Graphics and Visual Computing 3 (2020) 200013

Contents lists available at ScienceDirect

Graphics and Visual Computing

journal homepage: www.elsevier.com/locate/gvc

Technical Section

CrossVis: A visual analytics system for exploring heterogeneous multivariate data with applications to materials and climate sciences[☆]

Chad A. Steed[a,*], John R. Goodall[b], Junghoon Chae[a], Artem Trofimov[c,1]

[a] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[b] Cyber and Applied Data Analytics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[c] Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

ARTICLE INFO

Article history:
Received 20 November 2019
Revised 7 April 2020
Accepted 23 April 2020
Available online 12 May 2020

Keywords:
Information visualization
Multivariate data analysis
Statistical analytics
Materials science
Climate science
Explainable neural networks
Parallel coordinates

ABSTRACT

We present a new visual analytics system, called CrossVis, that allows flexible exploration of multivariate data with heterogeneous data types. After presenting the design requirements, which were derived from prior collaborations with domain experts, we introduce key features of CrossVis beginning with a tabular data model that coordinates multiple linked views and performance enhancements that enable scalable exploration of complex data. Next, we introduce extensions to the parallel coordinates plot, which include new axis representations for numerical, temporal, categorical, and image data, an embedded bivariate axis option, dynamic selections, focus+context axis scaling, and graphical indicators of key statistical values. We demonstrate the practical effectiveness of CrossVis through two scientific use cases; one focused on understanding neural network image classifications from a genetic engineering project and another involving general exploration of a large and complex data set of historical hurricane observations. We conclude with discussions regarding domain expert feedback, future enhancements to address limitations, and the interdisciplinary process used to design CrossVis.

1. Introduction

Forming a comprehensive understanding of patterns and relationships in multivariate data, where the phenomena under investigation are influenced by multiple factors, is integral to unlocking the full potential of today's vast data sets, especially in scientific domains. Whether interpreting the output of deep learning algorithms or exploring historical climate observations, scientists require interactive tools to develop a comprehensive understanding of large, multivariate data sets.

Developing new and improved multivariate visualization techniques for data exploration has captured the attention of many researchers as evidenced by a recent survey from Liu et al. [1]. However, real world analysis of such data remains a significant challenge for several reasons. One challenge is rooted in the technical difficulties of exploring increasingly large volumes of data. Another lies in equipping scientists with effective sense-making techniques (e.g., visual representations, interactive queries) for multivariate patterns. In practice, multivariate data sets often contain heterogeneous data types, missing values, and quality issues, which exacerbate the problem. Even with moderately sized data sets, heterogeneous data types represent a significant barrier to thorough data exploration. But when data sets are large and contain a mixture of data types (e.g., numeric, categorical, temporal, and image data) with problematic values, developing an adequate understanding of the data with the tools at hand becomes increasingly difficult, especially when coupled with the scientist's desire to freely navigate alternate paths during their analytical journey. To develop a comprehensive understanding, scientists need comprehensive solutions that address these challenges.

This article has been certified as Replicable by the *Graphics Replicability Stamp Initiative:* http://www.replicabilitystamp.org

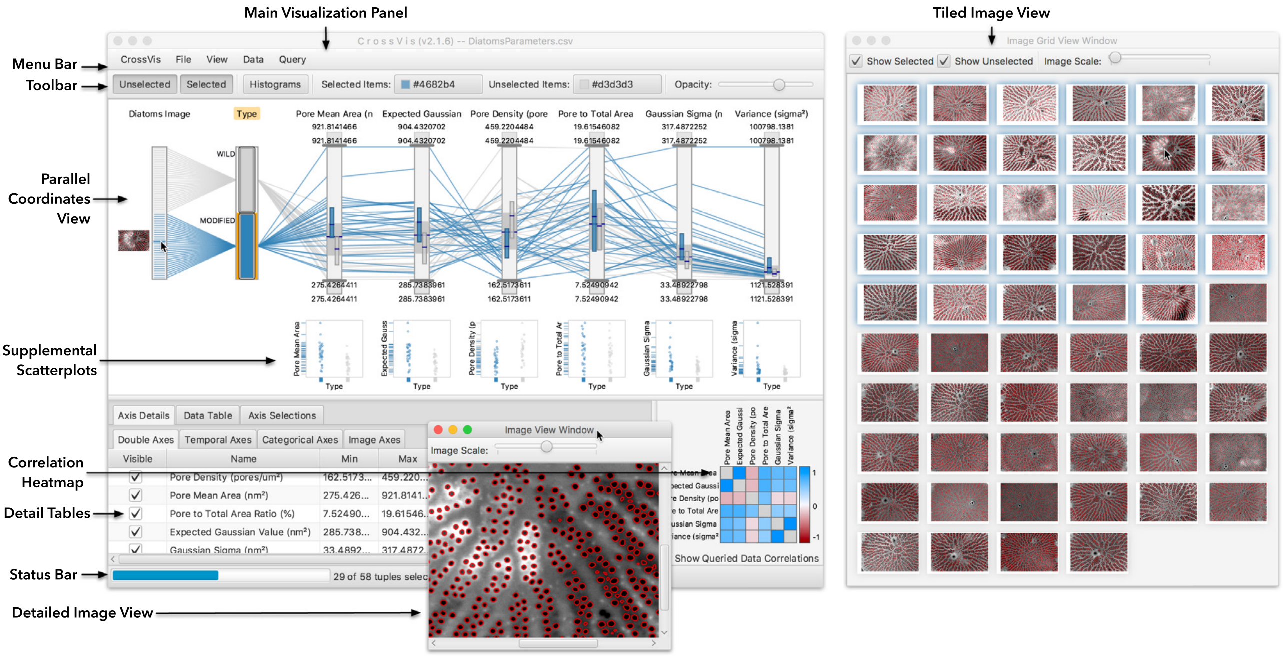
[☆] This article was recommended for publication by Prof. A. Gomes.

[*] Corresponding author.

E-mail addresses: csteed@acm.org (C.A. Steed), jgoodall@ornl.gov (J.R. Goodall), chaej@ornl.gov (J. Chae), trofimovaa@ornl.gov (A. Trofimov).

[1] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

https://doi.org/10.1016/j.gvc.2020.200013



Visible	Name	Min	Max
☑	Pore Density (pores/um²)	162.5173...	459.220...
☑	Pore Mean Area (nm²)	275.426...	921.8141...
☑	Pore to Total Area Ratio (%)	7.52490...	19.61546...
☑	Expected Gaussian Value (nm²)	285.738...	904.432...
☑	Gaussian Sigma (nm²)	33.4892...	317.4872...

29 of 58 tuples selec

Fig. 1. The CrossVis system presents several interactive views of multivariate data that are coordinated using a single tabular data model. The main view, an extended version of the parallel coordinates plot, is supplemented with scatterplots, correlation visualizations, and image views to foster creative exploratory data analysis. CrossVis's interactive query capabilities help uncover key patterns, which are revealed through the various visual representations.

Visual analytics offers a viable approach for coping with these issues [2,3]. By blending interactive visualizations with computational guidance, well-designed visual analytics systems harness human and computational strengths to improve the outcomes of data-driven studies. Yet, the number of visual analytics techniques that are readily available to non-visualization experts for practical use with large, heterogeneous, and multivariate data, which are ubiquitous in modern scientific studies, is low, especially when a combination of capabilities is desired.

In light of these and other practical challenges, we have developed the CrossVis visual analytics system (shown in Fig. 1) in collaboration with domain experts. CrossVis expands the parallel coordinates plot (PCP) [4] to support new axis representations for several non-numeric data types, embedded bivariate PCP axes, linked supplemental visualizations, focus+context axis scaling, and views that vary the level of detail. A progressive rendering algorithm and an optimized data model provide support for large data sets. These features are motivated by feedback from domain experts in multiple scientific domains. This paper presents the design and integration of these methods into a flexible system that enables comprehensive multivariate data exploration. The main contributions of the current work include the following:

- New visual representations of numerical, categorical, temporal, image, and bivariate PCP axes that, when combined with linked supplemental visualizations and interactive queries methods, reveal trends and patterns in heterogeneous, multivariate data
- Functional design considerations related to the visual representations provided by CrossVis including discussion of alternative approaches
- An overview of CrossVis describing the incorporation of several PCP extensions (visual representations and interaction techniques) into a comprehensive system that is greater than the sum of its parts
- Feedback from domain experts following their application of CrossVis to practical scientific data analysis scenarios

2. Related work

CrossVis employs multiple visualization methods, such as scatterplots, tiled image views, and correlation heatmaps, but the focal point is an extension of the classic PCP technique. Inselberg [4] initially popularized the PCP as a method for visualizing hyper-dimensional geometries and later Wegman [5] applied it to the analysis of multivariate data. The standard PCP method yields a compact two-dimensional representation of multidimensional data sets by mapping the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ to points on N parallel axes, which are joined using a polyline [6]. The PCP is attractive for exploratory data analysis because it transforms high-dimensional data sets into a two-dimensional plot without dimensionality reduction. Although the number of variables that can be shown is only geometrically restricted by the resolution of the display, axes that are located next to one another yield the most obvious insight. To analyze relationships between variables that are separated by one or more axes, interactions and representations of information derived from analytical algorithms are necessary.

As recent surveys of PCP methods [7,8] and a book on PCPs by Inselberg [6] demonstrate, the drive to improve and apply PCPs has attracted considerable attention. In addition to extensions of the technique, a significant portion of the prior work involves the application of PCPs to a wide variety of domains, such as climate science [9–11], cybersecurity [12], computer forensics [13], genetics [14], biomedical [15], healthcare [16], and environmental pollution [17].

CrossVis implements several common PCP interactions building on the direct interaction techniques described in Siirtola [18] and visual data mining methods reviewed by Inselberg [6]. These interactions enhance exploratory data analysis and include polyline selections, reorderable axes, and details on demand. CrossVis also extends the standard PCP axis with graphical indicators of various summary statistics following earlier designs by Hauser et al. [19]. The current work describes extensions to these interactions as well as linkages to new visual representations.

PCPs and scatterplots [20] are two of the most popular multivariate data visualization techniques. Due to complementary characteristics, some prior work has combined the two into a single layout [17,21] using a coordinated multiple view (CMV) strategy [22]. Yuan et al. [23] introduced the scattering points technique to embed scatterplot points between PCP axes. Both PCPs and scatterplots excel at showing correlation relationships between variables [24]. Some previous work directly augmented standard PCPs with graphical indicators that encode correlation metrics, such as the Pearson correlation coefficient, to guide users to potentially significant trends [18]. Zhou and Weiskopf [25] delved deeper into correlation analysis using PCPs by introducing an indexed point representation of multivariate correlations as opposed to most PCP systems that focus on the relationships between pairwise variable combinations. In addition to supplemental scatterplots and correlation indicators, CrossVis includes the ability to interactively embed a scatterplot between axes in the main PCP view to investigate pairwise correlations.

Although the vast majority of PCP methods focus on numerical data, the desire to represent other data types has inspired PCP extensions. Kosara et al. [26] introduced the ParallelSets technique to allow interactive representations of categorical data. Fernstad and Johansson [27] demonstrated that the ParallelSets method is superior to common quantitative encodings of categorical data in frequency related tasks. More recently, Vosough et al. [28] described the parallel hierarchies technique, which used parallel Icicle Plots to display hierarchical categorical data. CrossVis includes a variation of the ParallelSets approach for representing categorical data as well as new representations for temporal and image data. To the best of our knowledge, CrossVis represents the first PCP system with support for numerical, temporal, categorical, and image-based data in a single PCP framework.

When dealing with some moderate and most large scale data sets, PCPs are prone to polyline overplotting and occlusion issues [24,29]. Clustering [29,30] and binning methods [31] alleviate these issues by reducing the number of polylines that are rendered. Other approaches include alpha blending and displaying statistical representations (e.g., summary statistics, histograms), in lieu of or in combination with polylines, to represent the data at a higher level of detail [18,19,32,33]. Finally, both graphical processing units (GPUs) [34] and distributed computing infrastructure [35] have been harnessed to improve the rendering speed and scalability of PCPs. CrossVis uses a progressive rendering algorithm that leverages system GPUs for improved performance. In addition, CrossVis uses statistical representations of raw data to provide summarized levels of detail, thereby reducing the need to represent every individual polyline for large data sets. CrossVis also integrates a focus+context technique for PCPs, which allows users to zoom into ranges of interest on numerical and temporal axes with dense clusters of polylines while maintaining contextual awareness, similar to work by Novotný et al. [31] and more recently Richer et al. [36] where a focus+context approach is formalized with abstract PCPs.

3. Design requirements

For over a decade, we have collaborated closely with scientists from climate, materials science, manufacturing, and other fields that engage in multivariate data analysis. Through these engagements, we have observed two fundamental limitations. At the onset of these collaborations, scientists often state that *they fail to examine enough of their data*. This issue is partially due to large data volumes, but other factors come into play such as inadequate visualization support for heterogeneous data types and cumbersome query support. Scientists also state that although they are good at finding patterns they already know, *new discoveries are slow to occur*. This issue can be tied to an inability to interactively explore the full data set as well as the application of automated methods or workflows that focus on known relationships, which can lead to anchoring bias.

We postulate that more flexible human-centered interactions, scalable visualizations, and comparative techniques that are tailored to specific data types are viable solutions to these issues. In the remainder of this section, we consider these issues in greater detail and present the design requirements that have guided the development of CrossVis.

R1: The visualizations should show the distinguishing characteristics of heterogeneous data types. Visualizations of multivariate data often transform temporal and categorical data into numerical values because a wider range of numerical representations are available. However, this process often leads to misleading statistical summaries and informative characteristics of the native data types are discarded. Therefore, CrossVis includes visual representations that are tailored to key data types (e.g., temporal, categorical, images) to enable more comprehensive analysis.

R2: The system should support flexible comparative analysis of variables and subsets. The ability to compare different variables and/or subsets of values empowers scientists to freely ask questions and explore more of the data. This capability is challenging for multivariate data sets, especially heterogeneous data, due to the number of ways items can be compared. To meet these demands, scientists require the ability to quickly select data. Furthermore, visual representations must clearly highlight variations between selections. Often both direct and indirect selections are necessary; precise adjustment capabilities for selecting parameters through indirect controls can complement fast and approximate direct selections.

R3: The system should clearly link selections in separate views of the data through highlights in the visual representations. Separate visualizations should be linked through interactions so that changes in one view are propagated to all views. Each view should be designed to clearly communicate specific aspects of the data to increase the probability of finding new insights. A viable way to achieve this requirement is to provide coordinated multiple views where interactions are managed using a separate data model that shares selection state through an event-based listener interface.

R4: The system should maintain responsive interactions and visualizations. Exploratory data analysis techniques must maintain responsive interactions to avoid disrupting cognitive flow. By spawning threads to handle different aspects of interaction, summarization, and rendering, the system can orchestrate processing demands and prioritize tasks vital to interactive performance. Rendering threads are launched to render subsets of the updated data and progressively refine of the visualization. The result is an increase in perceived scalability with large data sets. In addition, computational optimization, preprocessing, and caching strategies help sustain interactive performance.

R5: The visualizations should support views at multiple scales. A complementary approach to progressive rendering is to summarize the raw data at different levels of detail to reduce the number of graphical elements that are needed to display the essence of the data. In addition to reducing rendering times, summarized views reduce occlusion and clutter. In CrossVis, variable detail displays are implemented using a hierarchy of statistical summaries that drill down to raw data views.

R6: The visualizations should display magnified views with context. Focus+context techniques that allow users to expand or zoom into specific regions of interest in the overall data space are helpful for selectively probing higher levels of detail on demand. Focus regions magnify data within a particular range and provide contex-

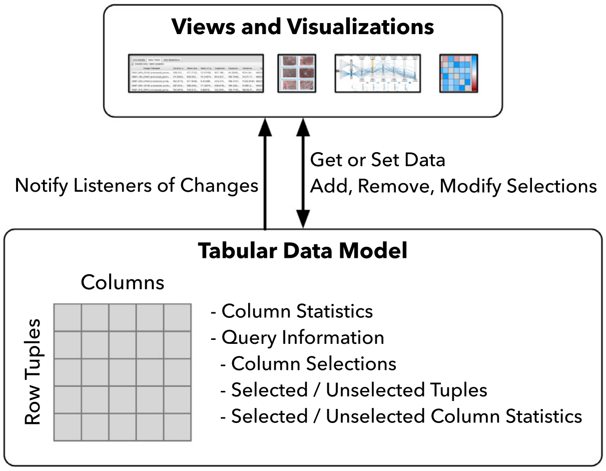


Fig. 2. A tabular data model supports multiple data types and provides access to both raw data and statistical summary information. The data model also coordinates user queries across the linked visualizations.

tual awareness by preserving the proximity of the focused region within the whole. These capabilities make dense clusters of shapes more legible.

4. Introducing the CrossVis system

As shown in Fig. 1, CrossVis is comprised of a main visualization panel that is supplemented by other linked views. In addition to detailed table views, the main panel is augmented by a correlation matrix, scatterplots, and image-based visualizations. The main visualization panel is an extension of the PCP featuring a unique combination of axis representations and embedded visualizations of statistical information.

CrossVis is an open source application[2] with performance enhancements that address the large scale analysis needs mentioned in **R4**. The JavaFX graphics library is used to render geometric shapes and the JavaFX user interface library is used for layouts, menus, and windows. These libraries automatically utilize system GPUs to boost rendering performance in a platform independent manner. In addition, the rendering algorithms use parallel threads to prioritize the display of more salient visual features and progressively reveal more details. In the remainder of this section, the data model and key data visualization techniques are described.

4.1. Tabular data model

CrossVis is supported by a custom-developed, tabular data model (see Fig. 2) that stores raw data, derived statistics, and selection criteria using collections of row and column objects. This data model is a critical component in CrossVis, and it is integral to fulfilling all of the previously mentioned requirements (**R1–R6**). Data structures are allocated in working memory as files are loaded, but these structures are not serialized to disk like a typical database system. The data model provides optimized statistical summaries, fast data access, and modular support for columns of numeric, categorical, temporal, and image data. Internally, row data are stored in native data types using a generic object array. A row-based data structure is implemented to match the access patterns of the PCP rendering algorithms, which display row tuples as polylines. The column objects store metadata and summary statistics, and they provide convenience methods for accessing data elements as native values. Caching mechanisms are also integrated for improved performance.

The data model manages subset selections using column selection criteria (e.g., value ranges, value sets) and an event-based listener interface to propagate changes. Data views register as listeners with the data model and implement a set of interface methods to respond to changes. Data views transmit user interactions to the data model, which notifies registered listeners. For quick access, the data model also maintains a query object that holds column selection criteria and references to the currently selected and unselected rows.

Column objects store summary statistics for the overall data distribution. The query object also stores column summary statistics for both the selected and the unselected rows. As selection criteria change in the visualizations, the data model detects the changes and updates the statistical summaries, which triggers the event-listener interface and forces other views to redraw. The summaries supply the visualizations with a fast level of detail hierarchy that enables multiple scale views. Standard descriptive statistics (e.g., mean, median, standard deviation, interquartile ranges) are calculated for numerical columns, and frequency-based statistics (e.g., histograms) are calculated for numerical, categorical, temporal, and image data.

The CrossVis data model performance is highly dependent on the host system's processor, graphics cards, and memory configurations. Most development occurred on a MacBook Pro with 16GB of random access memory, a 3.1GHz Intel Core i7 processor, and a 4GB AMD Radeon Pro 5600 GPU. With data files containing less than 10,000 rows, CrossVis maintains responsive interactions and rendering. As row counts are increased, responsiveness tends to suffer. However, drawing such a large number of lines in PCPs and points in scatterplots is often not useful due to overplotting side effects that make it difficult to see patterns. In these situations, CrossVis is designed to show histograms and / or summary statistics. These higher level views are capable of revealing patterns for larger segments of data and the user can select subsets using the interactive query techniques to draw polylines on demand. This scheme leverages a level of detail hierarchy to allow exploration of data sets containing upwards of 100,000 rows and a dozen or more columns. Further performance enhancements can be achieved, but these data scales represent a sweet spot for our targeted users.

4.2. New axis representations for parallel coordinates

The main visualization panel extends the PCP method to include new representations for specific data types and statistical information. PCP polylines correspond to row tuples in the data model and vertical axes to columns. In addition to subtle refinements (e.g., incremental line rendering, automated axis layouts), the CrossVis PCP design includes new axis representations supporting additional data types (addressing **R1**), supplemental displays of statistical information (addressing **R5**), and focus+context scaling (addressing **R6**). In the remainder of this section, these extensions are presented.

4.2.1. Numerical axis representation

Numerical PCP axes are augmented with graphical statistical summaries (see Fig. 3) and enhanced with the focus+context scaling technique described in Section 4.2.5. Both the typical value (mean or median) and dispersion range (two times the standard deviation range centered on the mean or the interquartile range) are represented in the axis bar interior. The user can toggle between mean- or median-based statistics using the application menu. These statistics are calculated for the overall distribution of values, the selected data, and the unselected data. For the

[2] CrossVis is available at https://github.com/ORNL/CrossVis.

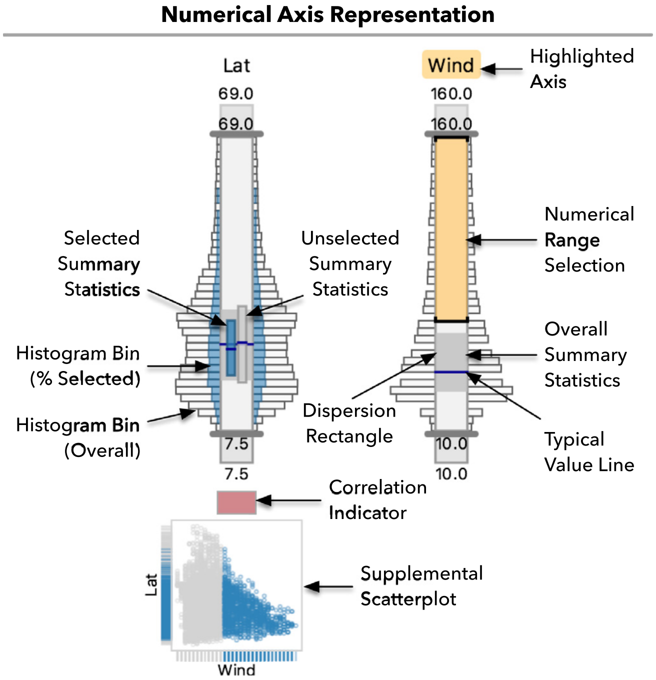


Fig. 3. CrossVis numerical axes are augmented with graphical statistical summaries using both descriptive and frequency statistics. Overall summaries, selected / unselected summaries, and scatterplots are visually represented providing an aggregated view to complement the detailed PCP polylines.

overall distribution (see the 'Wind' axis in Fig. 3), the height of the gray rectangle spanning the full axis bar width encodes the dispersion value and a blue line encodes the typical value. The overall statistical indicators serve as a baseline for comparisons against smaller subsets.

Narrow versions of the overall statistical indicators (see 'Lat' axis in Fig. 3) summarize the selected and unselected data. Statistics for the selected data are shown on the left and unselected on the right. The selected and unselected statistical indicators are visually linked to the selected and unselected polylines using color; the two dispersion rectangles are filled with the current selected or unselected polyline colors, which the user can modify using buttons on the toolbar above the PCP panel (see Fig. 1). These indicators are drawn over the overall indicators with a semi-transparent fill color to avoid completely masking the overall statistical information.

Numerical axes can also show frequency-based statistics as vertical histograms on the exterior of the axis bar (see Fig. 3). A standard histogram is computed for both the overall distribution and the selected data using bins covering equally sized value intervals. The number of bins is initially set to $k = \lceil \sqrt{n} \rceil$, where n is the number of rows in the data model, but the user can adjust the bin count through the application menu. A symmetrical layout is used with histogram bin rectangles shown on both sides of the axis. The rectangle width encodes the number of values falling within the bin's range. The percentage of values that are currently selected for each bin is encoded as the width of a semi-transparent rectangle drawn over the overall bin rectangle. The selected bin rectangle is filled with the current selected polyline color for consistency.

When both histograms and polylines are shown (see Fig. 9b), dense polyline clusters can negatively affect the decoding of bin counts. To deal with this situation, a thin white line is added to the bin rectangle's outer edge and the bin outline stroke color has high value contrast with the fill color. These color effects are intended to make the histogram silhouette more salient. Some clutter is introduced due to the inability of the semi-transparent histogram bins to completely mask polylines connecting to the axis behind them, but the polyline colors are muted and often the ability to see individual polyline axis intersections is useful.

Before settling on the symmetrical histogram design, we experimented with an alternative approach where bins were drawn on only one side of the axis. This unbalanced design made it difficult to see trends when an axis was positioned between two other axes. Furthermore, the unbalanced design was visually inconsistent with the other visual elements, which are mostly symmetrical. We also experimented with mapping bin counts to the fill colors of a band of smaller rectangles on the edge of the axis bar resembling a vertical heat map. Although using color required less space and avoided polyline occlusion, it was more difficult to compare relative bin counts, especially subtle differences. The superiority of positional to color encoding techniques is reported by Mackinlay [37].

Histograms can be shown instead of the polylines to improve performance with large data. Histograms provide more detail than the summary statistics, but less than individual lines. Thus, a series of increasingly detailed views is formed by showing statistical graphics, histograms, and then individual polylines providing a level of detail scheme that addresses **R5**.

4.2.2. Temporal axis representation

Temporal axes (see Fig. 4) are similar to numerical axes. Because the data model stores the values as time instants, labels, hover values, and range selections are shown using date-time formatted strings addressing **R1**. Temporal axis bars show a continuous value range and use the focus+context axis scaling technique (see Section 4.2.5). Instead of descriptive statistics, the axis interior displays a vertical temporal histogram where bin rectangles are centered horizontally. To provide more space for the histogram, temporal axis bars are wider than those of numerical axes. Locating the histogram inside the axis bar helps avoid the occlusion of polyline intersections. Similar to histograms on the numerical axes, the overall histogram bins are augmented with rectangles showing the percentage of selected values.

The symmetry of the temporal histogram visually unifies it with histograms on numerical axes. Unlike numerical histograms, temporal histograms are always displayed. In earlier designs, the temporal histograms were drawn outside the axis bar. This alternative design had a stronger correspondence to the numerical histograms, but it left an empty axis interior. We compensated for the empty space by making the axis bar more narrow, but this approach disrupted the overall consistency of axis bar treatments. We settled on the interior representation to capitalize on the opportunity to avoid occluding polyline intersections while sacrificing some visual unity. In the future, we plan to revisit the temporal axis design to explore methods that encode additional statistical information in the axis bar interior (e.g., dynamic time warping similarity metrics [38]).

4.2.3. Categorical axis representation

The categorical axis representation (see Fig. 4) emphasizes the relative frequency of categories. Categories are represented as rectangles inside the axis bar. In Fig. 4, the 'Status' axis has five categories. This figure also shows the two ways that category rectangles are displayed. On the left, the height of a category rectangle is mapped to the percentage of rows associated with its category. In this mode, the overall frequency trends are emphasized to support comparisons, but categories with a small percentage of values can be hard to see. The right 'Status' axis in Fig. 4 shows the equal height mode, which divides the axis bar height by the number of categories making smaller categories more visible. Users can also enable the display of category names as labels drawn to the left of the axis.

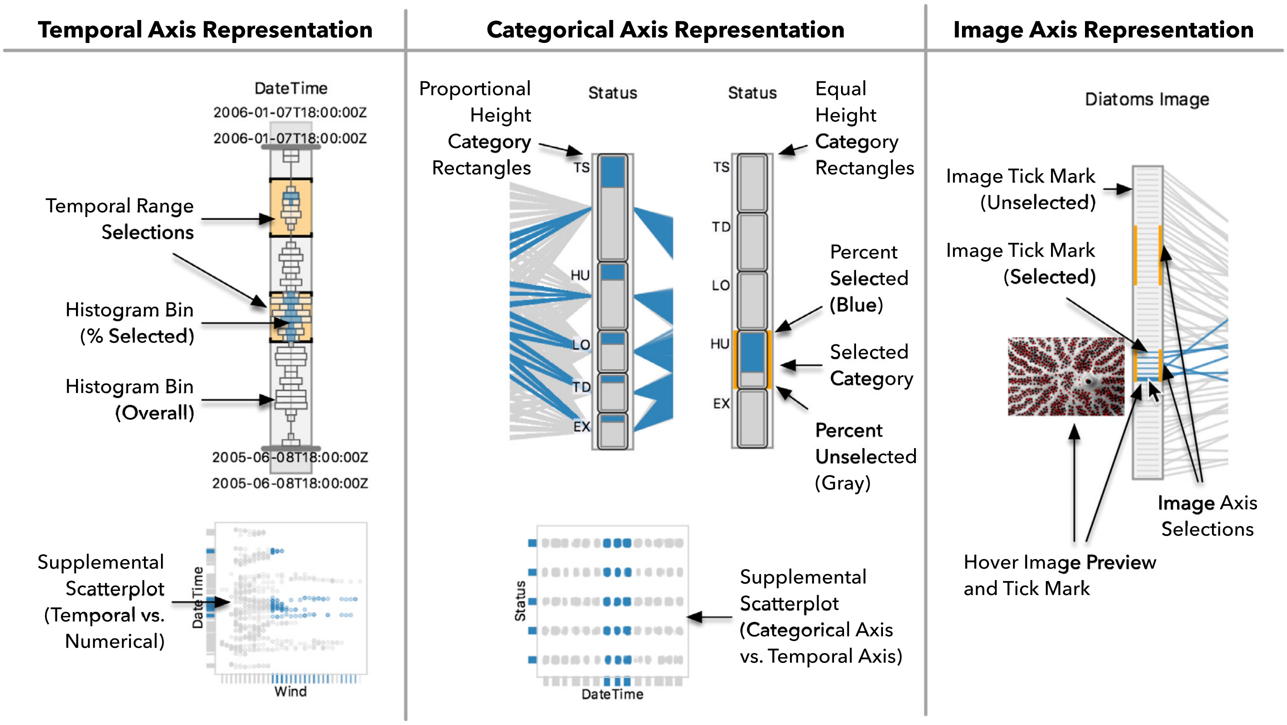

Temporal Axis Representation
Categorical Axis Representation
Image Axis Representation
DateTime
2006-01-07T18:00:00Z
2006-01-07T18:00:00Z
Temporal Range Selections
Histogram Bin (% Selected)
Histogram Bin (Overall)
2005-06-08T18:00:00Z
2005-06-08T18:00:00Z
Supplemental Scatterplot (Temporal vs. Numerical)
DateTime
Wind
Proportional Height Category Rectangles
Status
TS
HU
LO
TD
EX
Status
Equal Height Category Rectangles
TS
TD
LO
HU
EX
Percent Selected (Blue)
Selected Category
Percent Unselected (Gray)
Status
DateTime
Supplemental Scatterplot (Categorical Axis vs. Temporal Axis)
Diatoms Image
Image Tick Mark (Unselected)
Image Tick Mark (Selected)
Image Axis Selections
Hover Image Preview and Tick Mark

Fig. 4. CrossVis includes new axis representations for temporal, categorical, and image data. The unique characteristics of each data type inform the designs, which are augmented with statistical graphics and scatterplots. All the axes support interactive visual queries.

When polylines are selected, category rectangles are split into two smaller rectangles with heights that encode the ratio of selected (on the top and filled with the current selected polyline color) to unselected (on the bottom and filled with the current unselected polyline color) polylines. For example, the 'HU' category of the right 'Status' axis in Fig. 4 shows that about 75% of polylines associated with the 'HU' category are selected. To select polylines associated with a category, the user clicks on the category rectangle. On the right 'Status' axis in Fig. 4, the 'HU' category is selected. A category is removed from a selection by clicking the category rectangle a second time. When the user hovers over a rectangle, a tooltip reveals detailed information about the category (see Fig. 8).

Polylines are connected to the vertical centers of the overall category rectangles. In earlier designs, we evaluated representations that used polygonal shapes covering the full height of the category rectangles in a manner similar to the ParallelSets [26] design. However, during developer testing with data sets that had a large number of rows the polygonal shapes were difficult to read, especially between numerical (or temporal) and categorical columns. By adopting the polyline representation over the polygonal approach, we maintain visual consistency with representations between the other axis types and avoid clutter. However, we recognize that there is room for future improvements of the polyline representations on categorical axes.

4.2.4. Image axis representation

For columns consisting of images, CrossVis features a unique axis representation. As shown in Fig. 4, images are visually represented by horizontal tick marks inside the axis bar. The ordering of the image tick marks is determined by the file name. Tick marks are colored using either the selected or unselected polyline color. When the user hovers over a tick mark, its line's thickness increases and a small thumbnail copy of the corresponding image is shown to the left of the axis bar.

Images are selected by clicking on tick marks or dragging a selection range. To deselect images, the user presses the control key modifier while clicking or dragging. Selected images are indicated by orange highlights on the left and right of the axis (see Fig. 4). Because selections are combined using an OR operation between different axes, it is possible that images can be included in selections on the image axis, but those images are not associated with the current overall set of selected polylines (see the upper axis selection on the 'Diatoms Image' axis in Fig. 4). In such cases, the highlighting of selected images prevents selections on image axes from becoming invisible.

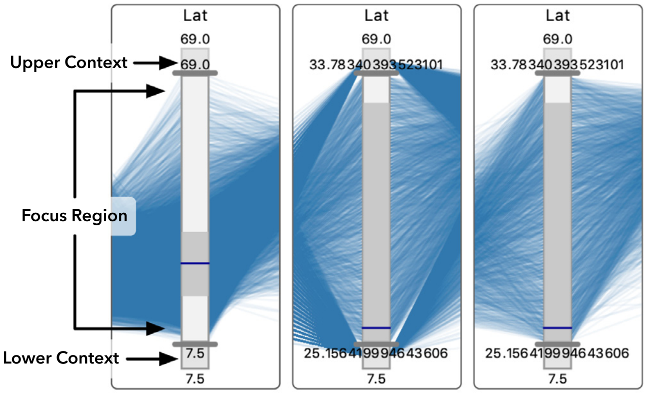
4.2.5. Focus+context axis scaling

Dense PCP polyline clusters make it difficult to decipher patterns due to overplotting. Adjusting the opacity of polylines helps, but the problem is not completely eliminated, especially with large data sets. To cope with this issue and address **R6**, CrossVis builds on the dynamic axis scaling technique introduced in MDX [11]. The CrossVis implementation provides more focus range control and adds support for temporal axes.

As shown in Fig. 5, upper and lower context regions are located above and below the main focus region. The focus region extents are adjusted by dragging the thick gray lines at the focus range edges. Moving the maximum value boundary line down decreases the extent value and upward movement increases it. After the extents are modified, the display is redrawn, which spreads polylines in the focus region out and pushes some polylines into the context. Especially when lines connect from an extreme edge of a neighbor axis to the opposite edge of another axis (e.g., top of one axis to the bottom of another), the context polylines can occlude the display after axis scaling. As shown in Fig. 5c, the user can choose to hide polylines in either of the context regions to further reduce clutter.

4.2.6. Bivariate axis representation

Bivariate PCP axis representations are supported because some variables are more easily understood in relationship to another variable. For example, geographic patterns are more apparent



(a)
Initial Focus Range Set to Overall Extents

(b)
After Focus Range Compression

(c)
After Hiding Context Polylines

Fig. 5. Focus+context axis scaling is provided for temporal and numerical axes. In (a), the focus range, marked by the thick gray lines between the focus region and the two outer context regions, is set to the overall data extent. In (b), the focus range lines are moved in magnifying data between the 50% and 75% percentile. In (c), polylines in the context region are hidden to reduce clutter.

when latitude and longitude values are shown in a scatterplot (see Fig. 9a). Univariate PCP axes make it hard to analyze such variables.

The bivariate axis is represented as a scatterplot using the same design as the supplemental scatterplots described in Section 4.3. However, the bivariate axis is embedded in the PCP with polylines of neighboring axes connecting to the y-axis of the scatterplot. By embedding the bivariate axis in the PCP, we alleviate potential perceptual issues associated with separated views, such as change blindness [39]. Any univariate axis can be combined with another to form a bivariate axis. The user can add bivariate axes manually by specifying the x and y columns from the data model, or the user can drag and release the x-axis on the target y-axis.

4.2.7. Additional interactive axis selection considerations

The ability to query and filter data is essential for efficient exploratory data analysis. The selection capabilities in CrossVis meet comparative analysis needs in **R2**. As shown in Fig. 3, users can select polylines that fall within value ranges on numerical and temporal axes. Multiple selections on multiple axes are supported (see Fig. 4) by directly dragging a range selection using the mouse over an axis bar. The user can select a category by clicking on its associated rectangle on categorical axes and an image by clicking on its associated tick mark. Users can set bivariate selections by dragging rectangles within the scatterplots. For precise control, the user can also manually add selections for all axis types using the Axis Selections tab (see bottom of Fig. 1).

Axis selections are visually unified using an orange highlight color. For range selections, the fill color of the selection rectangle uses the highlight color. For categorical and image selections, the rectangle or tick mark is augmented with an orange halo. The user can directly interact with the selection indicators to remove items or adjust the extents.

To compute the subset of selected polylines when multiple selections are present, a disjunction (OR) operation is first applied to selections on individual axes and then a conjunction (AND) operation is applied to the selected values between axes. This logic allows users to consider values spread out over non-contiguous value ranges for individual axes as well as relationships between different axes.

4.3. Supplemental scatterplots

In addition to the ability to embed scatterplots directly in the PCP, small scatterplots are shown below the PCP axes (addressing **R3**). When an axis is highlighted (see Fig. 8), the supplemental scatterplots show the pairwise relationship of the highlighted axis with all other axes. That is, the x-axis of each scatterplot is mapped to the highlighted axis and the y-axis is mapped to the axis above the scatterplot. When no axis is highlighted (see Fig. 9b), the scatterplots are shown in the space between the axes where the x-axis is the left axis and the y-axis is the right axis. Both the supplemental scatterplots and the embedded bivariate axis scatterplots can be configured to show tick marks on the axis boundary and convey univariate distributions.

Scatterplots excel at conveying non-linear trends and clusters. Furthermore, scientists are usually familiar with scatterplots and displaying these in conjunction with the PCP, which is often new to them, can increase understanding. Thus, the combination of these two techniques is more valuable than showing either in isolation.

4.4. Axis correlation coefficient representations

Supporting **R2** and **R3** requirements, correlations between numerical axes are shown in several ways. The user may glean correlations from polyline configurations in the PCP (e.g., 'X' shaped crossings indicate negative correlations and more horizontal crossings indicate positive) and point configurations in the scatterplots. In addition, direct encodings of the Pearson correlation coefficient, r, are shown above the supplemental scatterplot as color-filled rectangles (see Fig. 4). The r values are mapped to a color scale where the most saturated blue represents a perfect positive correlation, the most saturated red represents a perfect negative correlation, and white represents no correlation. The user can hover the mouse over a cell to see the exact r value. As shown in Fig. 1, CrossVis also shows a linked correlation matrix using the same color coding scheme as the indicators in the parallel coordinates plot. To the right of the matrix heatmap, the r value color scale is shown.

4.5. Image Set and Individual Image Views

CrossVis provides two separate image views (addressing **R3**): a tiled image set view and a detailed image view (see labels in Fig. 1). A slider at the top of both views allows adjustments of the image(s) size. The image set view is linked to the other visualizations in the main window. When polylines are selected, the images in the selection set are haloed with the selected polyline color. Likewise, the unselected images are haloed with the unselected color. Furthermore, selected images are located at the top of the image set view. In Fig. 1, the cutoff between selected and unselected images is visible after the fifth image of the fifth row.

When the user double-clicks on an image in the tiled view, a detailed view appears (see Fig. 1 at bottom) making it easier to focus on a specific image. In Fig. 6, the detailed image view shows a magnified region of the image where some of the dark circular shapes (see yellow box) of the microscopic view are not enclosed by red outlines that were added by a pore detection algorithm.

5. Two practical scientific use cases

In this section, we present two scientific use cases of CrossVis to demonstrate its data exploration capabilities. The first focuses on understanding the results of an artificial neural network (ANN) designed to classify microscopic imagery; the motivating scenario that inspired the development of CrossVis. The second describes exploration of a historical hurricane observation data set, and it

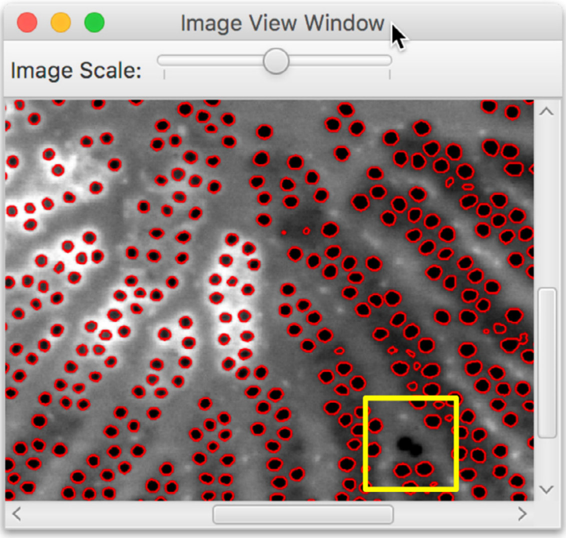


Fig. 6. The detailed image view allows scientists to visually examine images associated with polylines in the PCP. Here the view shows a diatom image where a pore detection algorithm failed to label two pores, outlined by the yellow box annotation, possibly because the boundary between the pores is blurred (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

illustrates CrossVis's capacity to discover, investigate, and validate patterns in a larger and more complex data set as well as its suitability as a general purpose exploration system.

5.1. Use Case 1: Understanding neural network image classifications in genetic engineering

Scientists at Oak Ridge National Laboratory's Center for Nanophase Materials Science (CNMS), one of whom is a co-author of this paper, used an ANN to automatically classify scanning electron microscope (SEM) images of diatoms, some of which were genetically modified. The ANN predicts whether images correspond to genetically modified ('MODIFIED') diatoms or not ('WILD'). A diatom is a unicell alga with a silica cell wall. Diatoms are attractive candidates for functional systems of materials with applications ranging from photonics, sensing, filtration, and drug delivery [40,41]. The scientists used CrossVis to analyze the ANN results and the following narrative captures some of their findings.

In addition to the ANN classification of 'WILD' or 'MODIFIED', the scientists computed a number of parameters that quantify characteristics of pores detected in the images: density of pores, mean area of pores, and the percentage of area occupied by pores relative to the total area of the valve captured in an image. Additionally, pore area distribution was extracted and fitted with a Gaussian distribution to yield two more parameters: Gaussian value and Gaussian sigma. These parameters, in combination with a variance metric, yield a total of six quantitative values that supplement the categorical value output from the neural network classification. The goal of this study was to understand the significance of these parameters for distinguishing between modified and unmodified diatom images.

We begin by selecting the 'MODIFIED' category on the 'Type' axis (see orange highlight on 'MODIFIED' category in Fig. 1). This action associates the selected polylines and summary statistics (shown in median/IQR mode) with images marked as 'MODIFIED' and the unselected with 'WILD' images to allow comparative analysis. In the tiled image view, the 'MODIFIED' classified images at top are haloed with the selection color and the 'WILD' images are below. This view reveals that 'WILD' images appear more uniform and have more pores (pores are outlined in red by a separate process) as compared to the 'MODIFIED' images, which exhibit more pixel variance and wider gaps between pores.

The variance is particularly evident in the hover image on the 'Diatoms Image' PCP axis in Fig. 1. The hover image is located at the intersection of row 2 and column 5 in the tiled image view. Fig. 6 shows a magnified view of the detailed image window at the bottom of Fig. 1. The image shows that two pores with fuzzy edges at the lower right corner (see yellow highlight box) were missed by the pore detection step. Several other missed pores are apparent in top detail image view on the right in Fig. 7. Visual inspections using CrossVis give scientists the ability to drill-down and find such subtle patterns, which in this case provides an opportunity to improve the pore detection process. Furthermore, CrossVis's ability to display the images at multiple scales helps scientists see that the two image sets are visually distinct.

We shift our attention to the PCP to explore quantitative trends. In Fig. 1, the IQR rectangles for the 'Pore Mean Area', 'Expected Gaussian Value', and 'Pore Density' axes show the most separation between the two image categories, which suggests that these are distinguishing features. The 'Pore Mean Area' and 'Expected Gaussian' axes exhibit a strong positive correction ($r = 0.97$), as indicated by a saturated blue square on the heatmap and nearly horizontal polylines between the two PCP axes. This finding suggests that one of the two variables can be removed since together they fail to add additional value. Having more overlap between the selected and unselected IQR rectangles, the 'Expected Gaussian' axis is a good candidate for removal. Both 'Pore Mean Area' and 'Pore Density' show clear separation between the two image categories, with 'Pore Mean Area' showing less overlap and fewer polylines crossing the median line.

In Fig. 8, the 'MODIFIED' category selection on the 'Type' axis is removed and a range selection on 'Pore Mean Area' for values greater than the median is added. The selection captures 29 of the 59 images, most of which are of the 'MODIFIED' category. However, some 'MODIFIED' images are excluded and some 'WILD' images are included. The tooltip shows 7 of the 29 'WILD' images are selected. The tooltip for the 'MODIFIED' category (not shown) shows that 22 of the 29 images are selected. This finding confirms the significance of pore sizes and density that we observed from the earlier visual inspections and it reinforces the importance of a multivariate process to accurately classify the images. As the materials scientist who led this analysis stated: "*Such information suggests that not every single diatom in the 'MODIFIED' set underwent genetic modification, which was clearly revealed using CrossVis analysis.*"

A glance at the 'Variance' axis in Fig. 8 reveals two severe outliers in the upper range. These two polylines are selected in Fig. 7 with the two associated images shown on the right. The 'MODIFIED' image discussed previously is shown and the wide range of pixel fluctuations in both images confirms the pixel variance.

5.2. Use Case 2: analyzing historical tropical cyclone observations

The National Oceanic and Atmospheric Administration (NOAA) maintains the Atlantic Hurricane Database (HURDAT2),[3] which contains information on the location, winds, central pressure, and size (since 2004) of all known tropical and subtropical cyclones in the Atlantic basin between the years 1851 and 2017 [42]. HURDAT2 is important for understanding historical tropical cyclone trends, but the number of records (over 50,303 rows), number of variables (21 columns), and heterogeneous data types (categorical, temporal, and numerical) make it a challenge to analyze.

[3] HURDAT2 is available at https://www.nhc.noaa.gov/data/.

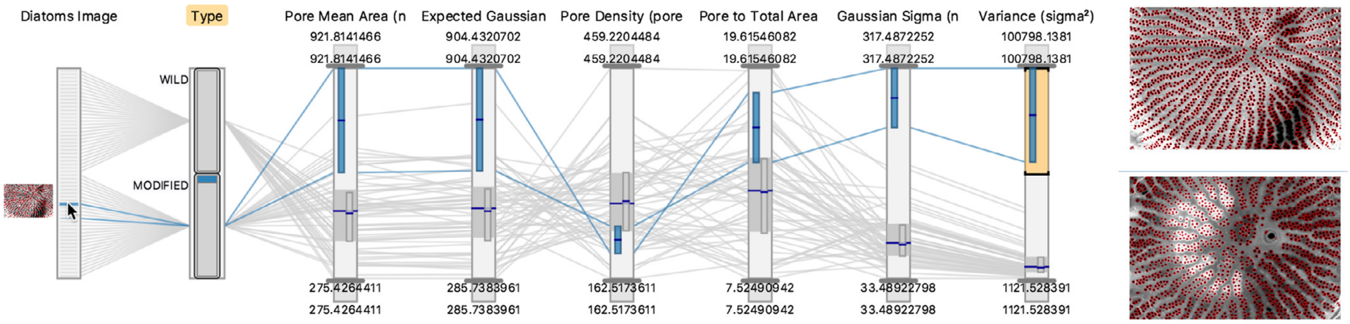


Fig. 7. Two outlier images are selected on the 'Variance' axis. The pixel variance of the images, both classified as 'MODIFIED', is evident in the thumbnails on the right. These images are outliers on all but the 'Pore Density' axis.

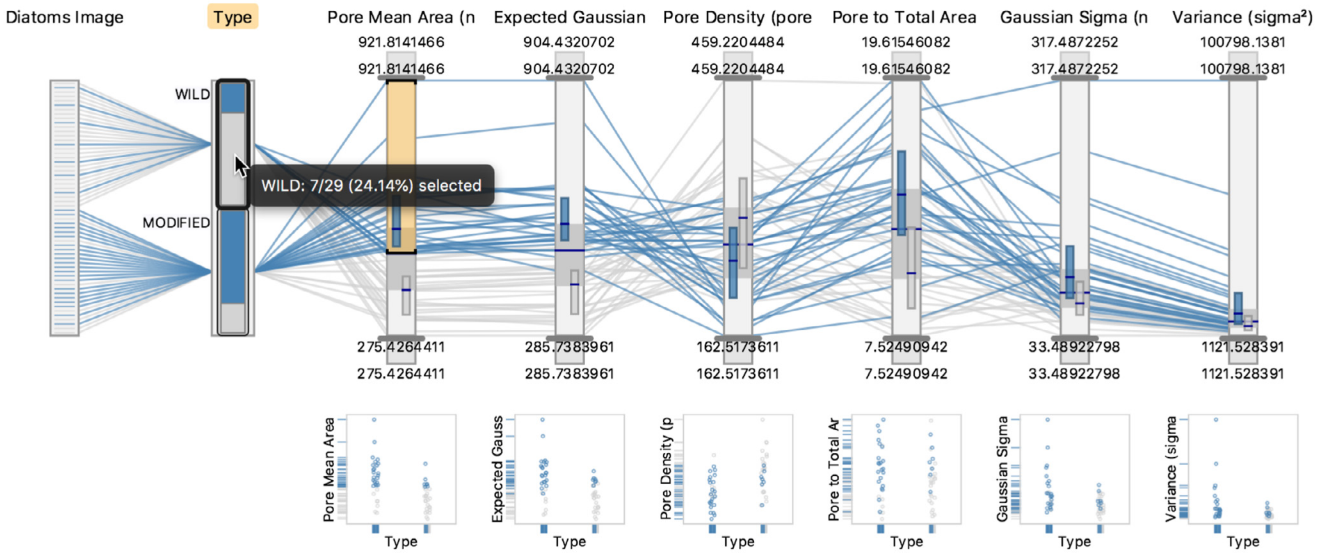


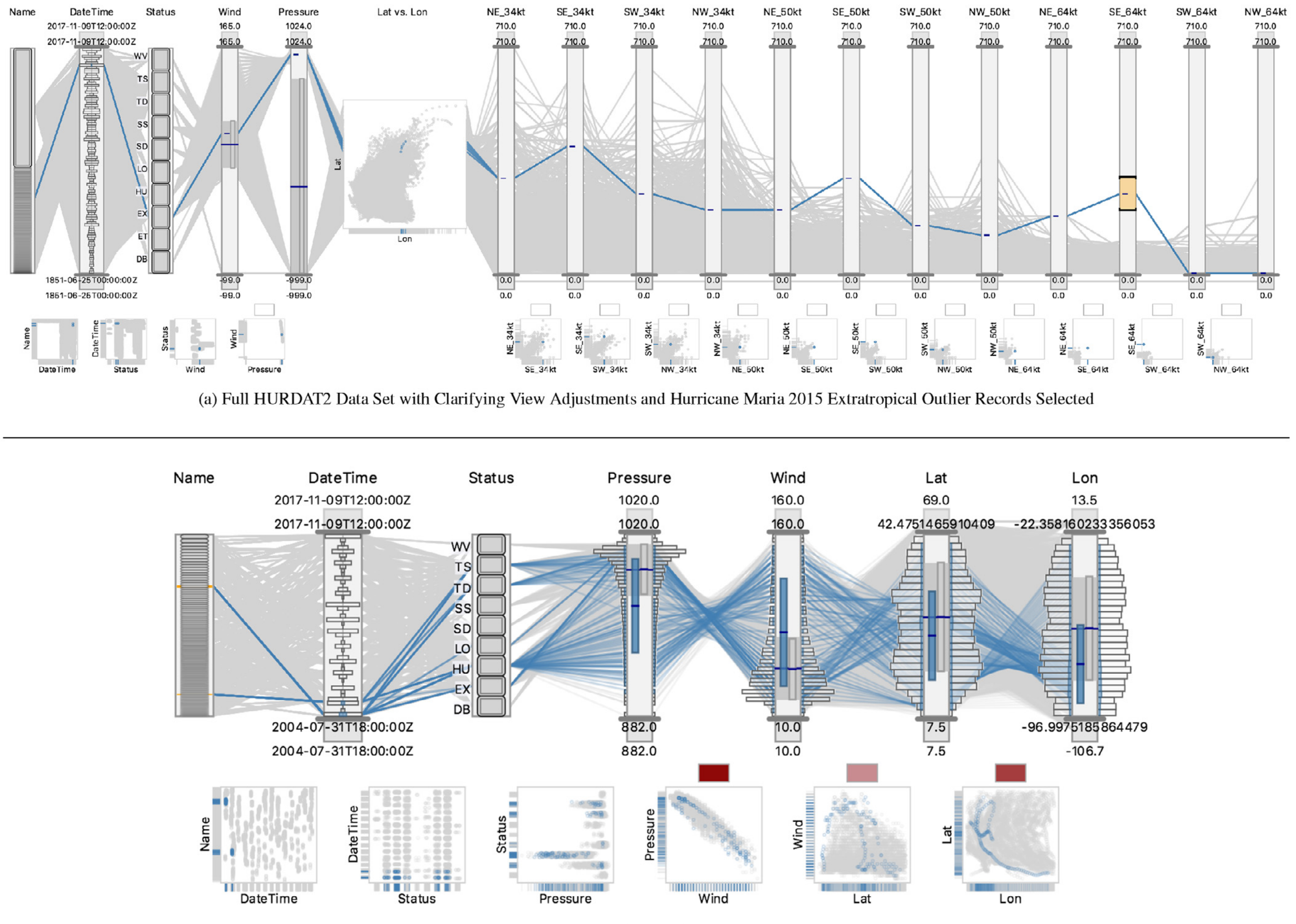
Fig. 8. A numerical range selection is used to compare images with high and low 'Pore Mean Area' values. The separation of the IQR rectangles on the 'Pore Mean Area' axis suggests the importance of these axes for distinguishing between the two 'Type' categories ('WILD' versus 'MODIFIED'). The 'Type' axis is highlighted, resulting in the display of scatterplots below the other axes.

In Fig. 9a, the full HURDAT2 data set is shown. Despite the size and number variables, CrossVis maintains interactive performance ($< 1sec$) during visual investigations. In the figure, scales for the 12 wind radii axes (located on the right side of the figure) are synchronized to a common range and several values with 'no data' wind radii values (wind radii fields were omitted for storms prior to 2004) are pushed into the context regions to achieve clearer views. The wind radii values provide wind swath size information for the 34 knot (34kt), 50 knot (50kt), and 64 knot (64kt) maximum wind ranges. For each range, four radii values are provided in nautical miles (nm) for the four quadrants: northeast (NE), southeast (SE), southwest (SW), and northwest (NW). The view reveals that wind swaths grow tighter (less dispersed) for fields with higher wind speed since the distance values decrease and associated wind speeds increase from left to right.

The resulting view highlights five records (selected in Fig. 9a) with remarkably large 'SE_64kt' values (see lines captured by the range selection on this axis). Using the numeric data table view (not shown), we find that these records are from Hurricane Maria in 2005. With these records selected, it is evident that the southeast quadrant is larger than normal on the other wind region axes ('SE_34kt' and 'SE_50kt'). The selection also shows that the western side of the storm swath has collapsed in the most intense wind region (0 nm for the 'SW_64kt' and 'NW_64kt' axes).

On the 'Status' axis, we see that the records for Hurricane Maria 2005 are assigned the 'EX' category, which denotes an extratropical system in the middle latitudes (between 30° and 60° latitude) of Earth. The latitudinal location is visible in the 'Lat vs. Lon' axis. These observations are consistent with the tendency of extratropical storms to become less symmetric in the middle latitudes [43] and the straightforward way these complex relationships are explored with CrossVis demonstrates its effectiveness in data verification and validation tasks.

Fig. 9b focuses on storm records between 2004 and 2017, which offer the most detail. We select the 'IVAN' and 'KATRINA' categories on the 'Name' axis highlighting records for Hurricane Ivan (2004) and Hurricane Katrina (2005), which were both major hurricanes that caused tremendous destruction. A high correlation between 'Pressure' and 'Wind' ($r = -0.94$) is indicated by the highly saturated red correlation indicator, scatterplot point configuration, and PCP polyline crossings. The statistical indicators also show that pressure values for these storms were lower than normal and wind values were higher. Both storms progressed through different stages of development, which accounts for the variation of values on the 'Status' axis. After adjusting the focus range of the 'Lat' and 'Lon' axes to drill into on the affected geographic range, we observe that on average these storms were more southern and western in latitude and longitude, respectively. To put it another way,



(a) Full HURDAT2 Data Set with Clarifying View Adjustments and Hurricane Maria 2015 Extratropical Outlier Records Selected

(b) Records between 2004 and 2017 with Categorical Selections of Hurricanes Ivan and Katrina

Fig. 9. CrossVis provides interactive techniques to cope with large multivariate data and the quality issues, such as missing data, that are common in scientific applications. In (a), the full NOAA HURDAT2 tropical cyclone data set is shown after we push 'no data' flagged values into the lower context region (see lines below 0.0 on the wind radii axes) and synchronize the 12 wind radii axis scales. Five outlier values on the 'SE_64kt' axis are selected originating from extratropical records of Hurricane Maria 2015. In (b), a view of records between 2004 and 2017 is shown with a selection on the 'Name' axis for Hurricanes Ivan and Katrina.

these storms lingered longer in the Caribbean and Gulf of Mexico regions, as can be seen in the supplemental scatterplot below the 'Lat' and 'Lon' axes.

6. Discussion

CrossVis increases scientists' capacity to develop a more comprehensive understanding of multivariate data, particularly when heterogeneous data types are present, by providing flexible techniques for quickly querying data and visually representing results. The system offers a unique collection of analysis capabilities that correspond to specific challenges scientists face. These challenges were uncovered through prior collaborations and the process of developing techniques to address them was executed in close collaborations with the scientists who now use the system. As we developed CrossVis and evaluated iterations with domain experts, we observed indicators of its effectiveness and limitations. In this section, we discuss these observations and reflect on the interdisciplinary design process.

6.1. Domain expert feedback and observations

The effectiveness of CrossVis is perhaps most apparent in a more comprehensive discussion of the results of the ANN diatom image classification project that our materials science collaborators recently published [44]. That work focused on the ANN design and scientific implications to genetic engineering as opposed to a detailed description of CrossVis, which the current work provides. Some findings related to explaining the diatom images classifications were revealed with an earlier version of CrossVis and communicated in the publication with figures from the main PCP visualization panel. The scientists stated that CrossVis significantly increased their understanding of the complex ANN process and, as the quote in Section 5.1 states, "*clearly revealed*" complex relationships in the data. Prior to using CrossVis, scientists were forced to flip between static plots (e.g., scatterplots, histograms) and file system image viewers to compare features. Querying and filtering the data involved running scripts to regenerate static plots, which slowed investigations and limited the number of different combinations of conditions that could be realistically viewed. Dimensionality reduction processes were also employed, but results were difficult to translate back to the original parameters.

Scientists also noted that CrossVis helped them think in a more multivariate manner during analysis. For example, it became clear that multiple pore measures were instrumental in classifying the images. Although some variables were more correlated than others, it was clear that no single variable could be tied to the results of the ANN process. The expanded PCP visualization efficiently conveyed this condition through the interactive representations of images and categorical values with quantitative metrics. One domain expert mentioned that CrossVis "*enabled faster pairwise comparative variable comparisons because we don't have to pick through and cycle* *between a set of scatterplots*." Referencing the diatom image classification example in the previous section, this same expert noted that the correlation between 'Pore Mean Area' and 'Pore Density' was not apparent until the data was viewed in CrossVis. Moreover, the PCP revealed that the 'Pore to Total Area' variable does not change between 'WILD' and 'MODIFIED' sets, which led to additional investigations. In the words of one expert: "*Only by using CrossVis were we able to find that an increase in mean area of pores caused a decrease in density of pores for the 'MODIFIED' set, and vise versa for the 'WILD' set, causing both sets to maintain approximately the same ratio of pore area to total area*."

The scientists also found the ability to view categorical and numerical data in a single system, especially with the image representations, particularly helpful in distinguishing between the two different image categories. The ability to select images of one category and see the separation in values on the other numerical axes was key to their identification of the most sensitive parameters for the ANN classification process.

The image visualization capabilities in CrossVis were added after the previously mentioned publication on the ANN diatom image classification project. By integrating an image-based PCP axis and the linked image view panel, scientists noted increased productivity because they didn't have to rely on a separate image viewer and manually link lines in the PCP to individual images. They felt that the image views helped to supplement the mostly quantitative analysis, especially for investigating outliers and visually exploring specific image features (e.g.pores, skeleton structures).

By viewing their data in new visual representations, scientists were encouraged to consider their data from fresh perspectives, which led to more creative analytical discourse. After an initial training session to help them understand the views, they were free to explore the data using direct interaction techniques. Because it was no longer a requirement to manually run scripts to query the data, CrossVis increased the number of combinations they could consider and expanded their depth of understanding of the data. Although the total time they invested in data analysis may have been about the same as before, the direct query capabilities allowed them to consider more of the data and find unforeseen patterns; two main issues we have observed in most of our collaborations with domain scientists.

6.2. Limitations and future enhancements

Although CrossVis works well in several scenarios, limitations and opportunities for improvement remain. In the remainder of this section, we describe the most salient observations.

6.2.1. Providing more active computational guidance

CrossVis relies on the user to form hypotheses and drive most of the analysis. Statistical analytics provide hints at potentially significant trends, but these are predominately passive requiring the user to see a visual indicator and follow the lead to deeper investigation. In the current application, the statistical analytics primarily support the level of detail rendering scheme.

To improve CrossVis, we envision the integration of automated machine learning techniques that suggest key patterns and more actively guide the user during the analysis process. We have already explored the integration of multiple linear regression in our prior work with the MDX system [11]. We are planning new methods to couple CrossVis visualizations with other approaches, such as dynamic time warping to highlight similar time series trends and other anomaly detection routines. These techniques could be designed to capture user interactions, either implicitly or explicitly, and feed the examples as labels to the automated algorithms for highlighting similar patterns in the full data volume, especially in unseen sections of the data.

6.2.2. Increasing scalability

CrossVis is a standalone application designed to run on laptops and workstations, which works well for typical data sets we have encountered but doesn't scale smoothly to data sets that exceed 10s of gigabytes. To support larger data sets, like those generated from computer simulations in climate science, we are investigating a distributed approach where the full data set resides on remote high performance systems. Analytical processing and query operations will be executed on the remote system and only the results will be transmitted to the client, where the interactive data visualization components operate. Such a system could enable web-based visualizations of larger and more complex data sets, thereby improving accessibility and maintenance.

6.2.3. Refining the PCP axis representations

We evaluated several variations of the overall CrossVis design before settling on the current version. One objective was to pack as much useful information as possible into the axis representations without overloading the user. For example, the symmetrical presentation of histograms on the numerical axes involved evaluations of alternative solutions using various visual feature mappings. We recognize that there is additional room for exploring enhanced axis representations. Some of these (e.g., reclaiming the axis bar interior for time series plots or other statistical metrics, representing polyline connections for categories) are mentioned in the previous sections. Other future work involves supporting vertical PCP axis orientations, as opposed to strictly horizontal, which may be easier for deciphering histogram and summary views. We are also interested in new ways to represent multiple focus regions using the focus+context axis scaling technique. This expansion could enable exploration of multiple focus ranges on the same axis as well as cascading axes that drill down into specific ranges.

We are actively expanding the concept of embedding additional multivariate views into PCPs in a manner similar to the bivariate axis scatterplots. Alternatives include leveraging additional visual features (e.g., color, size, shape) to encode additional variables, providing three-dimensional views of volumes with plane slicing capabilities, and embedding additional visualization techniques such as miniature treemaps or graphs. Using a modular axis design, a wide range of possibilities exist.

6.3. Reflections on the interdisciplinary design process

CrossVis was developed in close collaboration with materials scientists following a participatory design process where our collaborators were co-designers and primary users of the system. We met with them regularly to discuss new developments, mock-ups, and evaluated use of the tool with their data sets. The design was also influenced by prior collaborations with scientists in climate, cybersecurity, health care, and other domains. Such projects benefit from clear objectives related to domain specific challenges, which ground feature developments and help fulfill the central promise of data visualization; bringing the latest data visualization advances to data rich domains where they are needed.

By integrating experts from data visualization and other domains, interdisciplinary projects also benefit from a diversity of ideas. Domain experts teach data visualization experts about their analysis procedures, which often stimulates new interactive visualization designs. On the other hand, data visualization experts enlighten domain experts about new data visualization techniques and trends, thereby bridging the gap between theoretical data visualization and practical applications. Our experience is that both sides welcome the engagements and the results are almost always positive.

Data visualization experts often strive to generalize their techniques to maximize impacts on broader endeavors. At the same time, it can be hard to avoid degrading performance in the motivating domain specific scenario. For example, CrossVis supports reading data from CSV files while also supporting custom file formats for our domain collaborations to enhance performance with their data sets. If generalization impacts domain specific performance, the payoff must be carefully weighed. Perhaps by sacrificing some performance, domain experts will recognize the opportunity to use the technique with other data sets and the sacrifice will be welcomed. In such cases were performance degradation is unacceptable, the team may consider deploying specific builds of the tools; one for general purpose use and others for specific applications.

Interdisciplinary projects are challenging for data visualization experts because they must exhibit agility in learning new domains and concepts. By including the experts in the development team to validate assumptions and algorithmic decisions, the knowledge gap can be spanned and in the process data visualization researchers gradually gain a more comprehensive understanding of the domain. Furthermore, this process engages domain experts and can increase adoption rates as they share the techniques with others in their field. Perhaps the greatest reward from such an activity is when the experts publish results found with the new techniques, an outcome that essentially validates the effectiveness of the approach and one that we experienced the development of CrossVis with our materials science collaborators.

7. Conclusion

CrossVis extends the PCP concept by adding a range of new axis representation techniques, interactions, and scalability extensions to enable large scale, multivariate exploration of heterogeneous data. The overall design requirements were derived from key challenges uncovered during close interactions with experts in a variety of fields using prior multivariate visualization tools. The resulting system helps scientists look at more of their data and find new, and often unexpected, insights; two needs that are common in most scientific domains.

By working close with materials scientists to develop CrossVis, we observed how it improved the depth of their analysis as well as certain limitations and opportunities for future improvement. Important patterns were uncovered by domain experts who used CrossVis to interpret a neural network process for classifying microscopic images in a genetic engineering study. In the process of developing and applying the system, we gained additional insight into interdisciplinary collaborations to develop data science systems, particularly visual analysis systems. Another scientific use case described in the current work involves exploring a complex hurricane observation data set and demonstrates the suitability of CrossVis in general purpose data exploration.

CrossVis is a general purpose visual analytics framework that has also proven useful in other fields, such as cybersecurity, earth system modeling, health care, and algorithmic performance analysis. In the future, we will continue to expand and apply CrossVis to other data rich domains while continuing our interdisciplinary development strategy.

Declaration of Competing Interest

The authors declare that they have no known competing financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Chad A. Steed: Conceptualization, Methodology, Software, Project administration, Methodology, Visualization, Funding acquisition, Writing - review & editing. **John R. Goodall:** Conceptualization, Methodology, Writing - review & editing. **Junghoon Chae:** Conceptualization, Methodology, Visualization, Writing - review & editing. **Artem Trofimov:** Conceptualization, Formal analysis, Investigation, Data curation, Writing - review & editing.

Acknowledgments

The authors wish to express our appreciation to the reviewers and editor for their excellent feedback. The authors also thank Dr. Alison Pawlicki of Oak Ridge National Laboratory's (ORNL) Center for Nanophase Materials Sciences (CNMS) for acquiring the raw Scanning Electron Microscope (SEM) diatom imagery used in the CrossVis evaluation. The HURDAT2 data set is a product of the NOAA National Hurricane Center.

This research was sponsored by the U.S. Department of Energy (DOE) under the Scientific Discovery through Advanced Computing (SciDAC) RAPIDS project and the Laboratory Directed Research and Development Program of ORNL, managed by UT- Battelle, LLC, for the U.S. DOE. The SEM analysis was partially conducted at ORNL's CNMS, which is a DOE Office of Science User Facility.

References

[1] Liu S, Maljovec D, Wang B, Bremer P, Pascucci V. Visualizing high-dimensional data: Advances in the past decade. IEEE Trans Visual Comput Graph 2017;23(3):1249–68. doi:10.1109/TVCG.2016.2640960.

[2] Thomas JJ, Cook KA. A visual analytics agenda. IEEE Comput Graph Appl 2006;26(1):10–13. doi:10.1109/MCG.2006.5.

[3] Keim D, Kohlhammer J, Ellis G, Mansmann F, editors. Mastering the Information Age: Solving Problems with Visual Analytics. Goslar, Germany: Eurographics Association; 2010.

[4] Inselberg A. The plane with parallel coordinates. Visual Comput 1985;1(4):69–91. doi:10.1007/BF01898350.

[5] Wegman EJ. Hyperdimensional data analysis using parallel coordinates. J Am Stat Assoc 1990;85(411):664–75. doi:10.2307/2290001.

[6] Inselberg A. Parallel coordinates: Interactive visualization for high dimensions. In: Zudilova-Seinstra E, Adriaansen T, Liere R, editors. Trends in Interactive Visualization. London, UK: Springer-Verlag; 2009. p. 49–78. doi:10.1007/978-1-84800-269-2_3.

[7] Heinrich J, Weiskopf D. State of the art of parallel coordinates. In: Proceedings of the eurographics state of the art reports. The Eurographics Association; 2013.

[8] Johansson J, Forsell C. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. IEEE Trans. Visual. Comput. Graph. 2016;22(1):579–88. doi:10.1109/TVCG.2015.2466992.

[9] Wang J, Liu X, Shen H, Lin G. Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. IEEE Trans Visual Comput Graph 2017;23(1):81–90. doi:10.1109/TVCG.2016.2598830.

[10] Steed CA, Ricciuto DM, Shipman G, Smith B, Thornton PE, Wang D, et al. Big data visual analytics for exploratory earth system simulation analysis. Comput Geosci 2013;61:71–82. doi:10.1016/j.cageo.2013.07.025.

[11] Steed CA, Swan JE, Jankun-Kelly TJ, Fitzpatrick PJ. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In: Proceedings of the IEEE symposium on visual analytics science and technology; 2009. p. 19–26. doi:10.1109/VAST.2009.5332586.

[12] Choi H, Lee H, Kim H. Fast detection and visualization of network attacks on parallel coordinates. Comput Secur 2009;28(5):276–88. doi:10.1016/j.cose.2008.12.003.

[13] Wang WB, Huang ML, Lu L, Zhang J. Improving performance of forensics investigation with parallel coordinates visual analytics. In: Proceedings of the IEEE International Conference on Computational Science and Engineering; 2014. p. 1838–43. doi:10.1109/CSE.2014.337.

[14] Boogaerts T, Tranchevent L, Pavlopoulos GA, Aerts J, Vandewalle J. Visualizing high dimensional datasets using parallel coordinates: Application to gene prioritization. In: Proceedings of the IEEE International Conference on Bioinformatics Bioengineering; 2012. p. 52–7. doi:10.1109/BIBE.2012.6399706.

[15] Keefe D, Ewert M, Ribarsky W, Chang R. Interactive coordinated multiple-view visualization of biomechanical motion data. IEEE Trans Visualizat Comput Graph 2009;15(6):1383–90. doi:10.1109/TVCG.2009.152.

[16] Caat MT, Maurits NM, Roerdink JBTM. Design and evaluation of tiled parallel coordinate visualization of multichannel EEG data. IEEE Trans Visual Comput Graph 2007;13(1):70–9. doi:10.1109/TVCG.2007.9.

[17] Qu H, Chan W, Xu A, Chung K, Lau K, Guo P. Visual analysis of the air pollution problem in hong kong. IEEE Trans Visual Comput Graph 2007;13(6):1408–15. doi:10.1109/TVCG.2007.70613.

[18] Siirtola H. Direct manipulation of parallel coordinates. In: Proceedings of the IEEE conference on information visualization; 2000. p. 373–8. doi:10.1109/IV.2000.859784.

[19] Hauser H, Ledermann F, Doleisch H. Angular brushing of extended parallel coordinates. In: Proceedings of IEEE symposium on information visualization; 2002. p. 127–30. doi:10.1109/INFVIS.2002.1173157.

[20] Cleveland WC, McGill ME. Dynamic Graphics for Statistics. Boca Raton, FL, USA: CRC Press, Inc.; 1988. doi:10.1214/ss/1177013104.

[21] Claessen JHT, van Wijk JJ. Flexible linked axes for multivariate data visualization. IEEE Trans. Visual. Comput. Graph. 2011;17(12):2310–16. doi:10.1109/TVCG.2011.201.

[22] Roberts JC. Exploratory visualization with multiple linked views. In: Exploring Geovisualization. Elsevier; 2005. p. 159–80. doi:10.1016/B978-008044531-1/50426-7.

[23] Yuan X, Guo P, Xiao H, Zhou H, Qu H. Scattering points in parallel coordinates. IEEE Trans Visual Comput Graph 2009;15(6):1001–8. doi:10.1109/TVCG.2009.179.

[24] Cuzzocrea A, Zall D. Parallel coordinates technique in visual data mining: Advantages, disadvantages and combinations. In: Proceedings of the international conference on information visualisation; 2013. p. 278–84. doi:10.1109/IV.2013.96.

[25] Zhou L, Weiskopf D. Indexed-points parallel coordinates visualization of multivariate correlations. IEEE Trans Visual Comput Graph 2018;24(6):1997–2010. doi:10.1109/TVCG.2017.2698041.

[26] Kosara R, Bendix F, Hauser H. Parallel sets: interactive exploration and visual analysis of categorical data. IEEE Trans Visual Comput Graph 2006;12(4):558–68. doi:10.1109/TVCG.2006.76.

[27] Fernstad SJ, Johansson J. A task based performance evaluation of visualization approaches for categorical data analysis. In: Proceedings of the international conference on information visualisation; 2011. p. 80–9. doi:10.1109/IV.2011.92.

[28] Vosough Z, Hogräfer M, Royer LA, Groh R, Schulz H-J. Parallel hierarchies: A visualization for cross-tabulating hierarchical categories. Comput Graph 2018;76:1–17. doi:10.1016/j.cag.2018.07.009.

[29] Johansson J, Ljung P, Jern M, Cooper M. Revealing structure within clustered parallel coordinates displays. In: Proceedings of the IEEE symposium on information visualization INFOVIS; 2005. p. 125–32. doi:10.1109/INFVIS.2005.1532138.

[30] Palmas G, Bachynskyi M, Oulasvirta A, Seidel HP, Weinkauf T. An edge-bundling layout for interactive parallel coordinates. In: Proceedings of the IEEE Pacific Visualization Symposium; 2014. p. 57–64. doi:10.1109/PacificVis.2014.40.

[31] Novotný M, Hauser H. Outlier-preserving focus+context visualization in parallel coordinates. IEEE Trans Visual Comput Graph 2006;12(5):893–900. doi:10.1109/TVCG.2006.170.

[32] Andrienko G, Andrienko N. Blending aggregation and selection: Adapting parallel coordinates for the visualization of large datasets. Cartogr J 2005;42(1):49–60. doi:10.1179/000870405X57284.

[33] Janetzko H, Stein M, Sacha D. Enhancing parallel coordinates: Statistical visualizations for analyzing soccer data. In: Proceedings of the visualization and data analysis conference; 2016. p. 1–8. doi:10.2352/ISSN.2470-1173.2016.1.VDA-486.

[34] Blaas J, Botha C, Post F. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. IEEE Trans Visual Comput Graph 2008;14(6):1436–51. doi:10.1109/TVCG.2008.131.

[35] Sansen J, Richer G, Jourde T, Lanlanne F, Auber D, Bourqui R. Visual exploration of large multidimensional data using parallel coordinates on big data infrastructure. Informatics 2017;4(3):1–21. doi:10.3390/informatics4030021.

[36] Richer G, Sansen J, Lanlanne F, Auber D, Bourqui R. Enabling hierarchical exploration for large-scale multidimensional data with abstract parallel coordinates. In: Proceedings of the international workshop on big data visual exploration and analytics; 2018. p. 8pp..

[37] Mackinlay J. Automating the design of graphical presentations of relational information. ACM Trans Graph 1986;5(2):110–41. doi:10.1145/22949.22950.

[38] Salvador S, Chan P. Fastdtw: Toward accurate dynamic time warping in linear time and space. In: Proceedings of the ACM KDD workshop on mining temporal and sequential data; 2004. p. 70–80.

[39] Rensink RA. Change detection. Ann Rev Psychol 2002;53:245–577. doi:10.1146/annurev.psych.53.100901.135125.

[40] Hamm CE, Merkel R, Springer O, Jurkojc P, Maier C, Prechtel K, et al. Architecture and material properties of diatom shells provide effective mechanical protection. Nature 2003;421:841–3. doi:10.1038/nature01416.

[41] Delalat B, Sheppard VC, Ghaemi SR, Rao S, Prestidge CA, McPhee G, et al. Targeted drug delivery using genetically engineered diatom biosilica. Nature Commun 2015;6:8791. doi:10.1038/ncomms9791.

[42] Landsea CW, Franklin JL. Atlantic hurricane database uncertainty and presentation of a new database format. Monthly Weather Rev 2013;141(10):3576–92. doi:10.1175/MWR-D-12-00254.1.

[43] Evans JL, Hart RE. Objective indicators of the life cycle evolution of extratropical transition for atlantic tropical cyclones. Monthly Weather Rev 2003;131(5):909–25. doi:10.1175/1520-0493(2003)131<0909:OIOTLC⟩2.0.CO;2.

[44] Trofimov AA, Pawlicki AA, Borodinov N, Mandal S, Mathews TJ, Hildebrand M, et al. Deep data analytics for genetic engineering of diatoms linking genotype to phenotype via machine learning. NPJ Comput Mater 2019;5(1):67. doi:10.1038/s41524-019-0202-3.