International Workshop on Visual Analytics (2011)
S. Miksch and G. Santucci (Editors)

Steerable Clustering for Visual Analysis of Ecosystems

Zafar Ahmed[†1,2], Patrick Yost[‡1], Amy McGovern[§1], Chris Weaver[¶1,2]

School of Computer Science[1] and Center for Spatial Analysis[2], The University of Oklahoma

Abstract

One of the great challenges in the geosciences is understanding ecological systems in order to predict changes and responses in space and time at scales from local to global. Ecologists are starting to recognize the value of analysis methods that go beyond statistics to include data mining, visual representations, and combinations of these in computational tools. However, the tools in use today rarely provide means to perform the kinds of rich multidimensional interaction that hold promise to greatly expand possibilities for effective visual exploration and analysis. As part of a project to develop a cyberCommons for collaborative ecological forecasting, we are developing ways to integrate highly interactive visual analysis techniques with data mining algorithms. We describe here our work in progress on steering mixed-dimensional KD-KMeans clustering using multiple coordinated views. Contributions include more flexible interactive control over clustering inputs and outputs, greater consistency of cluster membership during interaction, and higher performance by caching cluster results as a function of interactive state. We present our current tool that implements these improvements for visual analysis of Terrestrial ECOsystem (TECO) data collected from FLUXNET towers, with feedback on utility from our ecologist collaborators.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1. Introduction

Ecological systems are extraordinarily complex. They vary widely in structure within and across spatial scales, and both drive and respond to global changes non-linearly [ZWL08]. An essential component of ecological forecasting is understanding these drivers and impacts, with an eye toward providing sustainable services such as clean water, carbon sequestration, and preserving biodiversity.

FLUXNET [BFG*01] is a geographically diverse network of remote sensing towers that uses a variety of micrometeorological techniques to measure various kinds of ecological parameters such as precipitation, air and soil temperature, and net CO_2 exchange. Despite the rich mixed-multidimensional nature of FLUXNET data, exploration of this data still typically involves primarily non-visual statistical analysis of individual tower time series.

The cyberCommons is an ongoing $6M+ NSF-funded collaboration between researchers in informatics and ecology located at the major public universities in Oklahoma and Kansas. A major goal of the collaboration is to establish new analysis methods for use by the ecologists in understanding and forecasting natural and anthropogenic impacts on ecological systems in the Central Plains. We are focusing specifically on integration of visual analysis and data mining techniques that involve combinations of geospatial, temporal, and relational information. Balancing the research interests and needs of both sides of our collaboration led us quickly to focus our efforts on Terrestrial ECOsystem (TECO) data collected from the AmeriFlux network of FLUXNET tower sites in the United States, Canada, and South America.

We are using Improvise [Wea04] to develop and evaluate an evolving collection of tools for visually exploring and analyzing TECO data. Improvise is a desktop application for building and browsing visualizations with richly coordinated

† ahmed.zafar@ou.edu
‡ patrick.yost@gmail.com
§ amcgovern@ou.edu
¶ weaver@cs.ou.edu

DOI: 10.2312/PE/EuroVAST/EuroVA11/049-052

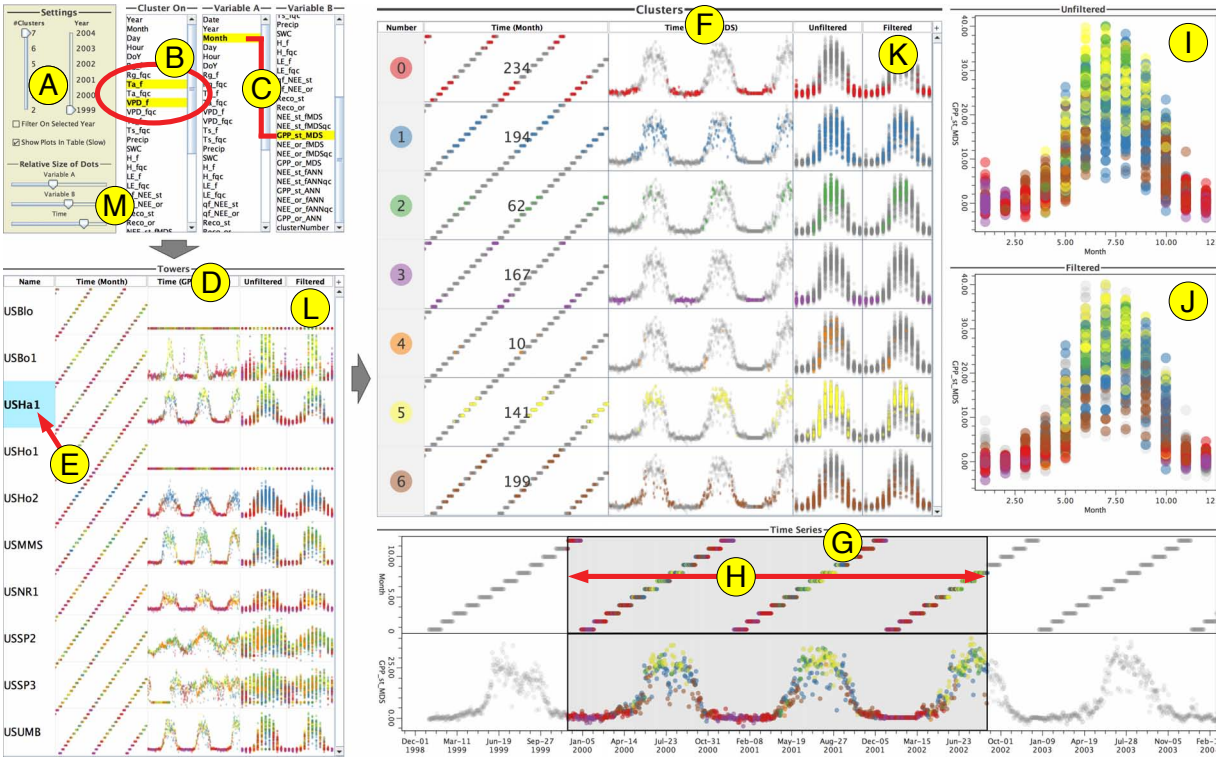

Figure 1: *Cluster visualization of six years of daily data from 10 AmeriFlux sites. (The labels are described in the text.)*

views of mixed-multidimensional data. At its foundation is an extensible library of interactively driven data processing algorithms and visualization techniques. This makes it a natural choice for developing highly interactive, flexible, visual clustering. Although what we describe here is an ongoing research effort, we have made progress in three directions:

- Multiple coordinated views provide control over dimensions to cluster, dimensions to display, and subsets of data values to include in clustering and/or display. Ecologists can select particular time scales, time ranges, and towers of interest in order to drill down into both bivariate relationships and univariate patterns over time. Interactive filtering can happen both before and after clustering, providing great flexibility of analysis.
- An extension of the k-means algorithm attempts to maintain reasonable assignment of tower observations to clusters in response to interaction, allowing for more consistent color-coding of cluster membership within and across views over the course of analysis. This allows the ecologists to visually track cluster shape by color, particularly while interactively selecting a time range to filter on by dragging and stretching in the time series views.
- Interactive performance is improved in two ways. First, the extended k-means algorithm calculates and caches a multi-resolution kd-tree over the dimensions of interactive variation, thereby substantially reducing subsequent cluster calculations, allowing sub-second interactive response. Second, cluster results are cached as a function of recent query parameter states, including while dragging the time range selection back and forth, providing even faster interaction when revisiting previous parameter states.

2. Tool Design and Use

Figure 1 shows the current Improvise visualization of TECO data. The user sets the number of clusters (A), then selects one or more driving variables to use as clustering attributes (B). They then select one independent and one dependent variable (C) to display separately in time series plots and together in bivariate plots. A table visually summarizes these relationships for each tower (D). Becoming interested in particular sites, the user can select any one tower (E) in order to explore temporal and bivariate patterns in the individual clusters for that tower (F), then explore temporal patterns more closely in larger time series plots (G). By dragging and stretching a time range slider (H), they can then compare bivariate patterns of clustering for the full data set (I) with clustering for data falling within the selected time range (J). The effects of time range filtering on cluster membership is also shown dynamically for both the individual clusters of the selected tower (K) and for all clusters of every tower (L).

Acceleration of k-means clustering often involves techniques such as approximation or choosing more useful starting points. We used the blacklisting algorithm by Pelleg and Moore [PM99]. It extends the multi-resolution kd-tree, a binary tree structure whose split nodes represent planes subdividing the data space into hyperrectangles. While building the tree, the space of data points is evenly partitioned and the relation of data points in space is easily discernable. As such, they are often used in nearest-neighbor algorithms. Using this data structure, it is usually possible to identify at upper nodes some centers to which no data points in the partition could be assigned, effectively blacklisting them. When points below that are assigned, there is no need to measure their distance to that center. Often, entire partitions can be assigned far from the leaf nodes, skipping many distance calculations. The result is an algorithm, KD-KMeans, that runs significantly faster on average. We added the KD-KMeans algorithm to the Improvise data transformation library. Because Improvise already caches transformation results as a function of query parameter states, this benefit accrued automatically to interactive steering of the algorithm.

The AmeriFlux network records observations at various time scales and then generates from these raw data four additional data sets of increasing levels of quality. We used processed observations of the highest (level 4) quality, which are gap-filled and contain flags to indicate the quality of both the original data values and those estimated for gap-filling. The AmeriFlux website (http://public.ornl.gov/ameriflux/) provides tower data from 1991 to 2010. However, the time required to process level 4 data means that it is not available for all towers in all years. Consequently, we chose ten sites from the North American Region that contain level 4 data from 1999-2004. Initial attempts with half hourly data over six years resulted in unreasonably slow interaction. The ecologists indicated more interest in exploring daily/seasonal characteristics than hourly characteristics anyway, motivating our switch to daily aggregated data.

Ideally, ecologists could select clustering attributes as they deem necessary for their analyses. The drawback is that selecting a large number of attributes reduces clustering speed substantially. Conveniently, there are two types of ecological variables in the AmeriFlux data: driving variables that are directly observed by the remote sensing towers (*DoY*, *Rg*, *Ta*, *VpD*, *TS*, *Precipitation*, *SWC*, *Le*, etc.) and dependent variables that are calculated as a function of observations (*Reco_sf*, *NEE*, *GPP*, etc.) The ecologists are strongly interested in looking at dependent variables over time. So, we limit possible clustering attributes to the set of driving variables. The clustering algorithm calculates similarity/distance using the selected subset of those attributes.

Whenever the user drags the time range slider, clustering is performed on the correspondingly filtered set of data points. Centroids for clusters can be randomly chosen and then converged iteratively depending on the similarity measure. For each new set of filtered data points, however, randomly chosen centroids not only increases calculation time but also leads to seemingly haphazard changes in cluster structure. We incorporated a naive cluster centroid caching scheme in which the previous cluster centroids are subsequently reused, with later runs iteratively converging to a new set of centroids from the cached ones. The new set of centroid points replaces the previously cached ones.

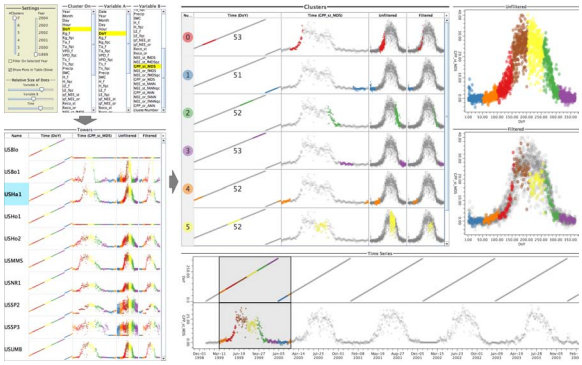

Figure 2: *Annual production and absorption of CO_2.*

A qualitative color scheme encodes cluster membership in all views. However, cluster identification often changes unpredictably during interactive filtering, meaning that the cluster-to-color encoding can update in a discontinuous and thus visually confusing manner. For this reason, we track cluster numbers as a function of cluster membership to maintain consistent cluster coloring.

Figure 2 shows an example of an interesting discovery by one of the ecologists. In order to explore the change in univariate patterns over time, he chose day of year with decimal fraction for time of day (*DoY*) as both the clustering and the driving variable, and gross primary production of CO_2 based on the net environmental exchange (*GPP*) as the dependent variable. In a particular year, *GPP* rises from zero to a peak value then back down again. In a broad sense, this means that the amount of CO_2 that the environment produces is absorbed by the environment in an annual cycle. Variation in clusters reveals different time stages throughout the year: flat (first cluster), slow accumulation (second cluster), rapid accumulation (third cluster), then to peak value (fourth cluster). In the fifth, sixth, and seventh clusters, the opposite (absorption) occurs. Despite using a naive similarity measurement—Euclidian distance—clustering still reveals distinct annual regimes. The ecologist became interested in how these regimes have changed over time. He dragged the time range back and forth; the larger plots showed updated cluster membership immediately. This revealed that *GPP* absorption was lower in summer 1999 than in other years. His expert conclusion was that this was due to drought, i.e., few rainfall events, low precipitation, high temperature, etc.

3. Related Work

Visual data mining is emerging as a mainstream methodology for analyzing large amounts of multidimensional data [KMSZ06]. Visual data analysis includes clustering for grouping and revealing trends in such data. A relatively recent survey [XW05] on this topic discusses several kinds of clustering algorithms and the similarity measures that are typically used with them.

Ankherst's [ABK98] heuristic optimization technique of grouping the dimensions of data and Choo's [CBP09] work on dimensionality reduction to preserve clustering is complementary to our work in that we can preprocess the dimensions of the data and provide visual interaction mechanisms in the dimension selection process. Nam [NHM*07] and iVibrate [CL06] are integrated interactive clustering frameworks in which clusters are calculated on a data set and then the user interacts with data points. Users can validate and label clusters but cannot perform multidimensional drill-down to look for and analyze patterns and trends. Jeong [JDN*08] has combined visualization with clustering to create tools for visual analysis of gene expression data, although without taking advantage of the benefits of rapid interaction across many dimensions in multiple views. Andrienko's [AA09] and Guo's [GCML06] work shows analysis using clustering and visual interactions of data with space and time attributes. The visualization system and analysis facilities are limited to spatio-temporal visual tools. Our work is unique in that we allow users to pose a wider range of pertinent multidimensional questions interactively within a single visual tool.

4. Conclusion and Future Work

Highly interactive clustering appears to provide powerful support for visual analysis of complicated ecological data. Our immediate next goal is to add geospatial capabilities to the current tool. Moving forward with the AmeriFlux data, our goal is to help ecologists materialize their discoveries from interactive clusters in a tangible way by supporting visual comparison of cluster structure over time and between towers, including a feature to record and annotate progressions of interactions that reveal interesting individual and differential cluster shapes. We are also working on a heuristic-based cluster centroid caching scheme to smooth interaction by reducing the number of iterations necessary in the clustering algorithm while preserving valid results. More generally, we plan to extend and implement additional clustering algorithms and evaluate their usage and knowledge extraction performance by comparing them with other visual and non-visual tools that use similar approaches.

5. Acknowledgments

This work is supported in part by National Science Foundation Award #0919466 and the State of Oklahoma Regents for Higher Education. We wish to express our gratitude to our collaborators Yiqi Luo and Shen Feng for providing access to and domain expertise on TECO data.

References

[AA09] ANDRIENKO G., ANDRIENKO N.: Interactive cluster analysis of diverse types of spatiotemporal data. *SIGKDD Explorations Newsletter 11*, 2 (2009), 19–28.

[ABK98] ANKERST M., BERCHTOLD S., KEIM D. A.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)* (Research Triangle Park, NC, 1998), IEEE Computer Society, p. 52.

[BFG*01] BALDOCCHI D., FALGE E., GU L., OLSON R., HOLLINGER D., RUNNING S., ANTHONI P., BERNHOFER C., DAVIS K., EVANS R., FUENTES J., GOLDSTEIN A., KATUL G., LAW B., LEE X., MALHI Y., MEYERS T., MUNGER W., OECHEL W., PAW K. T., PILEGAARD K., SCHMID H. P., VALENTINI R., VERMA S., VESALA T., WILSON K., WOFSY S.: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society 82*, 11 (2001), 2415–2434.

[CBP09] CHOO J., BOHN S., PARK H.: Two-stage framework for visualization of clustered high dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (Atlantic City, NJ, October 2009), IEEE, pp. 67–74.

[CL06] CHEN K., LIU L.: iVIBRATE: Interactive visualization-based framework for clustering large datasets. *ACM Transactions on Information Systems 24*, 2 (April 2006), 245–294.

[GCML06] GUO D., CHEN J., MACEACHREN A. M., LIAO K.: A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (November-December 2006), 1461–1474.

[JDN*08] JEONG D. H., DARVISH A., NAJARIAN K., YANG J., RIBARSKY W.: Interactive visual analysis of time-series microarray data. *The Visual Computer 24*, 12 (2008), 1053–1066.

[KMSZ06] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in visual data analysis. In *Proceedings of the International Conference on Information Visualization (IV)* (July 2006), IEEE Computer Society, pp. 9–16.

[NHM*07] NAM E. J., HAN Y., MUELLER K., ZELENYUK A., IMRE D.: Clustersculptor: A visual analytics tool for high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (Sacramento, CA, October 30–November 1 2007), IEEE, pp. 75–82.

[PM99] PELLEG D., MOORE A.: Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (1999), ACM, pp. 277–281.

[Wea04] WEAVER C.: Building highly-coordinated visualizations in Improvise. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)* (Austin, TX, 2004), pp. 159–166.

[XW05] XU R., WUNSCH D. I.: Survey of clustering algorithms. *IEEE Transactions on Neural Networks 16*, 3 (May 2005), 645–678.

[ZWL08] ZHOU X., WENG E., LUO Y.: Modeling patterns of nonlinearity in ecosystem responses to temperature, CO_2, and precipitation changes. *Ecological Applications 18*, 2 (2008), 453–466.