related to the El Niño phenomenon. Such a trend is also detectable for the salinity variable when we increase the number of clusters, as shown in Figure 2 (a).

5 Conclusion and Future Work

In this paper we presented a study on how visualization can be used to analyze complex climate data. In addition to presenting scientific data in a way that makes it easier for visual analysis, we also provided a correlation analysis tool. This tool, while still a work in progress, enables the scientists to get an overview of their data and to have a better understanding as to which regions of the data change according to similar temporal patterns. This new knowledge has been derived from the original data set based on TACs, PCA, and other statistical information.

Our current implementation of the normalized cut algorithm relies on `MatLab`, which has its strict memory limitation. In the future, we will reimplement it in C++ so that larger data can be processed. We also plan to incorporate the Nyström method [3] into our normalized cut implementation in order to significantly reduce the algorithm's memory requirements. Performance speed up can also be sought using the GPU implementation.

Acknowledgements

We thank Andrew Wittenberg from NOAA for providing the climate data set and the discussion of the El Niño phenomenon.

References

[1] W. J. Burroughs. *Climate Change: A Multidisciplinary Approach*. Cambridge University Press, second edition, 2007.

[2] Z. Fang, T. Möller, G. Hamarneh, and A. Celler. Visualization and exploration of time-varying medical image data sets. In *Proceedings of Graphics Interface 2007*, pages 281–288, 2007.

[3] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.