The Application of Clustering to Earth Science Data: Progress and Challenges

Michael Steinbach* Pang-Ning Tan† Shyam Boriah*
Vipin Kumar* Steven Klooster‡ Christopher Potter§

1 Introduction

The work described in this paper was conducted as part of the NASA funded project, *Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining*, which was part of the Intelligent Systems (NRA2-37143) program. The goal of this project was to better understand global scale patterns in biosphere processes, especially relationships between the global carbon cycle and the climate system. During this project, we developed new data analysis and knowledge discovery techniques to investigate changes in the global carbon cycle and climate system. This research has resulted in numerous joint publications in archival journals and major conferences [4, 10, 17–21, 23–28, 31–34], as well as two NASA press releases [14, 15].

More specifically, in this paper, we describe a novel clustering technique that we developed to identify regions of uniform behavior in spatio-temporal data. The clusters produced by this method are useful in discovering climate indices[1] because they identify significant regions of the ocean or atmosphere where the behavior is relatively uniform over the entire area. Some of the discovered clusters correspond to known climate indices, while other clusters are variants of known indices that appear to provide better predictive power for some land areas, and still other clusters may represent potentially new Earth science phenomena. Although this application of clustering to Earth science data has proven useful, many challenges remain. After a quick description of the data and our clustering work, we briefly describe one of these challenges, namely, the need for clusters that can represent dynamic phenomena such as those associated with climate indices.

2 Earth Science Data

The types of data shown in Figure 1 are representative of the data considered in this project, i.e., the basic data elements are individual co-registered cells in grids that cover the entire surface of the earth with resolutions between 0.25 km and 50 km. (Land

*Department of Computer Science and Engineering, University of Minnesota, {steinbac, sboriah, kumar}@cs.umn.edu.

†Department of Computer Science and Engineering, Michigan State University, ptan@cse.msu.edu

‡California State University, Monterey Bay, sklooster@gaia.arc.nasa.gov

§NASA Ames Research Center, cpotter@mail.arc.nasa.gov

[1] To analyze the effect of the oceans and atmosphere on land climate, Earth Scientists have developed **climate indices**, which are time series that summarize the behavior of selected regions of the Earth's oceans and atmosphere.

variables derived from EOS satellite data are available at resolutions as high as 0.25 km, while surface climatology data, e.g., temperature and precipitation, are only available for grids of resolution 50 km or greater.) At a particular moment in time, each grid point can be described by the values of different variables, e.g., Net Primary Production (NPP), temperature, precipitation, etc. The variable values for each grid point are available for periodic, discrete points in time with resolutions from 8 days to 1 month. These variable values can either be the result of observations (from satellites or other sources) or the result of model predictions, such as NPP predictions from the NASA-CASA model [22].

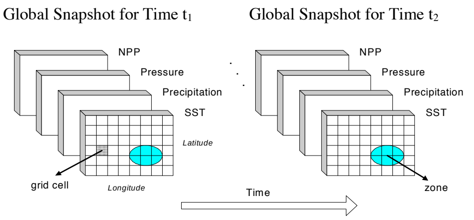

Figure 1: A simplified view of the problem domain.

3 Progress: Clustering to Discover Climate Indices

Our interest in climate indices [9] arises from a desire to improve our understanding of teleconnections involving ocean temperature/pressure and terrestrial carbon flux. In the past, Earth scientists have used observation and, more recently, eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices [29]. These techniques are only useful for finding a few of the strongest signals and impose a condition that all discovered signals must be orthogonal to each other. We have developed an alternative methodology [26–28] for the discovery of climate indices that overcomes these limitations and is based on clusters that represent geographic regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of these geographical areas.

Figure 2 shows the clusters produced by shared nearest neighbor (SNN) clustering [3] of sea level pressure data for the period 1958-1998 [26–28]. Many pairs of clusters in this clustering are highly correlated with the known climate indices. For example, clusters 13 and 20 are highly correlated with the Southern Oscillation Index (SOI), clusters 10 and 18 are correlated with the Arctic Oscillation index (AO), and clusters 7 and 10 are correlated with the North Atlantic Oscillation index (NAO).

We have also investigated clusters of SST. Four of these clusters are very highly correlated (correlation > 0.9) with well-known climate indices, e.g., NINO 1+2, NINO 3, NINO 3.4, and NINO 4, and were located in approximately the same location as where these indices are defined [26–28]. The SST clusters that are less well correlated with known indices may represent new Earth science phenomena or weaker versions or variations of known phenomena. Indeed, some of these cluster centroids provide better coverage, i.e., higher correlation to land temperature, for some areas of the land. This is illustrated in Figure 3, which compares the El Nino indices to that of cluster 62 (close to Brazil). Areas of yellow indicate where cluster 62 has higher correlation, while areas of blue indicate where the El Nino indices have higher correlation. Observe that cluster 62 outperforms the known indices for some areas of the land. The overall coverage of cluster 62 (measured in area weighted correlation) is similar to that of an El Nino based index, such as NINO 1+2, NINO 3, etc.

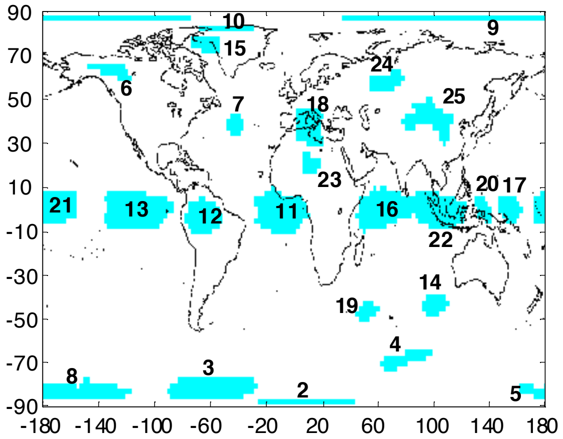

Figure 2: 25 ocean clusters produced by SNN clustering of sea level pressure data for 1958-1998.

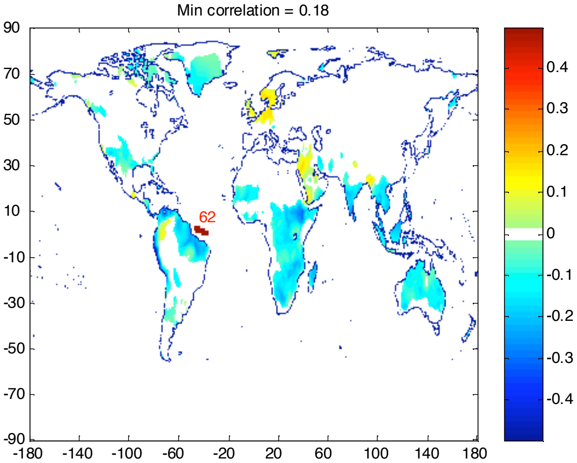

Figure 3: Comparison of correlation of Cluster 62 vs. El Nino Indices to land temperature.

4 Challenge: Dynamic Clusters

Most of the well known climate indices are based on meteorological data collected at fixed land stations. For example, NAO [7], which refers to swings in the atmospheric sea level pressure difference between the Arctic and the subtropical Atlantic, is computed as the normalized difference between SLP at a pair of land stations in these two regions of the North Atlantic Ocean. However, the phenomenon underlying NAO occurs at irregular intervals, and the exact location at which the phenomenon occurs varies over time. Specifically, Portis et al [16] showed that the high and low pressure fields in the North Atlantic are mobile from month to month. (See Figure 4, where the locations of the high/low pressure fields are marked by month and the land stations are marked by stars.) Therefore, a given land station may not always be in the right location to collect data. A climate phenomenon can be captured much more accurately by having a notion of a dynamic cluster that represents a dynamic phenomenon, based on the satellite data for the entire region.

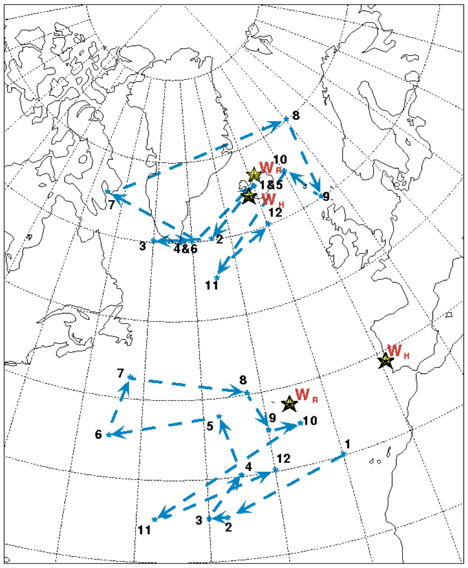

Figure 4: Mobile NAO

To extend the modeling of climate phenomena using clustering, it is necessary for us to investigate novel clustering approaches that can find clusters over different time periods and associate corresponding clusters from the different clusterings in order to determine how different "phases" of a dynamic cluster might move in space and/or expand or contract in extent. The successful development of techniques for dynamic clustering will provide better insight as to how climate indices and their impact on land climate change over time.

Previous Work Some spatio-temporal clustering approaches view the data as a collection of global snapshots taken at different time periods [5, 6, 11]. For each snapshot, spatial clustering [2, 13] is performed to group together regions that have similar values in their domain attributes. Because the data is clustered independently at each snapshot, one potential limitation of this approach is the lack of contiguity between clusters found at different time periods. Visualization techniques are therefore needed to aid the interpretation of clusters [5, 6].

Another approach for finding dynamic clusters assumes that the objects to be clustered are distinguishable from one another based on their unique identifiers (e.g., the RFID tags of animals or IP addresses of mobile devices). As the locations of these objects change over time, a group of closely situated objects that move together for an extended period of time is called a moving cluster. There are two common strategies for creating and maintaining these moving clusters. The first strategy is to cluster the objects at each snapshot independently using a similarity measure defined in terms of their trajectory profile (spatial coordinates and velocities) [11]. The correspondence between clusters in different time periods is then established by measuring the fraction of objects the clusters share. In the second approach, spatial clustering is initially applied to the data in the first snapshot. The clusters are then updated incrementally by taking into account changes due to moving objects that leave or join the cluster as time progresses [12].

Also, signal processing techniques, such as Kalman filtering [8] can be used to track moving objects [1]. However, the application of such approaches may be limited because (i) the time intervals between successive observations can be large, and (ii) objects can appear and disappear between successive time frames. Indeed, the problem of dynamic clusters in Earth science data may need techniques that are more related to those creating a correspondence of clusters between two or more different clusterings of the same data [30].

5 Conclusion

We briefly described our current progress in applying clustering to find climate indices and discussed one of the challenges still remaining. We plan to investigate several approaches to modeling dynamic clusters, including ones that take into account domain specific factors such as seasonality, land cover, and geographical boundaries. Other challenges that need to be addressed include those common to most clustering algorithms, such as determining a strategy for handling outliers, parameter initialization, and the need to scale to large data sets.

6 Acknowledgments

This work was partially supported by NASA grant # NCC 2 1231 and by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014. 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 the AHPCRC and the Minnesota Supercomputing Institute.

References

[1] E. Brookner. *Tracking and Kalman Filtering Made Easy*. Wiley-Interscience, April 1998.

[2] J. Conley, M. Gahegan, and J. Macgill. A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis*, 37:286–314, 2005.

[3] L. Ertöz, M. Steinbach, and V. Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In *Proceedings of Third SIAM International Conference on Data Mining*, San Francisco, CA, May 2003.

[4] J. Han, R. B. Altman, V. Kumar, H. Mannila, and D. Pregibon. Emerging scientific applications in data mining. *Commun. ACM*, 45(8):54–58, 2002.

[5] F. Hoffman, W. Hargrove, D. Erickson, and R. Oglesby. Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models. *Earth Interactions*, 9:1–27, 2005.

[6] F. Hoffman, W. Hargrove, and A. D. Genio. Multivariate Spatio-Temporal Clustering of Time Series Data: An Approach for Diagnosing Cloud Properties and Understanding ARM Site Representativeness. In *Proc. of the 13th ARM Science Team Meeting*, Bloomfield, CO, 2003.

[7] J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. V. (eds.). *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*. American Geophysical Union, 2003.

[8] E. Kalman, Rudolph. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[9] T. Karl, R. Knight, D. Easterling, and R. Quayle. Indices of Climate Change for the United States. *Bulletin of the American Meteorological Society*, 77(2):279–292, 1996.

[10] V. Kumar, M. Steinbach, P. N. Tan, C. Potter, S. Klooster, and A. Torregrosa. Mining Scientific Data: Discovery of Patterns in the Global Climate System. In *Joint Statistical Meeting*, 2001.

[11] P. Laknis, N. Mamoulis, and S. Bakiras. On Discovering Moving Clusters in Spatio-Temporal Data. In *Proc. of the 9th International Symposium on Spatial and Temporal Databases (SSTD 2005)*, Angra dos Reis, Brazil, 2005.

[12] Y. Li, J. Han, and J. Yang. Clustering Moving Objects. In *Proc. of the 2004 ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining*, 2004.

[13] E. Martin, H. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. In *Proc. of the 1996 ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining*, Portland, OR, 1996.

[14] NASA Press Release, 2003. http://www.spaceref.ca/news/viewpr.html?pid=12031.

[15] NASA Press Release, 2004. http://www.spaceref.ca/news/viewpr.html?pid=15673.

[16] D. H. Portis, J. E. Walsh, M. E. Hamly, and P. J. Lamb. Seasonality of the NAO. In *AGU Chapman Conference on "The North Atlantic Oscillation"*, Orense, Galicia, Spain, November 28–December 1 2000.

[17] C. Potter, S. Klooster, R. Myneni, V. Genovese, P. Tan, and V. Kumar. Continental Scale Comparisons of Terrestrial Carbon Sinks. *Global and Planetary Change*, 39:201–213, 2003.

[18] C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, and C. Carvalho. Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Flux. *Global Change Biology*, 10:693–703, 2004.

[19] C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni. Global Teleconnections of Ocean Climate to Terrestrial Carbon Flux. *Journal of Geo-*

physical Research, 108, 2003.

[20] C. Potter, S. Klooster, P. Tan, M. Steinbach, V. Kumar, and V. Genovese. Variability in Terrestrial Carbon Sinks over Two Decades. Part I: North America. *Earth Interactions*, 7, 2003.

[21] C. Potter, S. Klooster, P. Tan, M. Steinbach, V. Kumar, and V. Genovese. Variability in Terrestrial Carbon Sinks over Two Decades. Part II: Eurasia. *Global and Planetary Change*, 49:177–186, 2005.

[22] C. 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, August 1999.

[23] C. Potter, P. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, and V. Genovese. Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets. *Global Change Biology*, 9(7):1005–1021, July 2003.

[24] C. Potter, P. Zhang, S. Klooster, V. Genovese, S. Shekhar, and V. Kumar. Land Use—Understanding Controls on Historical River Discharge in the World's Largest Drainage Basins. *Earth Interactions*, 8:1–21, 2004.

[25] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices. In *Proceedings of the KDD Temporal Data Mining Workshop, Edmonton, Alberta, Canada*, August 2002.

[26] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of Climate Indices Using Clustering. In *KDD 2003*, pages 446–455. ACM Press, 2003.

[27] M. Steinbach, P.-N. Tan, V. Kumar, C. Potter, and S. Klooster. Data Mining for the Discovery of Ocean Climate Indices. In *Mining Scientific Datasets Workshop, 2nd Annual SIAM International Conference on Data Mining*, April 2002.

[28] M. Steinbach, P.-N. Tan, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Clustering Earth Science Data: Goals, Issues and Results. In *Proceedings of the Fourth KDD Workshop on Mining Scientific Datasets, San Francisco, California, USA*, August 2001.

[29] H. V. Storch and F. W. Zwiers. *Statistical Analysis in Climate Research.* Cambridge University Press, July 1999.

[30] A. Strehl and J. Ghosh. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.*, 3:583–617, 2003.

[31] P.-N. Tan, M. Steinbach, V. Kumar, S. Klooster, C. Potter, and A. Torregrosa. Finding Spatio-Termporal Patterns in Earth Science Data. In *KDD Temporal Data Mining Workshop, San Francisco, California, USA*, August 2001.

[32] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach. In *the Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2003.

[33] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries. In *the Proc. of the 8th Int'l. Symp. on Spatial and Temporal Databases*, 2003.

[34] P. Zhang, S. Shekhar, Y. Huang, and V. Kumar. Spatial Cone Tree: An Index Structure for Correlation-based Similarity Queries on Spatial Time Series Data. In *the Proc. of the Int'l Workshop on Next Generation Geospatial Information*, 2003.