Interactive Selection of Multivariate Features in Large Spatiotemporal Data

Jingyuan Wang*
University of Tennessee

Robert Sisneros†
National Center for Supercomputing Applications

Jian Huang‡
University of Tennessee

ABSTRACT

Selecting meaningful features is central in the analysis of scientific data. Today's multivariate scientific datasets are often large and complex making it difficult to define general features of interest significant to scientific applications. To address this problem, we propose three general, spatiotemporal metrics to quantify the significant properties of data features–concentration, continuity and co-occurrence, named collectively as CO_3. We implemented an interactive visualization system to investigate complex multivariate time-varying data from satellite remote sensing with great spatial resolutions, as well as from real-time continental-scale power grid monitoring with great temporal resolutions. The system integrates CO_3 metrics with an elegant multi-space user interaction tool to provide various forms of quantitative user feedback. Through these, the system supports an iterative user-driven analysis process. Our findings demonstrate that the CO_3 metrics are useful for simplifying the problem space and revealing potential unknown possibilities of scientific discoveries by assisting users to effectively select significant features and groups of features for visualization and analysis. Users can then comprehend the problem better and design future studies using newly discovered scientific hypotheses.

Keywords: Multivariate, Interactive Feature Selection, Large Data, Metrics

1 INTRODUCTION

Current computing power has greatly accelerated both simulation capabilities and the collection of experimental and observational data. Datasets with an increasing number of variables paired with greater spatial and temporal resolutions are now common, posing significant complications for data analysis. It is crucial for domain scientists to differentiate and extract important information from a complex problem space. Hence, an adaptable, effective, and interactive visualization system to accomplish this goal is valuable for scientific discoveries.

Traditional feature extraction techniques are commonly utilized in many data analysis applications that involve large-scale multivariate spatiotemporal datasets. With the growth of computing power and data size, extraction of features with much finer detail is more affordable than ever before. While more features potentially contain more information, the amount of extracted features has become overwhelming to users – simple enumeration through these features is no longer plausible for analyzing the features in most cases. Interactive feature selection is called for such that a user can navigate, evaluate, and separate a complex problem space based on application-specific interest and significance.

In this work we propose three spatiotemporal metrics to enhance the feature analysis process by quantifying the significance of individual features and the correlation among multiple features. The metrics are **Concentration**, **Continuity**, and **Co-occurrence**—known collectively as CO_3. Integrated into the traditional workflow for large-scale multivariate data analysis, the CO_3 metrics can be interactively explored in concert using our prototype system called CO_3 Inspector.

The CO_3 metrics are general across application domains and are applicable in both the spatial and temporal domains. They are useful in the following ways:

- Enabling users to better specify what is 'interesting'—both strong and weak properties among the three metrics can be potentially significant to an application;
- Enabling users to identify and group features that are inherently correlated and analyze them simultaneously for possible scientific discoveries.

While the metrics serve as the backend of the analysis, our prototype system CO_3 Inspector provides a visualization and analysis front end by multi-linking the data, metrics, and statistical information together such that the users can explore the feature space effectively. Additionally, our user interface provides access to differing levels of granularity by which a user may customize how features are generated and how the three properties are evaluated.

We illustrate the effectiveness of CO_3 by exploring two datasets: continental-scale time-varying phenology data captured with satellites at 250-meter resolution, and continental-scale power grid monitoring data collected at sub-second resolution. The phenology data is acquired by the NASA MODIS satellite which covers the entire globe every 8 days at 250 meter resolution and has been collecting data since the year 2000. The power grid data are collected using 49 synchrophasor sensors distributed across the Eastern Interconnect of North America (47 in the United States, and 2 in Canada). These sensors collect multivariate data at 10 times per second and timestamp each record using GPS, resulting in 864,000 timesteps per day. These data provide an unprecedented view into how a real-world complex system, such as the power grid, operates in a large variety of conditions, including how it recovers from failure.

CO_3 Inspector's interactive techniques have allowed our scientists to select meaningful patterns from large scale datasets which were not known *a priori*. For example, we were able to distinguish modes of multivariate variation that are characteristic of normal operational states of the power grid system as well as those specific to "Storm Periods" resulting from two simultaneous major generator failures. In the 500 GB MODIS dataset, we were able to quickly identify very rare characteristic patterns such as those corresponding to irrigation systems built within chronically dry areas. These small patterns, which are often mistaken for noise, easily stand out using the CO_3 metrics.

We describe user needs of the driving applications and related previous work in Section 2. We then define the attribute space and show how it is built in Section 3. We present the CO_3 spatiotemporal metrics in Section 4 and detail the interactive visualization component in Section 5. Major results demonstrating the use of our system in selecting significant features and feature groups in the two driving applications are described in Section 6.

*e-mail: jingyuan@utk.edu
†e-mail: sisneros@illinois.edu
‡e-mail: huangj@eecs.utk.edu

2 BACKGROUND

With today's data collection technology, the creation of large, high resolution datasets has become commonplace. As is the case with each of our driving applications, the ability to effectively handle such data is necessary to observe the behavior of a very complex system such as the earth or the power grid. The setting of our research is general to many other data-intensive applications.

The challenge is to derive previously unknown knowledge about the multivariate patterns in such complex physical systems. In a change from the past, with these new data intensive applications, it is quite possible to obtain millions of features using existing techniques, such as clustering [17, 21, 22] , geo-spatial-temporal queries [6] and variable space range queries [13, 5].

Ground truth about the relationships among these features, however, is largely lacking—it can sometimes be simulated, but only in a very limited manner. Hence the starting point of research involves questions like, "Which features are important?," "Which groups of features occur together?," and "In what order and with what consequence." These questions motivated us to develop the CO_3 metrics and the Inspector system to verify the efficacy of the CO_3 metrics. In the following we review the background of our applications and the relevant existing methods in the visualization literature.

2.1 Characterizing Phenology of Forest Ecosystems

Understanding and safeguarding the health of the planet's ecosystems is pivotal to our security, economical prosperity, quality of life, and the stewardship of our natural and cultural heritage. To this end, a key aspect is to understand and separate "normal" or healthy patterns of variation in an ecosystem from those that are abnormal and indicate threats to ecosystem health that may require intervention. Here we explore such spatiotemporal variations in forest ecosystems using remotely sensed vegetation patterns of growth and deterioration, or phenology. Previous works including Mills et al. [17] have successfully applied methods for geospatiotemporal data mining of multi-year land surface phenology data in detecting threats to forest ecosystems. The dataset we use consists of Normalized Difference Vegetation Index (NDVI) values, a measure of "greenness", from the Moderate Resolution Imaging Spectroradiometer (MODIS).

2.2 Power Grid Situation Awareness

The power grid is a critical fixture in our current industrial era. Our society depends on its consistent availability. Power grid failures could paralyze a city, region, or in the worst case, an entire country. Situation awareness visualization plays a significant role in helping grid operators to better monitor the current environment and to recognize, prevent or recover from major system failures [12, 18].

Our data was collected on the Eastern Interconnect of the U.S. on April 27, 2011 - a day when two major power generators temporarily went offline and caused widespread oscillation in the power grid. A total of 49 FNET (Frequency monitoring NETwork [25]) devices distributed across the Eastern Interconnect recorded data at 0.1 second resolution. We were given the rough time of major generator trips. The duration of load shedding and severe oscillation is referred to as the "Storm Period". Frequency, voltage and phase angle are three variables measured by the devices. We refer to this dataset as FNET.

2.3 Previous Work

The study of features in multivariate scientific data has been a central topic for visualization research. Broadly defined, a transfer function for volume rendering is a method of feature selection. There has been abundant work on extracting features from the attribute space as well as the spatial/temporal dimensions from multivariate spatiotemporal data. Due to the complexity, it has become prevalent to use multiple linked views to simultaneously show, explore, and analyze different aspects of multivariate data. Examples include SimVis [2] and follow-up research works like [3, 11] that demonstrates the ability of multiple linked views to enable iterative feature specification and hypotheses generation. Our work also follows the same practice.

Many previous works on feature extraction undertook the perspective of classification. A common goal is to classify voxels into a few classes, after which a user could interactively (but manually) enumerate through and control how they are rendered. For example, Tzeng and Ma [21] classified volume data using a clustering algorithm while Ip et al. [9] applied a hierarchical segmentation method. As the amount of potentially viewable features increases, the appeal of automatic feature extraction is likewise magnified. There are methods to automatically assign rendering settings based on regions of interest [23] and leverage non-parametric clustering in transfer function space to guide transfer function generation [14].

Many researchers have incorporated statistical properties of data to the workflow of data analysis. Recent examples include a rank-by-feature framework proposed by Seo et al. [20] that enables users with better understanding of subspaces of multidimensional data by ranking them using quantitative criterions, work to statistically analyze time activity curve by Fang et al. [4], a method to automatically select turbulent flow features using local statistical analysis by Janicke et al. [10], an approach to create a transfer-function space based on statistical properties derived from neighborhood of each sample point by Haidacher [7] and an approach to abstract attribute space by using information metrics detailing the relationship between attributes of the multivariate volume data by Maciejewski et al. [13]

Correlation within data becomes an interesting analysis subject as well as an assistive tool. Chen et al. [1] devised a sampling-based approach to correlation classification for time-varying multivariate data. Mehta et al. [16] derived three spatiotemporal relationships– directional, topological and navigational. They incorporated spatial and temporal graphs to display the spatial and temporal trajectories of scientific objects. Yang et al. [24] developed the Value and Relation Display method to effectively and efficiently explore large datasets with several hundred dimensions based on relationships among the dimensions.

This paper uses CO_3 metrics to analyze the properties and correlations of features extracted through hierarchical clustering. However, the metrics differ from the existing clustering metrics like homogeneity and completeness [19] since these existing clustering metrics are designed to measure the quality of clustering algorithms, whereas CO_3 measures the spatiotemporal properties of the clusters. Furthermore for these clustering metrics, there is an assumption that correct cluster assignment is known. Our research is complementary to the existing work in that our goal is to study how to select features when there are much more than just a few hundred. The aim is for users to explore a large number of features from a high data-rate real-time observation of a real-world system, such that they can hypothesize about which groups of features occur together, how those groups of features occur together, and consequently which groups of features are important for recognizing application domain issues. For each feature, CO_3 assigns its significance according to in which neighborhoods or among which group the feature consistently appears.

3 ATTRIBUTE SPACE

CO_3 operates in two different spaces. **Attribute Space** is where multivariate data is processed and abstracted into features based on similarity. **Physical Space** is where we distinguish how features are distributed across space and time and whether they are mutually coincident in the spatial or temporal neighborhoods.

Multivariate feature extraction in the attribute space is a separate

preprocessing module from the CO_3 Inspector system. CO_3 metrics can handle features extracted from any methods that produce spatially or temporally distributed features. This is an important process that requires high efficiency and accuracy, especially for large-scale datasets. In this work, we use a customized parallel hierarchical clustering algorithm to create abstractions of the dataset at multiple scales, offering the users the capability to analyze the problems at varying granularity. Our hierarchical clustering is implemented in a bottom-up fashion. Small grained clusters are merged together as long as the distance between cluster centroids is under a pre-set threshold. As hierarchical clustering progresses to coarser levels, the distance threshold increases linearly. Each cluster is a multivariate feature regardless of which level or scale.

MODIS: As the yearly vegetation variation is one of the research focuses of climate scientists, treating the vegetation indices collected at different times in the year as different variables is useful for the analysis purpose. The whole satellite observational data is structured as a regular grid of 19732 (longitude) x 13571 (latitude) x 11 (year) x 46 (variable). Utilizing sophisticated dimension-reduction techniques and hierarchical clustering algorithms, the whole dataset is abstracted into a hierarchy of clusters. The number of clusters varies from 14225 at the bottom level to 223 at the top level.

FNET: The whole power grid dataset is structured as a regular grid of 864000 (time step) x 49 (location) x 3 (variable). In addition to the three measured variables in the dataset, variation of these variables are also included in the feature extraction process as recommended by the domain experts. The resulting hierarchy contains 49642 clusters at the lowest level and 1680 at the highest one.

4 Spatiotemporal Feature Metrics

4.1 Multi-Scale Physical Space Overview

It has been a common assumption that all features can potentially play an important role. Hence, many techniques render features directly in their original spatiotemporal space and leave it to the users to determine what features deserve further exploration. That assumption is less than ideal for handling features that may be noise-corrupted, redundant or less informative.

The purpose of developing metrics is to provide a general way of quantifying significance among a large number of features. Our CO_3 metrics, concentration, continuity, and co-occurrence, encapsulate properties that are readily identifiable in the physical space, both spatially and temporally. The metrics represent three desirable properties when exploring for interesting features by domain experts.

CO_3 metrics are defined on a per-cluster basis and assume that the 4-dimensional space including the spatial and temporal domain has been partitioned into coarse grained bins, referred to as *regular bins*. All dimensions are treated equally in the partitioning. In general, the granularity of each regular bin is defined in the 4-dimensional space [x, y, z, t]. As for different analysis focuses of datasets, MODIS is partitioned spatially to study the distribution of yearly vegetation growing pattern in geographical space while FNET is partitioned temporally to study the dynamics of power grid over time. Example granularities could be [5 km, 5 km, −, 1 year] or [−, −, −, 1 second], where the '−' symbol denotes an undefined or unpartitioned dimension.

The CO_3 metrics are computed based on the distribution of clusters on the partition of the physical space, hence, the choice of bin size affects the values of the metrics. The CO_3 Inspector system empirically provides a pre-set of bin sizes: 5, 10, 15 and 20 km for MODIS and 1, 2, 5, 10 second for FNET. These pre-set bin sizes are based on the rough spatial/temporal scale of application problems that domain experts are interested in. For instance, 10 seconds is considered to be a long period of time in which power transmission on the grid would vary much.

To properly define these metrics for feature properties, we need the following notations and quantities:

F_i: Cluster i
E_i: Number of elements of F_i.
E_{ib}: Number of elements of F_i in regular bin b.
t: The percentage threshold for identifying significant bins.

For a given cluster F_i, the set of *significant bins* is the smallest set needed to represent some percentage t of all data points belonging to cluster F_i. For example, in Figure 1, a cluster F_i contains 28 data points and is spread over 5 bins, A through E. We sort the bins in decreasing order of E_{ib} and then traverse the array, computing the prefix sum of E_{ib}. We stop the traversal as soon as the prefix sum has reached t. For a value of $t = 90\%$, the significant bins would be A through D. The concept of significant bins elegantly handles noise-like anomaly data, the choice of t is application dependent.

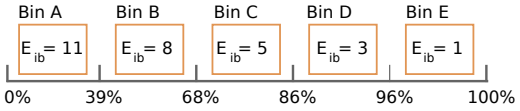


Figure 1: An illustration of determining significant bins. Given a cluster (F_i) and the number of its data points per bin (E_{ib}, in decreasing order), the set of significant bins is the smallest group of bins that can represent F_i's presence above a given percentage threshold (t).

In the following subsections, we describe the three CO_3 metrics.

4.2 Concentration

The concentration metric, C_1, denotes the average occupancy of bins in the set of significant bins for a given cluster. It indicates the properties of a cluster with respect to both physical distribution and size and is calculated as:

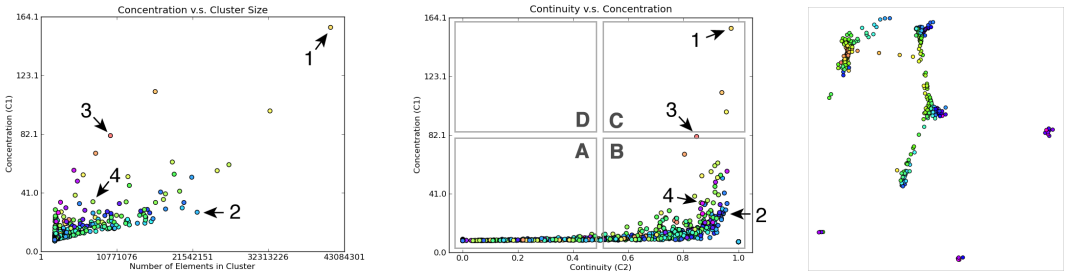
$$C_{1i} = \frac{E_i * \gamma}{K_i} \tag{1}$$

where γ is the percentage of elements within significant bins for a given cluster F_i.

Since C_{1i} depends on K_i, the number of significant bins, this guarantees that C_1 is unaffected by outlier data in the cluster. Highly concentrated features have a high representation in a small number of significant bins and will therefore have a high C_1 value. Clusters with a smaller representation across bins will stack on the lower end of the C_1 axis. A concentrated feature can be a dominant pattern across a large portion of the physical space because of its large volume of data elements. It can also be a smaller-sized feature representative of certain locale in the physical space.

Figure 2(a) illustrates the space formed by concentration *vs.* cluster size. The metrics are computed using a 5 km bin size. On a 250-meter resolution grid, this amounts to 400 geographic locations in every bin. A C_1 value of 200 or more indicates that a cluster monopolizes more than half of its significant bins. When a user examines highly condensed patterns such as vegetation damage due to insect infestation, those feature patterns are small yet highly concentrated. The clusters corresponding to them will not appear among the large ones. The search should start from the left side of Figure 2(a), populated by smaller clusters.

Also in Figure 2(a), several individual clusters are labeled for comparison. Cluster "1" and "2" are both large but have very different concentration properties. Cluster "1" is one of the largest features on the continental U.S., yet it is so concentrated that it takes up almost half of each physical bin. That cluster happens to correspond to the mountainous areas of the western United States. Cluster "2" is large but does not monopolize any 5 km-square geographical bins. Cluster "2" distributes over the middle and eastern part of the United States. Cluster "3" is similar in size to cluster "4", but is more concentrated with its spatial presence concentrated on



(a) Concentration *vs.* Cluster size (b) Concentration *vs.* Continuity (c) Co-occurrence graph layout

Figure 2: Statistics views based on CO_3 metrics. Various examples of utilizing CO_3 metrics in visualization and analysis. The utility of the visualizations in the subfigures (along with corresponding labels) are discussed in Sections 4.2, 4.3, and 4.4. (Year 2003, 5 km bin size)

lakes and other water bodies. Clusters appearing at the right-bottom corner of this plot are likely widespread noise in the data.

4.3 Continuity

C_2 denotes the continuity of significant bins for a given cluster. Bins are connected if they comprise spatiotemporally continuous regions. Connected significant bins are grouped into *significant regions* and C_2 is calculated as:

$$C_{2i} = 1.0 - \frac{R_i}{K_i} \tag{2}$$

where R_i is the number of significant regions and K_i is the number of significant bins. Hence, C_2 can range from 0.0 (no bins connected) to 1.0 (all bins connected, 1.0 not included).

When paired, continuity and concentration create an interesting space. We believe the C_1 *vs.* C_2 space can be divided into four areas in which clusters that fall in the same area share similar spatiotemporal properties. For example, Figure 2(b) shows a sample plot of C_1 *vs.* C_2 with labeled regions. In this space, low concentration and low continuity likely represent noisy data elements (A); high concentration and high continuity represent a cluster that is well represented in distinct spatial regions of the data (C); and low concentration and high continuity could easily represent elements of data that define "normal" data elements for given regions (B). Defining features of interest is entirely dependent on the application however. Figure 3 provides a map view of the clusters in regions (A), (B) and (C).

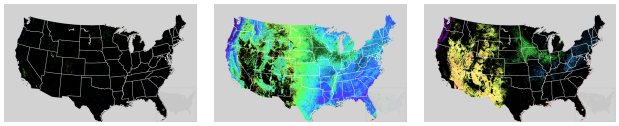


Figure 3: Clusters in quadrants A, B and C (left to right) in Figure 2(b).

Note that clusters "1" and "3" in Figure 2(a) are still clearly distinguishable in Figure 2(b). From that, we can tell both of those clusters are highly concentrated and continuous and are likely features representative of a geographic area. Cluster "2" in Figure 2(a) is also highly continuous as the agricultural growing pattern represented by Cluster "2" is more or less common in the middle and eastern US though not prevalent.

4.4 Co-occurrence

While concentration and continuity quantify global properties of a single cluster, we also desire to assess clusters locally and within the context of one another. Co-occurrence, or C_3, measures the degree to which clusters reside near each another (i.e. are collocated) and assists in the analysis of relationships between features. Unlike C_1 and C_2, C_3 is calculated from all bins, not just significant bins.

$$C_{3ij} = \frac{\sum_{b \in V_{ij}} min(E_{ib}, E_{jb})}{(E_i + E_j)/2} \tag{3}$$

For two clusters F_i and F_j, V_{ij} is the set of regular bins in which F_i and F_j overlap. C_{3ij} measures how much F_i overlaps F_j in spatial presence on the granularity of spatial bins. Hence, C_3 will range from 0.0 (no overlap) to 1.0 (perfect spatial overlap). This metric is very well-conditioned to be directly used for edge weights in a force-directed graph layout algorithm (discussed in Section 5.2). We threshold edge weights and filter out edges before performing the graph layout. Figure 2(c) is an example with a threshold corresponding to keeping only top 40% of edges and a bin granularity of 5 km. We omitted edges to reduce over-plotting.

With concentration and continuity, users can specify significant features based on the strong or weak properties; however, co-occurrence is more complex to understand because co-occurrence can not be examined using the concept of 'high' or 'low'. However, by employing a graph layout algorithm to embed the features into a two-dimensional graph, users can better visualize and analyze this metric. In the graph, the position of a particular feature has no physical meaning. The distances between features are the only measurement related to C_3. If features are close to each other in the graph, it means these features are near each other spatially or they occur in similar period of time.

The significance of C_3 is shown by our driving applications. Climate scientists are always interested in discovering exact causes of abnormal growing patterns. Two co-occurred features imply certain ecological scenarios. It could be that they are both consequences of the same event, like unexpected regional drought. Or it could be that one of them is the cause of other co-occurred features. Similarly for the power grid application, unusual events that occur shortly before or after abnormal power grid operation states, like large-scale frequency oscillation, are significant. Understanding the reasons for and the consequences of an abnormal event is crucial for handling similar occurrences in the future.

Although initially more complex to understand, C_3 actually presents a great deal of information about feature combinations which is often neglected or missed in traditional attribute analyses. In Section 6, we present some interesting groups of features discovered from the co-occurrence graph of the CO_3 Inspector system that were not known a priori.

5 Inspector - The User Interface

Figure 4 shows the initial view of CO_3 for MODIS. The interface has three components: a spatiotemporal view, a co-occurrence graph, and statistics plots.

All clusters are assigned different colors based on cluster centroids, and the same color scheme is used across all views and clustering levels. Since the system is designed for visualizing a large number of clusters and the color represents the multivariate properties of clusters rather than categories, repetitive color assignment,

as used in Dimstiller [8], is not a choice in this case. In MODIS, the colormap is indexed according to the primary and secondary principle components. In FNET, the red, green and blue channels are assigned according to changes in frequency, voltage and phase angle, respectively. Missing data is transparent.

5.1 Spatiotemporal Rendering

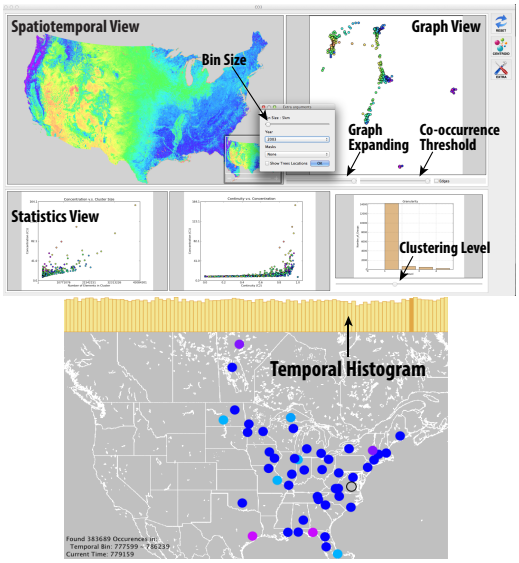


Figure 4: (Top) A snapshot of CO_3 Inspector showing the spatiotemporal view, graph view, and statistics view; (bottom) Spatiotemporal view adapted for the power grid application.

The spatiotemporal view is specifically designed for different datasets. For MODIS, 2D image-based rendering is implemented while FNET uses an adapted view with a temporal histogram above the map and sensor locations represented by colored disks. In the temporal histogram, the entire day's data is partitioned into roughly 15-minute intervals and the height of a bar corresponds to the number of occurrences of chosen clusters during the 15-minute interval.

In both cases, spatiotemporal renderings color each location according to its cluster membership.

5.2 Co-occurence Graph

We use a graph layout to visualize co-occurrence. At any level in the cluster hierarchy, we can consider each cluster as a node v in a graph $G(V,E)$ with edge weights assigned by the C_3 metric.

The graph layout is computed using a force-directed method with an energy barrier [15]. Proximal clusters are represented as proximal nodes in the final layout. Also, we capture the animated process during which a graph layout converges. Users find the functionality of being able to view at least the final steps of a converging graph layout to be very useful in examining subtle differences in co-occurrence. This is demonstrated in Section 6.2.

5.3 Statistics Plots

In the data exploration process, statistics are a classical way for domain experts to explore local or global characteristics of data. When coupled with more complex rendering techniques, this numerical exploration can effectively assist with user interaction. In our application, the statistics view offers an easy and flexible interface to control the multi-levels of clustering results. Furthermore, it provides users with useful quantitative feedback. CO_3 Inspector provides four widgets: a scatterplot widget of C_1 *vs.* C_2, a scatterplot widget of C_1 *vs.* cluster size, a histogram of the number of clusters in any hierarchical level and a parallel coordinates plot activated upon selection of clusters in any space (Figure 10(e)). The parallel coordinates plot is used to display the multivariate values of cluster centroids.

5.4 Multiple-view Coordination

Each view in the interface is fully coordinated with all other views such that any action taken in one view is immediately reflected in all others. In this context, analysis is an iterative and user-driven approach with each step providing instant feedback while refining focus.

During the interactive visualization phase of CO_3, only clusters are analyzed, oblivious of the raw multivariate time-varying data. Brushing is enabled to select clusters in any of the viewports. Selected clusters are highlighted with a semi-transparent plus sign. Brushing using the left mouse button makes 'fresh' selections whereas brushing done with the right mouse button selects a subset of the already selected clusters. Brushing operations can be arbitrarily chained together as a result of iterative user interactions.

5.5 Implementation

The Inspector system employs image-based rendering techniques. Matplotlib, a python plotting library, is used to generate statistics plots and co-occurrence graphs. These are pre-generated only once after the feature extraction and evaluation of the CO_3 metrics. Such preprocessing improves the speed of interaction and provides the system with comprehensive plotting features. The Inspector system then offers visualization functionalities interactively to provide immediate feedback on a single laptop computer. The rendering preprocessing, including the calculation of the co-occurrence graph layout, is executed in parallel. This takes about 7 minutes for FNET and 60 minutes for MODIS on a 12-core Linux workstation in the setting presented in the paper.

6 Results

With both application datasets, navigating through the multi-level feature space formed by hierarchical clustering is particularly difficult for domain users since the total number of features is beyond a person's ability for the traditional click-and-view analysis process. Analyses become even more complex when feature correlation is included. Highly correlated features are intuitive to analyze in groups and exhibit promising opportunities for scientific discoveries. Our CO_3 Inspector greatly reduces users' work by highlighting the important solitary features and, more importantly, groups of features. Domain experts are then able to carry out analyses following the visual hot spots that appear along the road to discovery. The usefulness of the CO_3 system is demonstrated in the following two categories of examples: selecting significant individual features and selecting significant groups of features. Our system is designed with an emphasis on new scientific discovery; the examples discussed in this paper are therefore focused on detecting outlier patterns over those commonly occurring.

The features extracted from MODIS dataset provide information on the vegetation growing pattern year-wide. Mills et al. [17] have termed these phenology class assignments **phenostates**. For FNET, the features describe sets of 1-second events that share the same operational behaviors in the power grid.

6.1 Selecting Significant Individual Features

Using the statistics view widgets of the CO_3 Inspector, users are able to specify significant features by selecting the strong or weak properties of the CO_3 metrics.

Example 1 (MODIS): Figure 5 shows an example of one unique feature extracted from the MODIS dataset. View A in Figure 5 shows the concentration *vs.* continuity space while View C in the same figure shows the concentration *vs.* cluster size space. In both views, dozens of phenostates stand out from the whole population in the space and spotting them is straightforward. The selected one

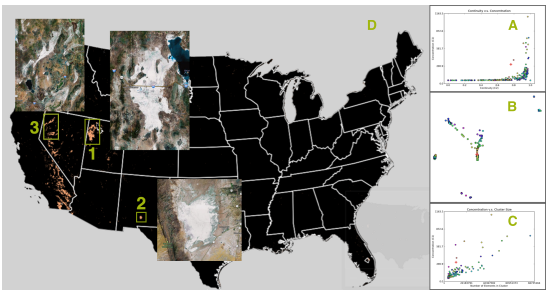


Figure 5: A highly continuous and concentrated feature in the MODIS data that captures areas of salt plains and white sands - The Bonneville Salt Flats (1), White Sands National Monument (2), and other salt flats (3) are highlighted. (Year 2000, 15km bin size)

(in red) is small but has a relatively high concentration and continuity. The map in View D shows the geographic locations with the phenological properties defined by the phenostate selected. The three labeled regions in the figure represent salt flats and areas with white sands. The Bonneville Salt Flats near the Great Salt Lake is the most contiguous and concentrated feature. The other areas including the White Sands National Monument capture similar phenological properties— areas of the United States that remain a very white color year round (in contrast with snow, which is seasonal) and have absolutely no vegetation.

Example 2 (FNET): Similarly, in the FNET dataset, a small number of features standing out from the majority in the metric space represent possibilities of discovery. A tiny, highly concentrated and continuous feature (in the top left of Figure 6(b) and the top right of Figure 6(c)) proves to be unique after further study. The corresponding operation state dominantly appears across most of the Eastern Interconnect but only within a very short period after the two generators' temporary failure (shown by the solitary tall bar in the temporal histogram). This feature has an exceptionally low frequency but a large phase angle shift. Its dominance indicates that this state is characteristic of the gradual process of the grid recovering to normal operation. Further study of this would assist in understanding the recovery process and help technicians respond quickly to severe power grid failures.

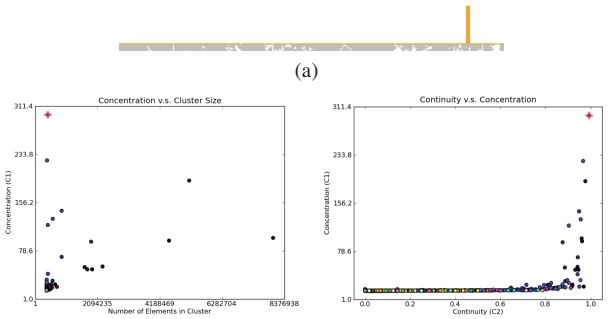


Figure 6: An example from the FNET data showing a highly concentrated and continuous feature, with an exceptionally low frequency but a large phase angle shift. Temporal histogram (a) shows this feature is exclusive to a small window of time after the "Storm Period". (1s bin size)

6.2 Selecting Significant Groups of Features

Selecting significant individual features can become overwhelming for a large number of features. As discussed earlier, simple enumeration is no longer plausible in actual analysis and discovering the important correlations among features in a complex feature space is very challenging. The co-occurrence graph layout of the CO_3 Inspector, linked with the other metric views, enables users to navigate through the feature space, and select both interesting individual features as well as interesting groups of features.

Example 1 (MODIS): Exploring the co-occurrence graph relating to the salt flat phenostate mentioned above, we find a related feature group spreading out sparsely over the whole continental U.S. By further drilling down on some of the co-occurring phenostates, a set of small phenostates shows very interesting spatial distribution and is illustrated in Figure 7. The three labeled areas in the figure are representative of arid lands that contain significant green areas because of human activity and irrigation. These clusters are phenologically similar to the salt plains in that they are regions with a severe lack of available water.

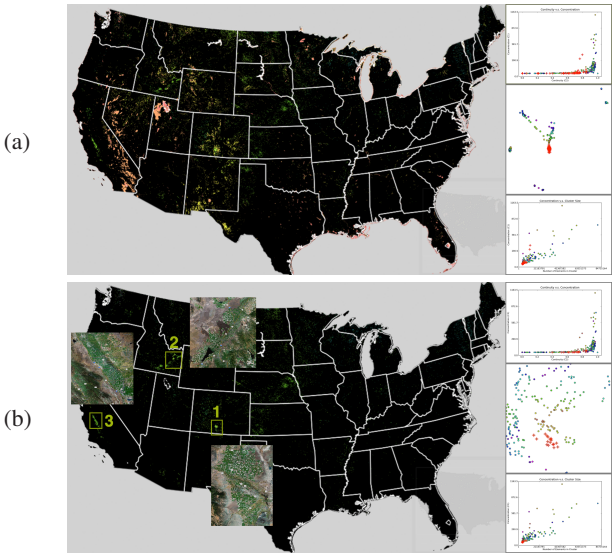


Figure 7: An example feature group in the MODIS data. The selection starts from the example feature of salt flats and white sands. After a series of navigation and selection refinement steps, users are able to discover such a feature group. These phenostates reside in some large irrigated lands in dry areas. (Year 2000, 15km bin size)

Example 2 (MODIS): Starting from the co-occurrence metric in the MODIS dataset with a 5km bin size, we notice a feature group in the co-occurrence graph which is highlighted in the spatial view (Figures 8(a) and 8(b)). It appears that this feature group represents the outline of the Central California Valley and other areas in the Southern Great Plains, both of which undergo rather irregular growing patterns due to the terrain of the Sierra Nevada mountain range and the highly varied weather patterns in the southern part of the Great Plains. By looking at an earlier stage of the graph layout convergence process, as shown in Figures 9(b) and 9(d), we can see that the original feature group is now expanded into two distinct areas. The top half of the group gives a near-exact outline of the Central California Valley.

We iteratively adjust the bin size to 20km (Figures 8(c) and 8(d)). We notice that the group of clusters gets tighter. However, the two main parts in the original group of features become separated when the bin size is 20km, leaving the Central California Valley directly selectable without going back to the earlier layout process. In this result, with larger bin sizes, we were able to select structures belonging to larger spatial scales.

Example 3 (FNET): With the FNET dataset, we can analyze the uncommon event called a "Storm Period" by computing the co-occurrence graph composed solely of the features that occurred during that time (Figure 10). The co-occurrence graph layout clearly reveals 6 characteristic groups that provide additional information upon further examination. Figure 10(a) reveals that one group (highlighted in red) involves features with both high frequency vari-

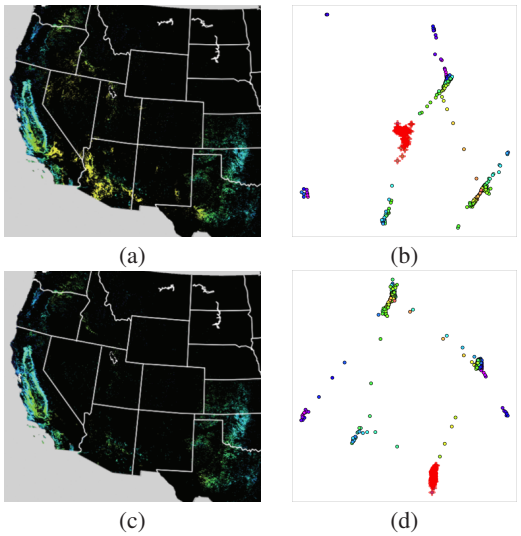


Figure 8: An example feature group in the MODIS data. The spatial distribution a the feature group containing areas in the Southern Great Plains and the outline of the Central California Valley (Year 2003, (a, b) 5km bin size, (c, d) 20km bin size).

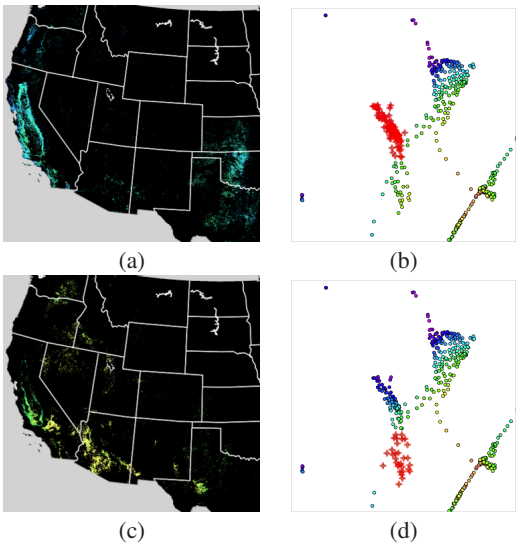


Figure 9: We expand the graph in Figure 8(b) and refine the selection to two individual parts that initially appeared together. The area showed on the top row is almost totally correlated to the Central California Valley. (Year 2003, 5km bin size)

ation and high voltage variation and occurred most frequently during the first half of the day (UTC time 00:43:12 to 08:52:48). Figure 10(b) shows another feature group that had a heavier presence in the second half of the day (UTC time 12:00:00 to 21:07:12) and exhibited smaller variations in frequency but larger ones in voltage. As visualized by the parallel coordinate rendering in Figure 10(e) *vs.* Figure 10(f), the contrasting behavior could help to motivate further domain science research to explain the cause and progression of a "Storm Period".

Discussion: The feature groups identified in the above examples show characteristic multivariate properties that might be significant to domain-specific users. However, without a proper tool, it is difficult to select these groups of features from the cluttered attribute space.

As an example, the first feature group representing salt flats and white sands contains 17 clusters, all of which are small with no distinguishing traits in terms of concentration or continuity. Selecting individual features from among the whole population is not straightforward. Even enumerating all features in the attribute space (223 clusters) would not help much in this case since each is too similar to the others to attract a user's attention. //none of them is unique enough to attract the user's attention. Also, these features might be well hidden among other small features and mistakenly considered to be noise. By selecting them, the spatial distribution reveals a meaningful and significant pattern.

There are many existing approaches for multivariate feature visualization. Using parallel coordinates, users are able to specify ranges for one or multiple variables to select a subset of all the features. Figure 11(b) is a parallel coordinates plot corresponding to the first selected feature group shown in Figure 7. All features in the same attribute space are plotted in orange and shown in Figure 11(a). Here, the selected features share similar variation patterns while the actual values of the vegetation indices are not particularly close to each other. In Figure 11(a), the multi-dimensional curves of all features do not show a clear structure or pattern to assist users in making a selection, as in Figure 11(b).

Similarly, Figures 11(c) and 11(d) demonstrate the advantage of selecting a significant group of features using our tool. In these two figures, the original features are colored in orange and the feature groups selected using our tool are colored in red. Figures 11(c) and 11(d) correspond to the example feature group illustrated in Figures 10(a) and 10(b), respectively. Additionally, we highlight two example selections of features using two different range queries in the same plot (blueish color). The queries are conducted on frequency (11(c)) and frequency variation (11(d)) variable, based on the exact distribution of the corresponding example feature group. In Figures 11(c) and 11(d), the total number of features is 1680. The queries on parallel coordinates results in selections of 1059 and 684 features respectively while the selection using CO_3 Inspector contains only 52 and 32 characteristically similar features. In both cases, simple compound range queries could not reveal the patterns found using our tool.

7 Conclusion and Discussion

CO_3 represents a new possible solution to facilitate visual and interactive feature analysis and selection by developing quantitative metrics that combine physical and attribute domains. Our capability to effectively select significant groups of features demonstrates the power of the CO_3 metrics in exposing previously unknown possibilities to users. The feature selection capability we demonstrate is crucial as datasets consistently and quickly grow in size and complexity. CO_3 metrics are especially useful to summarize physical space properties of features extracted from attribute space. Our use of coarse grained bins is general and novel for handling high resolutioned spatial and temporal datasets. Our domain experts from climate modeling and power systems find CO_3 metrics and the Inspector system to be useful for analyzing historical data. They were intrigued by the feature groups discovered by the Inspect system, and expressed a perception of a high level of utility and future potential. Our work also has a few limitations. First, our method requires non-trivial parallel preprocessing. Next, the color scheme we employed was chosen out of convenience. Finally, our method would not offer significant benefit over previous methods, if the dataset is relatively manageable or the feature set is already well understood.

8 Acknowledgement

We thank Dr. Richard Mills and Forrest Hoffman of Oak Ridge National Laboratory for inspiring us to undertake this research topic and for their insightful and substantive feedback. We also thank Dr. Wesley Kendall of University of Tennessee for his help with refining the scope of this work. Our work was supported in part by NSF

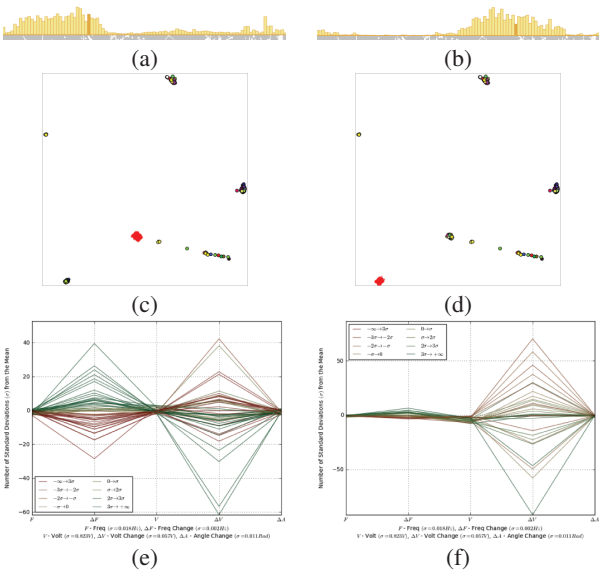


Figure 10: Two example feature groups in the FNET data; both occurred in a "Storm Period". Both are inherently correlated in terms of co-occurrence, but contain highly contrasting distribution patterns. (1s bin size)

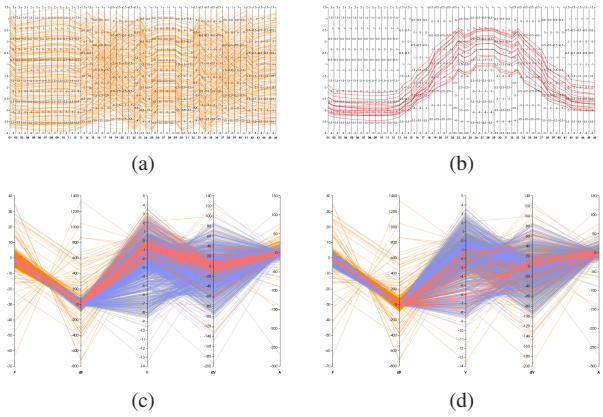


Figure 11: Traditional parallel coordinates plots of example feature groups (rendered with the other features in the same attribute space in different colors): (a, b) Figure 7, (c) Figure 10(a), (d) Figure 10(b). The features that selected using CO_3 Inspector are colored in red. The features selected by specifying variable ranges (c and d) are plotted in bluish color. The rest are colored in orange.

Office of Cyber Infrastructure under ARRA-NSF-OCI-0906324, DOE SciDAC Ultrascale Visualization Institute (DOE DE-FC02-06ER25778), and by the Engineering Research Center Program of NSF and DOE under NSF-EEC-1041877.

REFERENCES

[1] C.-K. Chen, C. Wang, K.-L. Ma, and A. Wittenberg. Static correlation visualization for large time-varying volume data. In *IEEE Pacific Visualization Symposium*, pages 27–34, 2011.

[2] H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *VisSym*, pages 239–248, 2003.

[3] H. Doleisch, M. Mayer, M. Gasser, P. Priesching, and H. Hauser. Interactive feature specification for simulation data on time-varying grids. *SimVis*, pages 291–304, 2005.

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

[5] M. Glatter, J. Huang, S. Ahern, J. Daniel, and A. Lu. Visualizing temporal patterns in large multivariate data using modified globbing. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1467–1474, 2008.

[6] M. Hadwiger, F. Laura, C. Rezk-Salama, T. Hollt, G. Geier, and T. Pabel. Interactive volume exploration for feature detection and quantification in industrial ct data. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1507–1514, 2008.

[7] M. Haidacher, D. Patel, S. Bruckner, A. Kanitsar, and M. Groller. Volume visualization based on statistical transfer-function spaces. In *IEEE Pacific Visualization Symposium*, pages 17–24, 2010.

[8] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE Symp. on Visual Analytics Sci. & Technology*, pages 3–10, 2010.

[9] C. Y. Ip, A. Varshney, and J. JaJa. Hierarchical exploration of volumes using multilevel segmentation of the intensity-gradient histograms. *IEEE Trans. Vis. Comput. Graphics*, 18:2355–2363, 2012.

[10] H. Janicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Trans. Vis. Comput. Graphics*, 13(6):1384–1391, 2007.

[11] J. Kehrer, F. Ladstädter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser. Hypothesis generation in climate research with interactive visual data exploration. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1579–1586, 2008.

[12] R. Klump and J. Weber. Real-time data retrieval and new visualization techniques for the energy industry. In *Proc. of the Annual Hawaii Int. Conference on System Sciences*, pages 712–717, 2002.

[13] R. Maciejewski, Y. Jang, I. Woo, H. Janicke, K. Gaither, and D. Ebert. Abstracting attribute space for transfer function exploration and design. *IEEE Trans. Vis. Comput. Graphics*, PP(99):1, 2012.

[14] R. Maciejewski, I. Woo, W. Chen, and D. Ebert. Structuring feature space: A non-parametric method for volumetric transfer function generation. *IEEE Trans. Vis. Comput. Graphics*, 15(6):1473–1480, 2009.

[15] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack. Openord: an open-source toolbox for large graph layout. In *Proc. SPIE 7868, Visualization and Data Analysis*, 2011.

[16] S. Mehta, S. Parthasarathy, and R. Machiraju. Visual exploration of spatio-temporal relationships for scientific data. In *IEEE Symposium on Visual Analytics Science And Technology*, pages 11–18, 2006.

[17] R. T. Mills, F. M. Hoffman, J. Kumar, and W. W. Hargrove. Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats. *Procedia CS*, 4:1612–1621, 2011.

[18] T. J. Overbye. Transmission system visualization for the smart grid (panel summary). In *Proc. of Power Systems Conference and Exposition*, pages 1–2, 2009.

[19] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proc. of Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, 2007.

[20] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):99–113, 2005.

[21] F.-Y. Tzeng and K.-L. Ma. A Cluster-Space Visual Interface for Arbitrary Dimensional Classification of Volume Data . In *Proc. of VisSym*, pages 17–24, 2004.

[22] J. Wei, H. Yu, J. Chen, and K.-L. Ma. Parallel clustering for visualizing large scientific line data. In *IEEE Symposium on Large Data Analysis and Visualization*, pages 47–55, 2011.

[23] I. Woo, R. Maciejewski, K. P. Gaither, and D. S. Ebert. Feature-driven data exploration for volumetric rendering. *IEEE Trans. Vis. Comput. Graphics*, 18(10):1731–1743, 2012.

[24] J. Yang, D. Hubball, M. Ward, E. Rundensteiner, and W. Ribarsky. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *IEEE Trans. Vis. Comput. Graphics*, 13(3):494–507, 2007.

[25] Y. Zhang, P. Markham, T. Xia, L. Chen, Y. Ye, Z. Wu, Z. Yuan, L. Wang, J. Bank, J. Burgett, R. Conners, and Y. Liu. Wide-area frequency monitoring network (fnet) architecture and applications. *IEEE Trans. Smart Grid*, 1(2):159–167, 2010.