EUROGRAPHICS 2015 / H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)

Volume 34 (2015), Number 3

VIMTEX: A *V*isualization *I*nterface for *M*ultivariate, *T*ime-Varying, Geological Data *Ex*ploration

A. Dasgupta [1] and R. Kosara [2] and L. Gosink [3]

[1]New York University
[2] Tableau Software [3] Pacific Northwest National Laboratory

Abstract

Observing interactions among chemical species and microorganisms in the earth's sub-surface is a common task in the field of geology. Bioremediation experiments constitute one such class of interactions which focus on getting rid of pollutants through processes such as carbon sequestration. The main goal of scientists' observations is to analyze the dynamics of the chemical reactions and understand how they collectively affect the carbon content of the soil. In our work, we extract the high-level goals of geologists and propose a visual analytics solution which helps scientists in deriving insights about multivariate, temporal behavior of these chemical species. Specifically, our key contributions are the following: i) characterization of the domain-specific goals and their translation to exploratory data analysis tasks, ii) developing an analytical abstraction in the form of perceptually motivated screen-space metrics for bridging the gap between the tasks and the visualization, and iii) realization of the tasks and metrics in the form of VIMTEX, which is a set of coordinated multiple views for letting scientists observe multivariate, temporal relationships in the data. We provide several examples and case studies along with expert feedback for demonstrating the efficacy of our solution.

1. Introduction

A key focus area in geology is the study of chemical reactions in the earth's subsurface and deriving insights about their environmental implications. In the sub-field of bioremediation, scientists aim to reduce soil pollutants through induced reaction among microorganisms and chemical species [DHHP98]. Carbon sequestration is an example of a bioremediation process for reduction of carbon and carbon compounds from soil. These processes involve complex reactions that are difficult to model. Often, scientists need to reformulate their hypotheses based on observations and change experimental settings and parameters. Currently there is a lack of exploratory data analysis tools that support the analysis of the generated experimental data.

Scientists need tools that let them visualize the dynamics of reactions over time, track salient temporal patterns, and detect expected and unexpected behavior of the chemical compounds. To address this need, by collaborating with geologists, we developed a visual analytics solution for letting scientists sift through the data of temporally changing concentrations of different chemical compounds (henceforth referred to as *variables*) and explore and form new hypotheses about the chemical processes.

In the bioremediation domain, we find only one instance of interactive visualization being used for scientific analysis [BCO01], where 3D volume rendering is used to show the different temporal relationships, but multivariate patterns are not captured. Current solutions for visualization of multivariate temporal data [Dol07, AM07, AMM*07] have two key shortcomings. First, there is a lack of explicit analytical abstraction for combining time and multiple variables, which would help scientists guide their attention towards the salient patterns. We explicitly encode [GAW*11] interesting relationships within the visualization for letting scientists efficiently find those patterns instead of sequentially searching for them. Second, existing tools do not adequately capture geologists' intents by letting them focus on variables and variable-pairs and letting them observe multivariate trends as context. Through our interface, scientists can observe how individual variable concentrations change over time and how combination of two or more variable concentrations affect each other.

We have three specific contributions: i) we characterize domain-specific questions and translate them to visual tasks and metrics to quantify important patterns of interest; ii) we extend previously proposed screen-space metrics and develop new metrics for characterizing visual features in relation to the tasks related to multivariate, temporal data analysis, and iii) we present VIMTEX, which is a set of coordinated multiple views (such as meta-level views and data-level views) that lets scientists navigate through different time steps and multiple variables by exploiting visual feedback and guidance based on the metrics.

Goals	Patterns of Interest	Views
Q1 Q2	Univariate distribution	Temporal Summary
Q3 Q4	Multivariate patterns	Detailed View
Q5 Q6	Bivariate Correlations, Aggregations	Temporal Summary
Q7 Q8	Clumping of data points Outlier data points	Detailed View

Table 1: *Translating the domain-specific goals into data properties and views needed to build a visual analysis model. This model guides the metrics encoded within the multiple views in VIMTEX.*

2. Domain Characterization and Task Analysis

In this work we collaborated with a team of geologists, all of them with more than ten years of experience. Our first step was to understand the problem domain through a series of informal interviews and discussions, and derive an analysis model. This model was then used to drive the design of the *VIMTEX* interface. In this section we first provide an overview of our data and then discuss our first contribution of characterizing the domain-specific goals and deriving an analysis model.

2.1. Data Complexity

The complexity of the bioremediation process itself is significant. The simulation begins with the injection of acetate into the subsurface. This injection initially stimulates the growth of the bacteria Geobacteraceae, and this bacteria uses the acetate substrate to reduce iron and aqueous uranium. The reduction of uranium produces a solid called uraninite, so that the reduction process effectively removes uranium from the groundwater. The high level goal of the scientists was to track this reduction process over time, and detect expected and unexpected patterns for the reaction among different chemical compounds. This data originates from a numerical simulation that models a complex, bioremediation [FYM*09] field experiment event that occurred over a two-month interim. The data consists of 10 different chemical compounds, i.e., variables, that are simulated on a 3*D* grid of size 56 X 40 X 43 (approximately 96000 locations), with 120 time steps.

2.2. Goals and Challenges

The high-level analysis questions which we distilled from our discussion sessions can be classified into the following categories:

Q1. Which compounds are stable or unstable over time?

Q2. At which time steps do we see significant rise or fall in concentration of a compound and why?

Q3. When concentration of a compound decreases or increases, what is the effect on others?

Q4. What is the effect of reaction between two compounds on others?

Q5. Which sets of compounds show similar levels of concentrations over time and why?

Q6. At which time steps do we see similar concentration levels and high reactivity? Are they expected?

Q7. Do the compounds behave similarly in particular locations?

Q8. Are reaction dynamics different in certain locations than others?

From these questions, we observed that the focus of analysis was on observing the temporal dynamics of the reactions, and how different locations respond to those reactions. Our initial interactions revealed that temporal summary of the relationships was essential in driving the analysis process. While a spatial view would explicitly reveal the location of the cells, our collaborators were initially not focused on identifying the locations, but on tracking if overall patterns were different for some of the locations.

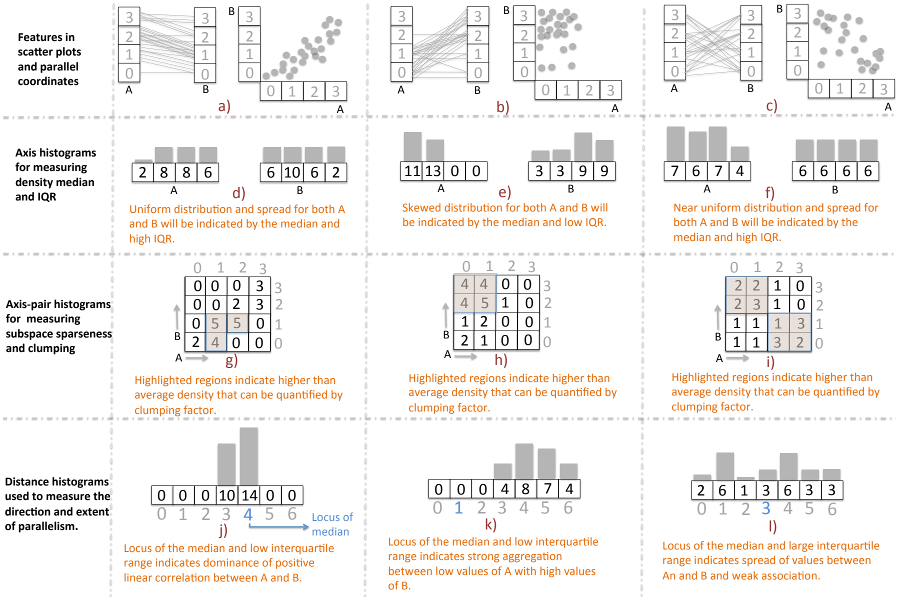
There were some key challenges with respect to visualization design addressing the analysis goals. The most important one was to reduce visual complexity in showing both multivariate and temporal patterns. The size of the data was too large to fit in a visualization display with a high-precision representation. Since this would lead to clutter, a level of abstraction was necessary. However, scientists were not keen on a very high level of abstraction, as they would like to drill down and identify if some of the cells behaved differently than the overall patterns. Therefore, the most important guiding factor of our design was to let scientists perform an efficient visual search for patterns of interest by gaining an overview of salient temporal relationships, and then drill down into subsets of data points and explore multivariate patterns in more detail.

2.3. Analytical Abstraction

For handling the data complexity and for facilitating an efficient visual search, our initial design decisions were centered upon identifying the key patterns of interest. This let us build a model based on screen-space metrics for expressing the patterns through multiple views of both the metrics and the data.

Patterns of Interest: As shown in Table 1, the main patterns of interest were univariate distributions, multivariate patterns and bivariate relationships like correlation and aggregation. While answering most of these questions scientists wanted to compare the behavior of one variable with respect to all others, or the reaction of two variables with respect to the rest. In most cases, they wanted to resolve the questions Q1, Q2, and Q3, Q4 simultaneously. For example, if they found the concentration of a compound falling at a particular time step, they would like to immediately see the effect on the other compounds. These led us to select variables and variable-pairs to be the building blocks of our system. Using these building blocks, we aimed to build a temporal summary view, which would capture the evolution of different relationships over time; and a detailed view showing the multivariate properties at any given time step. This was also necessary to answer Q7 and Q8, where scientists wanted to look at clusters and outliers.

Views: To satisfactorily answer the questions, we needed a multivariate data view, which could show the patterns among multiple variables at once, while allowing scientists to focus on variables and variable-pairs. To address these design issues we used parallel coordinates axes and axis pairs as the basic building blocks for showing, multivariate patterns. Previously, researchers have demonstrated that while analyzing multivariate data, users initially look for distributions and patterns between two variables [CvW11]. In the context of parallel coor-



Uniform distribution and spread for both A and B will be indicated by the median and high IQR.

Skewed distribution for both A and B will be indicated by the median and low IQR.

Near uniform distribution and spread for both A and B will be indicated by the median and high IQR.

Highlighted regions indicate higher than average density that can be quantified by clumping factor.

Highlighted regions indicate higher than average density that can be quantified by clumping factor.

Highlighted regions indicate higher than average density that can be quantified by clumping factor.

Locus of the median and low interquartile range indicates dominance of positive linear correlation between A and B.

Locus of the median and low interquartile range indicates strong aggregation between low values of A with high values of B.

Locus of the median and large interquartile range indicates spread of values between An and B and weak association.

Figure 1: *For designing VIMTEX we use different histograms for quantifying visual properties, common to scatter plots and parallel coordinates. As shown, each cell represents a pixel bin. Numbers within each cell in the first row represents the pixel coordinate of the bin and the numbers within each cell in the other rows represent the frequency of the bin.*

dinates, it has been argued [LMvW10] that users are mainly looking for patterns in the two-dimensional space between the axes. This is consistent with the scientists' analysis goals. The design of VIMTEX is centered around a parallel-coordinates based coordinated multiple views and metrics which let scientists shift between overview and details of multivariate, temporal patterns (Figure 2). It could be argued that for showing bivariate correlations, scatter plots or a scatter plot matrix is a better choice than parallel coordinates. However, other criteria, like looking at multivariate patterns and tracking anomalous cells across different variables, are better addressed by parallel coordinates. We consider this as a design trade-off and eventually our choice of a detailed view was using parallel coordinates for the aforementioned reasons. For reflecting temporal change of distribution of variables and relationship of variable pairs, we used standard time-series plots. In this case, the time series would not reflect the data, but metrics computed on the data.

Screen-Space Metrics: We decided to use screen-space metrics for computing patterns of interest. An alternative was to compute statistical metrics based on the data and represent their results visually. However, given the size of the data, computing metrics in the data space, was less efficient than computing them in the screen space. To alleviate that problem, we decided to first convert the data points to pixel-coordinates. This would make the computation faster as the time complexity would be independent of the data size. Since we were using pixel coordinates, we used existing screen-space metrics and devised new ones, as those were also perceptually more beneficial to the scientists.

The screen-space metrics are computed through a frequency-based representation of the data, where this representation is determined through an output-oriented binning technique [NH06] that relies on pixel binning. We adopt this strategy for two reasons: i) pixel binning directly addresses scalability issues that can arise when working with large numbers of data points, and ii) the screen-space metrics based on pixel bins directly reflect what the users see on screen ([RJTTJ03, AdOL04]). This second benefit is especially important in our work as our principal objective is to establish a direct correspondence between the perceptually motivated metrics and the pixel-based representation on screen.

There are alternatives to using a discrete model. For example, continuous data models, many based on a variety of density estimation strategies, can provide a highly accurate representation of high-dimensional data [HW09, HBW11]. These approaches can markedly reduce the data clutter that is inherent to visualizing a large number of points. In comparison to discrete models, however, continuous data models tend to have a higher computational complexity and their resulting density-based visual representations may not be intuitive to domain scientists. Additionally, it can be challenging with continuous data models to directly link our metrics with the pixel-based representation. In our approach, we deal with the problem of clutter by enabling metric-based brushing. When there is too much clutter, the metrics can be used for brushing, and those brushes would not only show subsets, but also reveal relationships like aggregation, outliers, etc. Using standard statistical metrics, this would be difficult. For example, we can quantify correlation using Pearson's correlation coefficient between two variables. But often those correlations are only exhibited by record subsets. When we use a screen-space measure, it becomes easier to parameterize the brushing mechanism and directly see how certain records (in this case spatial cells) behave differently from the rest.

3. Related Work

We discuss the related work in the context of analytical abstractions for multivariate temporal data analysis and the different variants of parallel coordinates for this purpose.

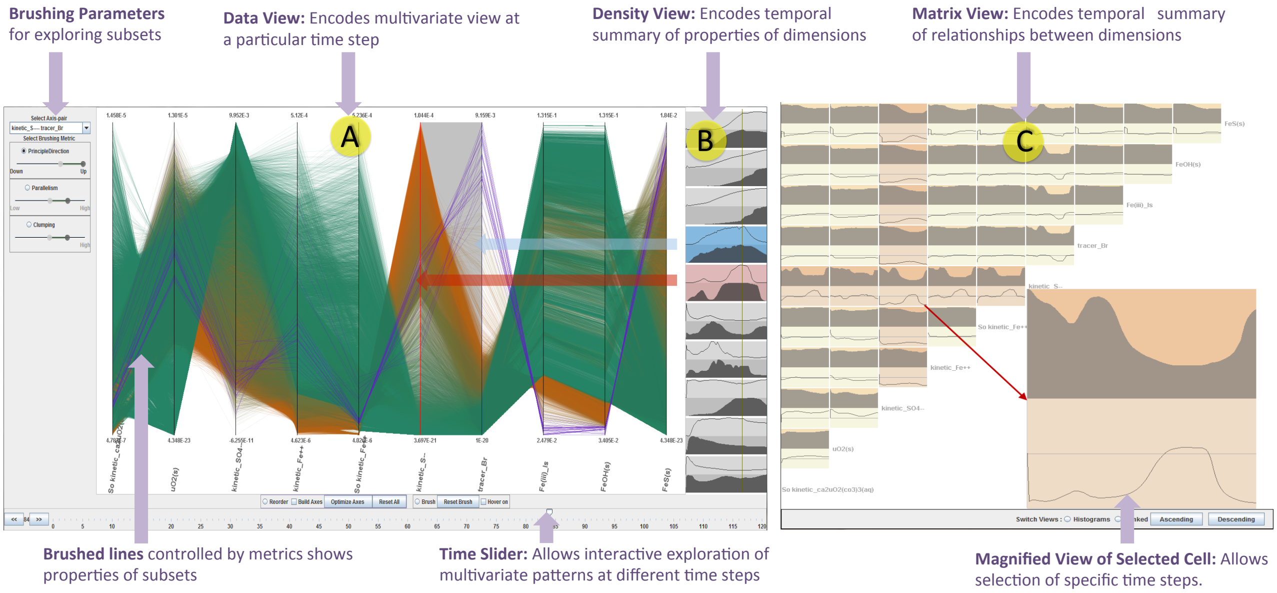

Brushing Parameters for exploring subsets
Data View: Encodes multivariate view at a particular time step
Density View: Encodes temporal summary of properties of dimensions
Matrix View: Encodes temporal summary of relationships between dimensions
A
B
C
Brushed lines controlled by metrics shows properties of subsets
Time Slider: Allows interactive exploration of multivariate patterns at different time steps
Magnified View of Selected Cell: Allows selection of specific time steps.
Select Axis-pair
Select Brushing Metric
PrincipleDirection
Parallelism
Clamping
Reorder
Build Axes
Optimize Axes
Reset All
Brush
Reset Brush
Hover on
Switch Views
Histograms
Ascending
Descending
So kinetic_ca2uO2(co3)3(aq)
uO2(s)
kinetic_SO4--
kinetic_Fe++
So kinetic_Fe++
kinetic_S--
tracer_Br
Fe(III)_ls
FeOH(s)
FeS(s)

Figure 2: *The different components of VIMTEX are: A)* ***Data view****, which is a multivariate, time-varying view of the data and reorderable, B)* ***Density view****, which shows univariate temporal distribution with the selected axis pair being highlighted, and C)* ***Matrix view*** *which shows the bivariate correlations as time-series.*

3.1. Analysis Models for Multivariate, Temporal Data

Multivariate temporal datasets require analytical abstractions for capturing the dynamic relationship among multiple variables. Integrating computational and visual aspects [GCML05] by using clustering based visualization techniques for modeling similarity-based temporal behavior have been proposed, where parallel coordinates have been applied for visualizing the results of clustering. Pre-processing of data for extracting trend sequences and subsequent visualization of those temporal trends through parallel coordinates have also been used [LS09]. Functional temporal plots for visualizing changes in correlation in a matrix layout have been used in the context of gene-sequence modeling [MMDP10]. These analytical abstraction methods focus on extracting information from the data and then using visualization tools for communicating that information to the user. In TimeSeer [DAW13], metrics based on scatter plot properties are used for guiding users. These metrics are however, not motivated by high-level analysis goals of domain experts. Also, for a large data size as in our case, using the metrics for showing the data while reducing clutter is a non-trivial challenge. In our approach, we use perceptually motivated screen-space metrics for guiding the scientists towards answering their analysis questions, by finding the salient patterns which can be hidden among a subset of the data points. The system presented by Glatter et al. [GHA*08] follows a similar principle, where a domain scientist specifies uncertain temporal patterns using a description language, and patterns can be formulated as queries using this language. Extracting importance based relationship using information theoretic metrics to describe the visual structures [WYM08] is another example of such an approach.

3.2. Temporal Parallel Coordinates

Parallel coordinates [ID90] has been a popular technique for visualizing scientific data and several variants have been proposed. One of the examples is the application for hurricane data analysis [SSJKF09] where statistical properties are mapped on to the parallel coordinates axes. In the interface proposed by Akiba and Ma [AM07], multivariate connections can be brushed using a parallel coordinates interface, which in turn is linked to time histograms and a direct volume rendering of selected attributes. But as observed in the case of tiled parallel coordinated display and min-max plots [CMR07] applied to time-varying EEG data, the overall temporal distribution is not conveyed in this approach. Johannson et al. use depth cues and variation of opacity to show temporal properties in parallel coordinates [JLC07]. This approach suffers from clutter in case of large number of time steps and data-points. Our approach is to look at the different recognizable visual features and use appropriate metrics to convey them. Blaas et al. [BBP08] use data quantization and compression to handle large number of data points in the context of parallel coordinates. While this works well at the overview level, detailed exploration of features is difficult using this approach. Frequency-based representations like use of angular histograms [GPSL*11] is similar to our approach. In our work, we apply some of the visualization design principles for simulation data as outlined by Doleisch et al. [Dol07], with focus on the exploration and analysis aspects. We have also incorporated some of the visualization strategies for dealing with time-series data [MS03], where we address the key issues of finding the temporal patterns and understanding the change in behavior over time through linking of different views of the data.

4. Computation Model

In VIMTEX the focus is on reducing the visual uncertainty due to pattern complexity [DCK12] by using metrics that describe the visual structures, thereby establishing a direct correspondence between the behavior of the metrics and structural change on screen. The computation model is based on pixel binning and a set of screen-space metrics. The metrics connect the domain specific intents to the visual features within parallel coordinates. They are generally applicable to scatter plots and parallel coordinates.

4.1. Data density

For quantifying univariate distribution, our goal was to characterize the data density in terms of the locus (where, on the axis, most data values are located) and randomness (amount of disorder among the values). For this purpose we propose two metrics for characterizing the data distribution, the *density median* and *interquartile range (IQR)*, both of which are computed from the one-dimensional axis histogram. The pixel coordinate of the density median is indicative of the degree of skewness of an unimodal distribution.

Computation: Density median μ is computed from the median of the frequencies of the pixel-bins in the one-dimensional axis histogram. The location, that is the pixel coordinate of the median β_μ, is then plotted over time. A high value of the median at a particular time step means dominant values at that time step are the high ones and a low value means dominant values are the low ones. Similarly, high IQR will indicate a spread of values and low IQR will indicate more concentrated values.

Implication: In Figures 1d and 1e the non-uniformity of the distribution can be quantified by axis IQR. The distributions being more skewed in Figure 1e, IQR will be lower than in Figure 1d for both axes A and B. Locus of the median towards the middle of the histogram in Figure 1d indicating an almost normal distribution. Locus of the median towards the ends of the histograms in Figure 1e indicates skewness of the distribution. Uniform distribution in Figure 1f can be indicated by high IQR.

4.1.1. Correlation and aggregation

For quantifying the change in temporal correlation and aggregation between adjacent axes, we apply the parallelism metric [DK10]. Parallelism among lines can reflect correlation and aggregation in parallel coordinates. Since the metric is essentially based on a distance histogram, it can also be directly applied to scatter plots. As shown earlier in Figure 1, the parallelism metric [DK10] can imply different relationships between data dimensions: positive linearity, aggregation and lack of linearity. The parallelism metric is composed of two elements: the range depicting the degree of correlation (if most data points conform to the trend or are loosely scattered) and the median or principal direction from left to right between two axes indicating if parallel lines are going up, down or staying horizontal. The direction is important in indicating association between high and low values.

Computation: To compute parallelism, a distance histogram is first constructed that records the distribution of pairwise vertical distances between data points on adjacent axes. From this histogram, the median distance value indicates the direction of parallelism, if lines are staying horizontal or going upward or downward. The direction is given by the median M_P, which is not normalized (the direction only makes sense in pixel coordinates): $M_P = q_{50}$. Here q_{50} represents the 50% quartile of the distance distribution.

The extent of parallelism is given by the interquartile range: a narrow interquartile range implies high parallelism. We normalize the distances between 0 and 1, by dividing by the highest possible distance. We then compute parallelism P_{norm} as follows based on the interquartile range between the 25% and the 75% quartiles, q_{25} and q_{75}, given by $P_{\text{norm}} = 1 - |q_{75} - q_{25}|$. The subtraction is done to get a higher parallelism value for a higher degree of parallelism (and thus a smaller interquartile range). M_P and P_{norm} are used to draw line-plots for the median and range for all the time steps that models the temporal trends over time. These are illustrated in Section 5.3.

Implication: The locus of the median of the distance histogram indicates the nature of association between A and B. As shown in Figure 1j, the median being in the centre and interquartile range being low, these imply positive linear correlation between A and B. Similarly, in Figure 1k, locus of the median towards the right of the histogram and low interquartile range indicates strong association of low values of A with high values of B, implying aggregation. On the other hand, in Figure 1l, higher interquartile range indicates a more spread of the values between A and B and thus implying a weaker aggregation. Higher interquartile range can also reflect negative correlation. Users can use brushing based on parallelism to confirm which of the two features is dominant.

4.1.2. Clumping

Converging and diverging structures have been shown to be interesting structures [DK10] in parallel coordinates. However, in case of large datasets, lines do not converge to (or diverge from) a precise pixel, but converge to a local neighborhood of pixels. We compute the density of these clumped spaces or neighborhoods by applying density-based clustering in the binned space and computing the *clumping factor*. The advantage of clustering in the screen-space, based on density of pixel bins is it relates exactly to what the user sees on screen: semantics is generated from the visualization, rather than the other way round. Computation of clumping factor requires two parameters: the number of contiguous neighbors, and the density of a neighborhood that is considered as clumpiness. The steps are described below.

When there are multiple modes, median alone is not an accurate estimator of density. A multimodal distribution implies data has a higher likelihood of being clumped in sub-spaces. To explore if there are hidden clusters that are not distinguishable from the overall patterns, we have developed a clumping metric, based on the two-dimensional distribution based on an axis pair. This is especially of relevance for temporal data, because at many time steps, certain data-points tend to cluster/clump together in local neighborhoods. Clumping metric captures this behavior.

Computation For determining the number of contiguous neighbors, we first compute sparse regions based on the frequency of the pixel bins. For this we use a two-dimensional axis-pair histogram. A clumping cluster is cut off once a sparse region is found. To compute clumping regions, we begin with the convergence-divergence metric. Let us denote a sparse bin by be indicated by $\bar{\beta}_i$.

Our clumping algorithm first selects the axis with higher average convergence/divergence, and iterate through the pixel-bins following the two subroutines that are described as follows:

Sparse Regions:

1. Set threshold for clumpiness to number of records divided by number of bins, i.e. $t = \frac{n}{h}$.
2. If the frequency of a bin is less than the quartile of the threshold, i.e., $\beta_{i,j} < 0.25t$, consider the bin as a sparse bin.
3. Compute the number of contiguous sparse bins (η) for each sparse region found.
4. Compute average sparseness as the number of contiguous sparse bins divided by the number of sparse regions (v). So we get

$$avg(\eta) = \frac{1}{v}\sum_{i=1}^{v}\eta_i$$

Average Clumping: Let the number of contiguous clumped bins be denoted by ζ.

1. If the frequency of a bin is greater than t, add it to the cluster.
2. Continue adding bins until a sparse bin ($\bar{\beta}_i$) is found.
3. If $v < avg(\eta)$ add bins to the cluster, else break.
4. Repeat steps 1 to 3 for all bins.
5. Add the number of bins in each clumped cluster. Divide by the number of such clusters (e). That is the clumping factor (CLF).

$$CLF = \frac{1}{e}\sum_{i=1}^{e}\zeta_i$$

Implication: As observed in Figure Figures 1g, 1h, and 1i, the threshold (t) for clumpiness is set to 6. In Figure 1g, there is one bin with clumpiness surrounded by two sparse regions on either side. The clumping factors in this case is 10. For the case illustrated in Figure 1h, once a clumped bin with frequency 6 is found, the next one with frequency 5 is added, and then the cluster is cut off because of the sparse region with higher than average sparseness. Thus in this case there is a clumped cluster with two bins, with CLF 6. In Figure 1i, there is no clumping, but high sparseness.

5. Multiple Views

We use the screen-space metrics as the basis for designing coordinated multiple views (Figure 2) that show different properties of the data. The detailed or the data view (A) based on parallel coordinates is complemented by two temporal summary views, or meta views (B, C), where metrics abstracted in the form of time-series, are displayed. Interaction with these views enables analysts to seamlessly navigate through interesting time steps and data dimensions.

The views facilitate exploratory data analysis based on: a) getting an overview of the temporal trends that evolve over time and b) interactively explore patterns at time steps of interest and drill down to details. Our design of the coordination among the views follow Schneiderman's visual information seeking mantra [Shn96] by enabling the analysts to seamlessly switch between gaining overview and exploring details.

The main data view (Figure 2A) provides a multivariate representation of the data at different time steps. The metrics are encoded within the meta views, which are i) the density view (Figure 2B) shows the temporal behavior of each axis, and ii)the matrix view (Figure 2C) that shows the pairwise temporal behavior of the axes.

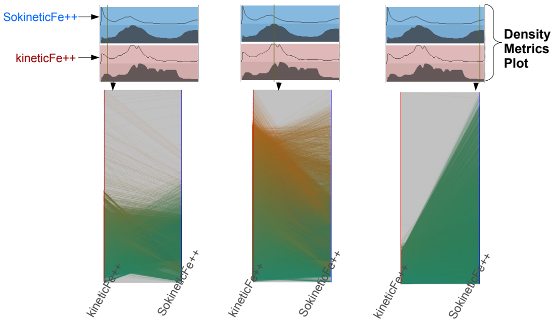


Figure 3: ***Illustrating the data density metrics****: Three different configurations of the axis pair (kinetic iron and sorbed kinetic iron) on interaction with the density view. Each blue box represents the density median plot for sorbed kinetic iron and the red box represents the same for kinetic iron. Low IQR clearly leads to more recognizable patterns due to low clutter.*

5.1. Density View

The density view is composed of sets of vertically stacked boxes (Figure 2B), where each box corresponds to a dimension and contains a line-plot and an area-plot. The line plot represents the density median and the area represents the IQR, time-axis being horizontal. The configuration of this view is invariant to the order of axes in the data view. Different colors (red for left axis and purple for right axis) are used to represent the selected axes. Even without interaction, the IQR and the pixel coordinate of the density median plots give an overview of which variables are stable or unstable.

5.2. Data View

The data view helps in addressing *Q1* by enabling the analysts to explore the features shown by the density view, that is, going from overview to the details. In this view we show the default parallel coordinates layout for our dataset, as shown in Figure 2A. The configuration is synchronized with the time-slider.
Global Scaling: We choose a global scaling for the variables, i.e., we compute the maxima and minima for the different dimensions over all the time steps and scale the data points accordingly. This helps us in handling ranges that can vary a lot from an initial time step to a later time step in a particular domain, which enables us to show how trends change within a fixed data range.
Color Gradient: We use a continuous color gradient of green to brown, to indicate the transition from low to high values on a particular axis. The application of color gradient to an axis, and not an axis pair, is motivated by the scientists' goal of observing the variance in concentration of the different variables with respect to a specific variable. The color gradient is applied to the left axis within the axis pair selected by a user, which is the pair with kinetic sulfide (kinetic S^{--}) and tracer bromine (tracer Br) in Figure 2A. The higher concentration of green lines on most other axes give an overview of difference in high and low concentrations of multiple variables even without adjacency. This also helps in reducing clutter when there is a large number of line crossings, thereby making the trends stand out. A purple color is used to indicate the brushed lines.

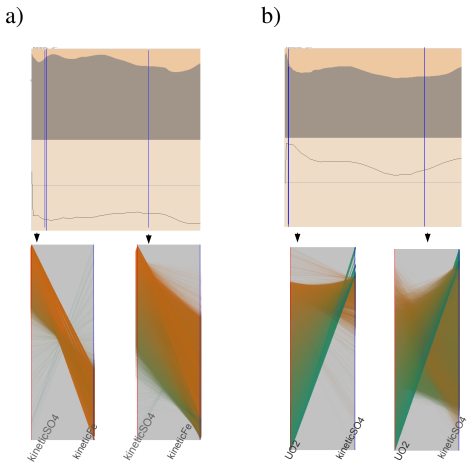


Figure 4: *Illustration of the parallelism metric, for two different axis pairs. In a) parallelism reflects aggregation, while in b) parallelism reflects positive linear correlation.*

5.3. Matrix View

A problem with parallel coordinates is that ordering of the variables has to be effective enough to convey the different conceivable properties that exist. This becomes an even bigger challenge for temporal data, because it is difficult to track the temporal pattern of all combinations of variables with the default layout. To address this issue, we build a matrix layout (Figure 2C) similar to a scatter plot matrix, and show the parallelism range and median plots in that view. In Section 6.2 we describe in detail how this view can be used for answering the questions about bivariate relationships.

6. Visually Guided Analysis

VIMTEX visually guides domain experts in finding patterns of interest in the data using the computation model described in earlier Section 4. The meta-level temporal summary views and the detailed data view of parallel coordinates helped our geologist collaborators in analyzing the data from multiple perspectives to answer their questions (Q1-Q8). In this section we highlight the functionality of VIMTEX by describing *how* our collaborators could use the metrics and views for answering these questions. Later in Section 7 we describe in more detail *what* they could find using VIMTEX and *why* they were important from a geological perspective.

6.1. Identifying Stable and Unstable Behavior (Q1, Q2)

In the analytical context, we define stability as the degree to which the data distribution remains unchanged over a period of time. Q1 and Q2 are both addressed using the metrics for data density by getting an overview from the density view, that are discussed below. By linking the density view with parallel coordinates, geologists could see compare the rise and fall in concentration at specific time steps to the overall temporal patterns. By focusing on variables and looking at multidimensional patterns, they could also answer Q3 and Q4.

As illustrated earlier in Figure 1, the pixel coordinate of density median and IQR are computed from the one-dimensional axis histogram. In Figure 3 we show a specific use case scenario where these metrics are useful in answering Q1 and Q2.

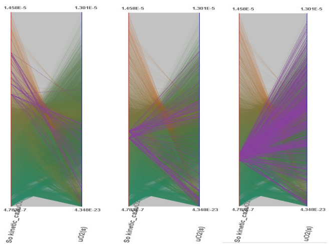


Figure 5: *Different degrees of clumping exhibited between uranium carbonate and uraninite at subsequent time steps. The clumping metric (C_f) returns a higher value when there are more clumped regions with higher density as in the rightmost image.*

The three boxes and their parallel coordinates representation are for three different time steps for kinetic iron and sorbed kinetic iron. For the first case, kinetic iron (red box) exhibits low value for the pixel coordinate of density median and low IQR, and same for sorbed kinetic iron (blue box). This is represented by highly dominant and less dispersed green lines (signifying dominance of low data values). In the second case pixel coordinate of density median for left axis is higher and IQR for both axes is higher. This is demonstrated by more brown lines, that signifies higher concentration of kinetic iron being dominant. While high IQR on both axes leads to a high dispersion among the lines. In the last case, low value for the pixel coordinate of density median and low IQR are indicated by the dominant green lines originating from kinetic iron. The difference from the first case is the lines are highly dispersed, shown by high IQR value for the right axis.

6.2. Exploring Bivariate Relationships (Q5,Q6)

The geologists were interested in observing bivariate relationships such as similar levels of concentration and reactivity among different combination of variables. In most cases they wanted to focus on variable-pairs and then look at the multidimensional patterns for answering Q3 and Q4. These properties are quantified in terms of correlation and aggregation between axis pairs in parallel coordinates, using the parallelism metric. The advantage of using the parallelism metric is that even if the aggregation or correlation patterns are not perceptible for the variable pairs for all records, using brushing based on the direction and extent of parallelism, one could find those relations easily. Intelligent brushing based on relationships rather than only data points, helped geologists find interesting hidden patterns.

An illustration of the parallelism metric is shown in Figure 4. The upper part of the box which is a filled area, shows spread of the parallel lines denoted by P_{norm}. Line plot in the bottom one depicts their direction, denoted by M_P. A large area under the curve corresponds to high parallelism (high P_{norm}) , i.e., less spread and smaller area means less parallelism (lower P_{norm}), i.e. more spread. The lower part of the box shows the line plot for M_P. An indicator horizontal line through the middle of the lower part indicates an M_P value of 0, i.e., lines remaining horizontally parallel to each other between adjacent axes. The line plot going above the indicator line denotes most lines in the parallel coordinates plot going upward ($M_P > 0$) and when below the indicator line, that denoted most lines going downward ($M_P < 0$).

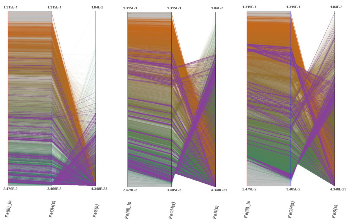


Figure 6: *Clumping pattern among Iron sillicate, Iron hydroxide and Iron sulphide that was an unexpected phenomena according to the scientists' initial hypothesis.*

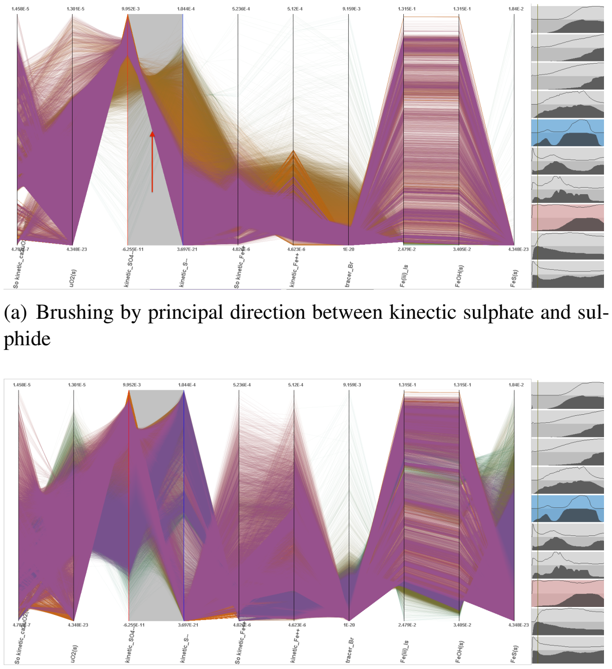
In Figure 4a, P_{norm} is high initially and M_P is much less than zero. So we see many lines going downward from kinetic sulfate to kinetic iron. This indicates many locations having similar high concentration of kinetic sulfate and low concentration of kinetic iron, and a correlation between these two. Later on, P_{norm} exhibits lower value but M_P increases slightly. Therefore we see more spread of the lines going downward. In Figure 4b, P_{norm} is high initially and M_P has a very high value, reflecting a strong aggregation of lines going up from uraninite to kinetic sulfate. In the second time step, M_P drops, reflecting a spread of more lines going downward, implying a lack of aggregation, and thus the concentration levels of uraninite being dissimilar over time.

6.3. Exploring Clumping of Concentration Levels (Q7)

The clumping metric described in Section 4.1.2 was especially important in judging the reactivity of different variables. The high-level goal being observing dynamics of reactions, a high degree of clumping between two variables usually meant low reactivity between them. Using this metric scientists could also analyze if certain locations exhibit similar levels of concentration than others (Q7) and if they are outlier locations (Q8). Axis pairs with different degrees of CLF are shown in Figure 5. The greater the clumping factor, the more the number of local neighborhoods with larger number of lines converging to or diverging from those neighborhoods. Brushing by high clumping on an axis pair enables an analyst to visualize the behavior of the clusters across multiple axes and examine the temporal change of those neighborhoods (Q3,Q4,Q6). In Figure 5a, we can observe only a few cell locations exhibiting a clumping pattern, indicating high reactivity between uranium carbonate and uraninite. However, as uranium carbonate got depleted over time, we can observe more cell locations exhibiting similarly low levels of concentration indicated by clumping of low values in Figure 5b and Figure 5c. The converging structures also indicated a low level of reactivity among uranium and uraninite over time. Thus the clumping metric was of great help to the geologists to explore subsets of variable concentration values and track their properties over time, which would otherwise be difficult due to clutter.

7. Case Study

In this section we demonstrate the usability of VIMTEX through one of the case studies that our collaborators performed. As this simulation is made up of over 29 unique reactions that are temporally distinct, we limit our discussion to a few selected motivating examples. The specific goals of were to identify the utility of our analysis methods to accomplish a subset of their original analysis questions: Task A) identify stable and unstable trends in variables in order to help verify certain aspects of the process models used to generate the data, and Task B) build visual evidence to help confirm certain hypotheses regarding the interactions between variables (specifically uraninite, sulfates and sulfides chemical species) and discover the trends and anomalies, if any. Task A mainly focused on answering questions Q1 and Q2. Task B mainly addresses the questions Q5–Q8. Q3 and Q4 could be combined with any of the other questions for looking at the multivariate trends.



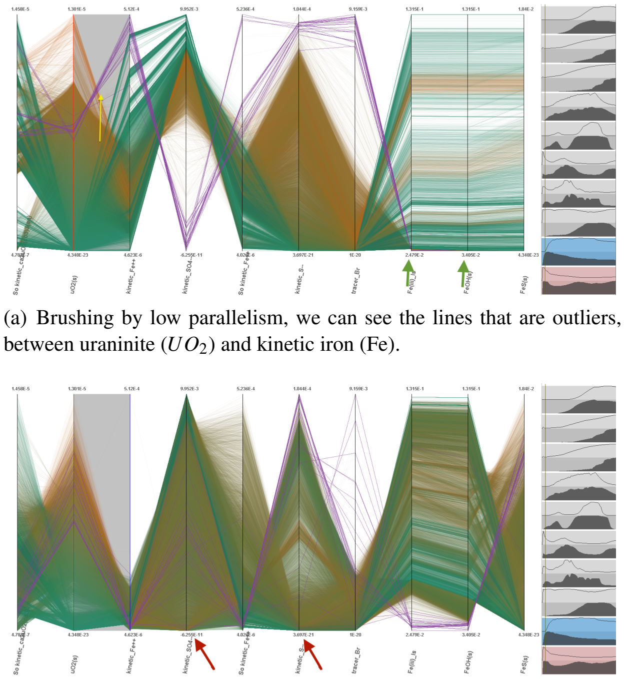
(a) Brushing by principal direction between kinectic sulphate and sulphide

(b) Sulphate concentrations fall and sulphur concentrations rise as the reaction approaches completion

Figure 7: *Brushing by principal direction enables an analyst to observe the association between low and high values of adjacent dimensions and then observe the multivariate patterns.*

Addressing Q1 and Q2: We began analysis by using the density view to provide a high-level overview of univariate data distribution for all variables. Addressing Task A (identifying stable and unstable variables), the density view in Figure 8 shows certain chemical species, like iron silicate, iron hydroxide, and iron sulphide maintain significant stability throughout the bioremediation process as indicated by their unchanged uniform distributions over time. In contrast, other chemical species display significantly more instability, indicating these variables are more involved in the remediation process (e.g., kinetic sulfate and kinetic sulfide). The scientists found that the additional metric views, apart from the main parallel coordinates view, provided a nice overview for them to gauge if what they saw matched their expectations.

Addressing Q5 and Q6: Next, we inspect the bivariate distributions between iron sillicate and hydroxide (matrix view in Figure 2C); In these distributions, note the strong parallelism between these species that remains more or less unchanged over time. This stable trend in parallelism indicates strong correlation throughout the simulation. This property was specifically



(a) Brushing by low parallelism, we can see the lines that are outliers, between uraninite (UO_2) and kinetic iron (Fe).

(b) The same data points that showed up as outliers now show similar patterns to the majority trend between uraninite (UO_2) and kinetic iron (Fe^{++}) towards the end of the simulation.

Figure 8: *Gaining an overview from the bivariate view and the univariate view allows an analyst to select outliers in the data view. Density view indicates stable and unstable variables. Indicated by green arrowhead: iron sillicate and iron hydroxide are the stable variables. Kinetic sulfate and sulfide are the unstable variables as indicated by the red arrow.*

of interest to the scientists who further explored the subspace between iron sillicate and hydroxide by brushing according to clumping. As shown in Figure 6, this brushing indicates strong clumping patterns between iron sillicate, hydroxide, and sulphide that remain largely unchanged across the temporal axis. The scientists expressed surprise on seeing this trait and identified this as significant, in that stable clumping (in this instance) implies a low level of reactivity between these variables that is indicative of initialization errors in the simulation itself. This was one of the major findings from using our tool, which the scientists where not aware of before.

Combining Q1 and Q2 with Q5 and Q6: To address Task B (build and confirm hypotheses), we look at the locus of density median plots (Figure 8) to get an overview of the interactions between sulfates and sulfides. In this figure, the corresponding boxes show gradually rising median trends for iron sulfide (Fes) and kinetic sulfide, while sulfate concentrations mostly remain high. From this visual information, the scientists hypothesized that cells with initially low sulfur concentrations would exhibit a rise in these concentrations, especially with respect to kinetic sulfate. This hypothesis was confirmed by brushing by principal direction (Figure 7(a)) where a selection of downward trend shows a strong cluster between high concentrations of kinetic sulfate, kinetic sulfide, and iron sulfide. This trend slowly gives way to more random patterns and the rising concentrations of sulfur are reflected with scattered brushed lines (Figure 7(b)). Our collaborators therefore confirmed their hypothesis and also concluded that concentrations of sulfate species become depleted in the middle of the reaction, and begin to rise again towards the end of the reaction.

Addressing Q7 and Q8: For examining anomalies, we select the axis pair involving uraninite and kinetic iron. Both density view and parallelism views show strong initial downward parallelism from uraninite to kinetic iron. We examine the behavior of outliers by brushing using low parallelism as the criteria which showed association between high values on both axes (Figure 8(a)). While this was flagged as a potential anomaly, the cells conformed to the expected trend of association between high values of uraninite and low values of iron, towards the end of the simulation, leading the scientists re-affirm their reaction model. This also increased their trust in using the metrics and the visualization for their analysis.

Feedback: From the first session we got feedback that were used for improving the usability of the tool, that got subsequently evaluated in the following session. The scientists found the idea of using information visualization tools to examine behavior of the variables, to be novel and commented that the tool could be useful even in cases where they do not know the model apriori. While it took them some time for grasping the concept of using visual abstractions and coordinated multiple views for showing salient relationships, with time they became comfortable with using the different views, and could visually detect patterns and outliers. They particularly appreciated how efficiently they could find the interesting patterns, which would help them form new hypotheses and conduct further experiments for running the simulation.

8. Conclusion and Future Work

Effectiveness of VIMTEX demonstrates how screen-space metrics can be applied in practice, which is a relatively new research direction. Assisted by the metrics, geologists could not only confirm their existing hypotheses about the chemical reactions in bioremediation, but also form new hypotheses through their interactions with VIMTEX. Our collaborators concluded that even for unknown data VIMTEX would be effective to build their hypotheses about the reactions in the simulation. Encouraged by the results, we want to conduct a formal user study in the future for comparing our tool with existing approaches. We will also apply our approach based on screen-space metrics to multivariate temporal datasets from other domains such as stock market analysis, cyber security, etc.

9. Acknowledgment

The research described in this paper is part of the Carbon Sequestration Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. The data in this paper was provided by the U.S. Department of Energy, Office of Science, Subsurface Biogeochemical Research Program through the Integrated Field Research Challenge Site (IFRC) at Rifle, Colorado.

References

[AdOL04] Artero A. O., de Oliveira M. C. F., Levkowitz H.: Uncovering clusters in crowded parallel coordinates visualizations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (2004), IEEE, pp. 81–88. 3

[AM07] Akiba H., Ma K.: A tri-space visualization interface for analyzing time-varying multivariate volume data. In *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization* (2007), pp. 115–122. 1, 4

[AMM*07] Aigner W., Miksch S., Müller W., Schumann H., Tominski C.: Visualizing time-oriented dataâĂŤa systematic view. *Computers & Graphics 31*, 3 (2007), 401–409. 1

[BBP08] Blaas J., Botha C., Post F.: Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1436–1451. 4

[BCO01] Baracca M., Clai G., Ornelli P.: Simulation and 3d visualization of bioremediation interventions in polluted soils. In *High-Performance Computing and Networking*, Hertzberger B., Hoekstra A., Williams R., (Eds.), vol. 2110 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2001, pp. 535–538. 1

[CMR07] Caat M., Maurits N., Roerdink J.: Design and evaluation of tiled parallel coordinate visualization of multichannel eeg data. *IEEE Transactions on Visualization and Computer Graphics 13*, 1 (2007), 70–79. 4

[CvW11] Claessen J. H., van Wijk J. J.: Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics 17* (2011), 2310–2316. 2

[DAW13] Dang T. N., Anand A., Wilkinson L.: Timeseer: Scagnostics for high-dimensional time series. *Visualization and Computer Graphics, IEEE Transactions on 19*, 3 (2013), 470–483. 4

[DCK12] Dasgupta A., Chen M., Kosara R.: Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum* (2012), 1015–1024. 4

[DHHP98] Dojka M. A., Hugenholtz P., Haack S. K., Pace N. R.: Microbial diversity in a hydrocarbon-and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Applied and Environmental Microbiology 64*, 10 (1998), 3869–3877. 1

[DK10] Dasgupta A., Kosara R.: Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1017–26. doi:10.1109/TVCG.2010.184. 5

[Dol07] Doleisch H.: Simvis: Interactive visual analysis of large and time-dependent 3d simulation data. In *Proceedings of the Winter Simulation Conference* (2007), IEEE Press, pp. 712–720. 1, 4

[FYM*09] Fang Y., Yabusaki S. B., Morrison S. J., Amonette J. P., Long P. E.: Multicomponent reactive transport modeling of uranium bioremediation field experiments. *Geochimica et Cosmochimica Acta 73*, 20 (2009), 6029 – 6051. doi:DOI:10.1016/j.gca.2009.07.019. 2

[GAW*11] Gleicher M., Albers D., Walker R., Jusufi I., Hansen C., Roberts J.: Visual comparison for information visualization. *Information Visualization 10*, 4 (2011), 289–309. 1

[GCML05] 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 (2005), 1461–74. 4

[GHA*08] Glatter M., Huang J., Ahern S., Daniel J., Lu A.: Visualizing temporal patterns in large multivariate data using textual pattern matching. *IEEE transactions on visualization and computer graphics 14*, 6 (2008), 1467–74. 4

[GPSL*11] Geng Z., Peng Z., S Laramee R., C Roberts J., Walker R.: Angular histograms: Frequency-based visualizations for large, high dimensional data. *IEEE Transactions on Visualization and Computer Graphics, 17*, 12 (2011), 2572–2580. 4

[HBW11] Heinrich J., Bachthaler S., Weiskopf D.: Progressive splatting of continuous scatterplots and parallel coordinates. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 653–662. 3

[HW09] Heinrich J., Weiskopf D.: Continuous parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 1531–1538. 3

[ID90] Inselberg A., Dimsdale B.: Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization* (1990), IEEE CS Press, pp. 361–378. 4

[JLC07] Johansson J., Ljung P., Cooper M.: Depth cues and density in temporal parallel coordinates. In *EuroVis* (2007), vol. 7, pp. 35–42. 4

[LMvW10] Li J., Martens J.-B., van Wijk J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2010), 13–30. 3

[LS09] Lee T.-Y., Shen H.-W.: Visualization and exploration of temporal trend relationships in multivariate time-varying data. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 1359–1366. 4

[MMDP10] Meyer M., Munzner T., DePace A., Pfister H.: Multeesum: A tool for comparative spatial and temporal gene expression data. *, IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 908–917. 4

[MS03] Muller W., Schumann H.: Visualization methods for time-dependent data - an overview. In *Proceedings of the Winter Simulation Conference,* (2003), vol. 1, pp. 737 – 745 Vol.1. 4

[NH06] Novotný M., Hauser H.: Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 12* (2006), 893–900. doi:10.1109/TVCG.2006.170. 3

[RJTTJ03] Rodrigues Jr J. F., Traina A. J., Traina Jr C.: Frequency plot and relevance plot to enhance visual data exploration. In *Brazilian Symposium on Computer Graphics and Image Processing* (2003), IEEE, pp. 117–124. 3

[Shn96] Shneiderman B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings Visual Languages* (1996), IEEE CS Press, pp. 336–343. 6

[SSJKF09] Steed C. a., Swan J. E., Jankun-Kelly T., Fitzpatrick P. J.: Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. *IEEE Symposium on VAST* (2009), 19–26. doi:10.1109/VAST.2009.5332586. 4

[WYM08] Wang C., Yu H., Ma K.: Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1547–1554. 4