Lit Review for handling numerical ties in PCPs

Denise Bradford

Introduction

Our data-driven modern civilization now depends heavily on the ability to precisely evaluate and exploit huge and complex datasets, for example, in small rural towns where resources are often sparse. This dissertation aims to make big datasets more understandable and valuable by utilizing three related innovations: generalized parallel coordinate plots (GPCPs), reliable ways to look at missing data, and efficient methods to elaborate on the detection of unusual trends. Each priority area is key for improving the GPCP technique to enable thorough and accurate data analysis across various datasets.

The accuracy of data interpretation is highest when dealing directly with numerical values or one-dimensional visual representations (Spence 1990). That is becoming more and more impractical as the size and dimension of datasets continue to increase. To make judgments based on data, it is necessary to find methods of simplifying and consolidating datasets without sacrificing significant information. Data visualization, such as graphs or tables, is often an essential and invaluable intermediate step between raw data and an observer's decision-making process. An effective data visualization communicates the gist of a data set in a way that is easy for an observer to understand and evaluate; ideally, data visualizations slightly sacrifice accuracy (relative to a table of values) in favor of a greater understanding of the relationship between observations or variables.

Most standard data visualizations work within the Cartesian coordinate system, with variables or functions of variables mapped to the x and y axis. Additional variables can be mapped to different properties of the plotted points, bars, or lines, and analysts can also create small multiples that show subsets of the data; even with these additions, viewers quickly become overwhelmed by the amount of information when more than p = 3 or 4 variables are shown (including the x and y coordinates). When it is necessary to understand the relationship between more than four variables, visualizations on a Cartesian grid no longer work as well; extensions to additional dimensions are also ineffective (**reference?**). Other approaches are necessary when it is necessary to understand $p > 4$ dimensions of data. Here, we examine parallel coordinate plots (PCPs) as a solution to $p > 4$ dimensional data visualization and assess the impact of different modifications of PCPs on their effectiveness for visualizing high N and/or high p dimensional data. We specifically evaluate the ability of each PCP version to

facilitate the identification of overall trends, outliers, and clusters within $N > 4$ dimensional data across different magnitudes of N (observations) and p (variables).

Stimulus-responsive attention often arises when an observer recognizes that a particular graphical element in a stimulus signals valuable information, guiding further search to improve judgment accuracy. Spence (1990) talked about the importance of visual psychophysics in figuring out how simple parts of graphs can be used to convey information. He stressed that how well we can understand graphs relies on our cognitive and visual abilities. Cleveland and McGill's study on graphical perception in 1985 found that people have different skill levels when decoding multidimensional data (Cleveland and McGill 1984). This shows that designs like generalized parallel coordinate plots must consider cognitive and perceptual variability. These plots show many dimensions within a single visual glyph, and the small changes between them make it much harder to understand complicated visual data. Additional studies, like Heer and Bostock (2010) and Simkin and Hastie (1987), built on these ideas by showing that mistakes happen at different points in information processing and that different graphs are better for different jobs depending on the user's accuracy. It was emphasized by Carswell (1992) and Shah & Freedman (2011) that task difficulty and perceptual processes interact. These articles say differences in how people think and perceive things should be considered when making generalized parallel coordinate plots to show multidimensional data.

Parallel Coordinate Plots (PCPs)

Why?

Parallel coordinate plots (PCPs) are an excellent way to show data with many dimensions since each is shown on its plane. This means that complex relationships between different factors can be studied repeatedly. PCPs are great for drawing attention to trends, clusters, and outliers and efficiently show many variables in a small area. However, as with their Cartesian equivalents, PCPs are vulnerable to overplotting and can become difficult to read or interpret when N is large, but because they overplot, they can get messy with large datasets. It can be more challenging to understand the data than with Cartesian coordinates, and you must go through a learning process to get the majority of them as well.

Cartesian coordinates, on the other hand, are straightforward to understand because they are simple and well-known. They make it easier to tell the difference between different information points and their exact values by displaying data points in an easily understood manner. Cartesian coordinates are an excellent method to present data clearly because they can handle up to three dimensions. They also need help with growth because adding more dimensions requires a lot of subplots or complicated three-dimensional plots, which can get boring and challenging to follow. Cartesian graphs additionally can take up a lot of room, and for data with more dimensions, they usually need more than one plot.

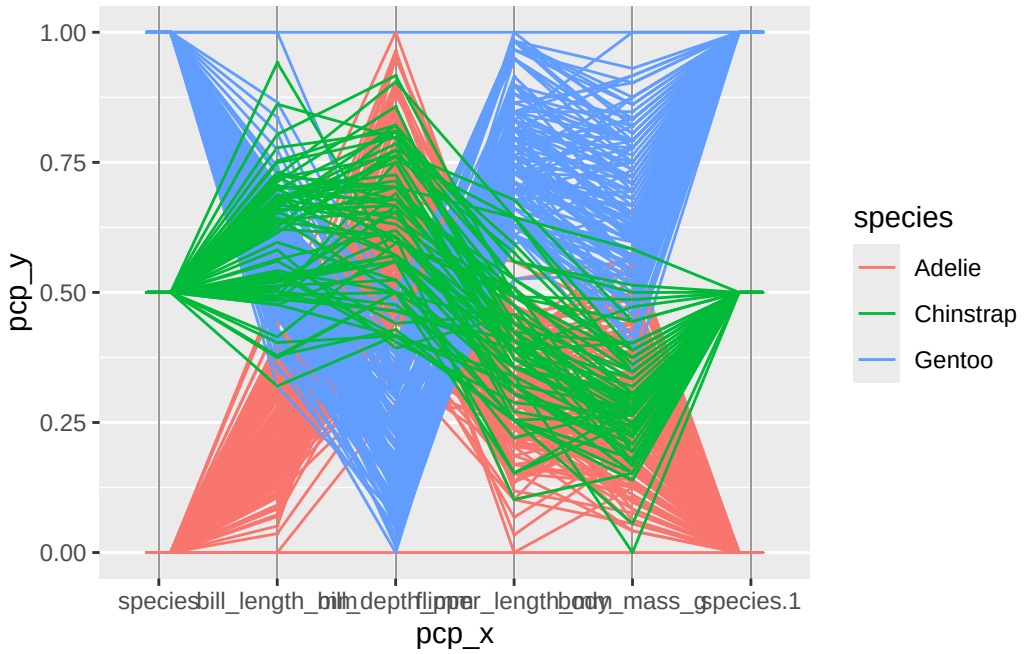

Figure 1: A parallel coordinate plot.

What?

Parallel coordinate plots (PCPs) leverage the projective space rather than the Cartesian coordinate system: each line in cartesian space is a set of points in projective space, and each point in cartesian space can be represented as a line in projection space (Inselberg 1985).
The result is that a single data point is For high-dimensional data, parallel coordinate plots (PCPs) are a standard visualization tool whereby data points are shown as a line intersecting a series of vertical axes representing the different variables in the dataset, therefore indicating a dimension. Conventional PCPs can be difficult to interpret because of overplotting - lines overlap and need to be distinguishable, and when there are large datasets, even non-identical segments can be hard to identify because of the number of line segments. It needs to improve its interpretability by having clutter and overlapping lines while handling big datasets. Generalized PCPs (GPCPs) introduced several different methods: By including axis reordering, bundling, and dimensionality reduction to solve these problems, generalized parallel coordinate plots (GPCPs) expand the conventional PCP. These improvements in clarity and usability of PCPs utilizing data pattern highlighting and visual clutter reduction help (Heinrich and Weiskopf 2009), for example, present several approaches to bundle comparable trajectories in GPCPs, so minimizing overlap and improving pattern identification; Johannsen et al. (2012) address the advantages of dynamic axis reordering to highlight different data characteristics.

XXX Show a PCP and a Scatterplot matrix representing the same data. Focus on numeric data and use a standard PCP for this. Ensure p>=4 and that all variables are numeric. XXX

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

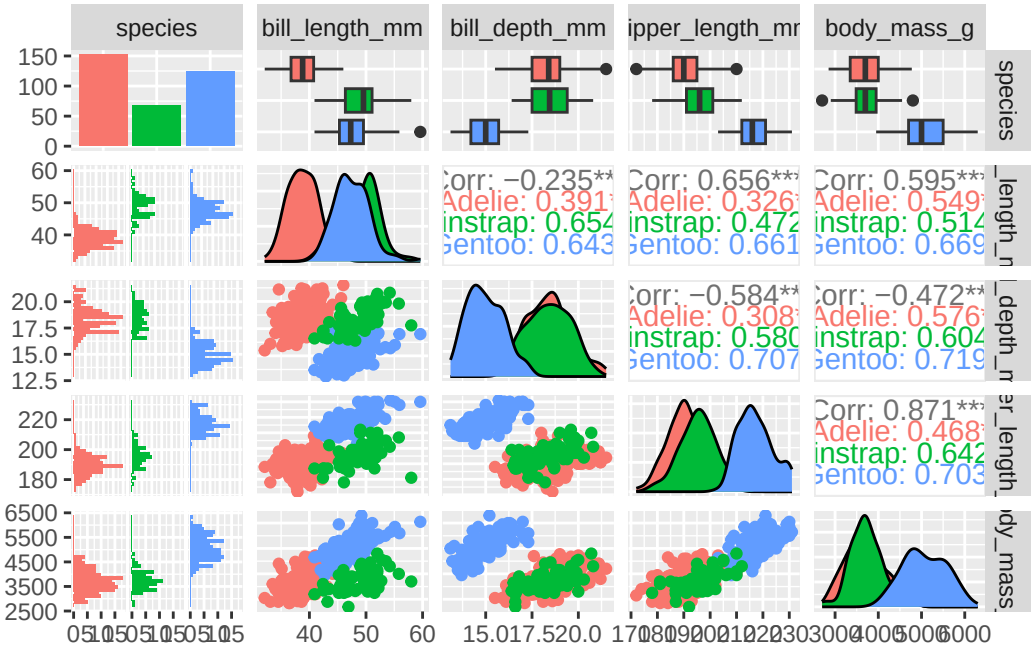

Figure 2: A generalized pairs plot

Modifications to Parallel Coordinate Plots

Since Alfred Inselberg introduced parallel coordinate plots (PCPs) in 1985, considerable advances have been made to address the original method's framework and improve its capacity to depict high-dimensional data effectively. The enhancements include changes to visual representation, handling various kinds of interactive data elements, and incorporating advanced computational approaches to increase the clarity and interpretability of the visualizations. Here are some specific enhancements and modifications that have been made:

Interactive and Dynamic

The addition of interactive features has substantially increased the usefulness of PCPs. Interactive PCPs enable users to dynamically change axes, reorder them, filter data based on specified criteria, and even invert axes to study data from various angles. These characteristics enable real-time data analysis, leading to a deeper understanding of patterns and correlations.

Brushing and connecting allow users to highlight specific data points across many axes, whereas filtering hides lines that do not satisfy set criteria, decreasing clutter. The interactive characteristics of PCPs make them more amenable to exploratory data analysis, particularly in big and complicated datasets (Heinrich & Weiskopf, 2013).

Heinrich and Weiskopf's 2013 paper "State of the Art of Parallel Coordinates" thoroughly reviews visualization methods for parallel coordinates. It includes a taxonomy that groups the different approaches into different categories. The writers discuss various ways to model, see, and work with parallel coordinates. They also show how these methods can be used for everyday tasks in knowledge discovery, like sorting, clustering, and regression. Some of the most important advances are discussing geometric models, interpolation methods, and the point-line duality that makes up the basis of parallel coordinate plots. The study discusses a better understanding of data using density-based visualizations, axis ordering, and improvements such as brushing and bundling.

The study points out some problems with parallel coordinates, like overplotting and the need for good axis arrangement. It also suggests ways to get around these problems, such as clustering and density estimation. To make it easier to see patterns and understand data, researchers look into different ways to map it, such as curves, shapes, and density plots. The writers also talk about how parallel coordinates can be used in real life in engineering and the life sciences. This shows how useful and flexible they are for high-dimensional data visualization tasks.

Initially, moodifications in parallel coordinate plots have dynamic features, which allow users to select, highlight, and filter data dimensions or specific data points in real-time. This has facilitated more precise and focused analysis within large datasets. Interaction allows users to engage with data intuitively, selectively focusing on areas of interest. "Interactive filtering and brushing make it easier to identify patterns within large datasets, aiding discovery and hypothesis generation" (Wegman 1990).

In 1990, Wegman et al. introduced fundamental techniques for interaction with multidimensional data, laying the groundwork for modern interactive PCPs. Siirtola and Räihä (2006) developed interactive features like brushing and linking to assist in dimensional analysis, making comparing attributes within large datasets easier. Inselberg (2009) expanded these features by enabling dynamic filtering and axis rearrangement to tailor the display to user needs.

Increased complexity can overwhelm novice users, requiring more computational power and potentially slowing down analysis in high-dimensional datasets. Inselberg (2009) noted, "While interactivity provides insight, it can lead to user fatigue in complex datasets due to the cognitive load required for multi-step filtering."

Bundling and Curving

Bundling and curving techniques have been introduced to handle visual clutter, especially with high-dimensional data. These techniques group similar paths, making patterns more visible and reducing overplotting. Bundling and curving significantly reduce clutter, making spotting overarching trends and common patterns easier. Holten and van Wijk noted that "bundling offers a compelling solution to mitigate the chaos often found in high-dimensional visualizations" (Holten and Van Wijk 2009).

Holten and van Wijk pioneered edge bundling for PCPs, which visually aggregates similar paths to reduce clutter. McDonnell and Mueller (2008) furthered this approach by introducing curvilinear PCPs, where curved lines help distinguish intersecting paths for clearer visual analysis. Johansson and Forsell (2016) evaluated the effectiveness of bundling and curving in PCPs, establishing criteria for when these modifications enhance interpretability.

Curved lines can be altered to vary in curvature dependent on data attributes, which improves visual separation and pattern recognition, particularly in scenarios with strongly correlated dimensions. This strategy has improved PCPs' ability to visualize complicated datasets with overlapping points (Kachhway, 2013).

Moustafa et al. combine parallel coordinate plots (PCPs) with parallel coordinate density plots (PCDPs) to reduce clutter when working with big datasets. As part of Moustafa's methodology, the standard PCP is mutated into a density plot. This creates plot areas with more observations that stand out and reduce visual clutter.

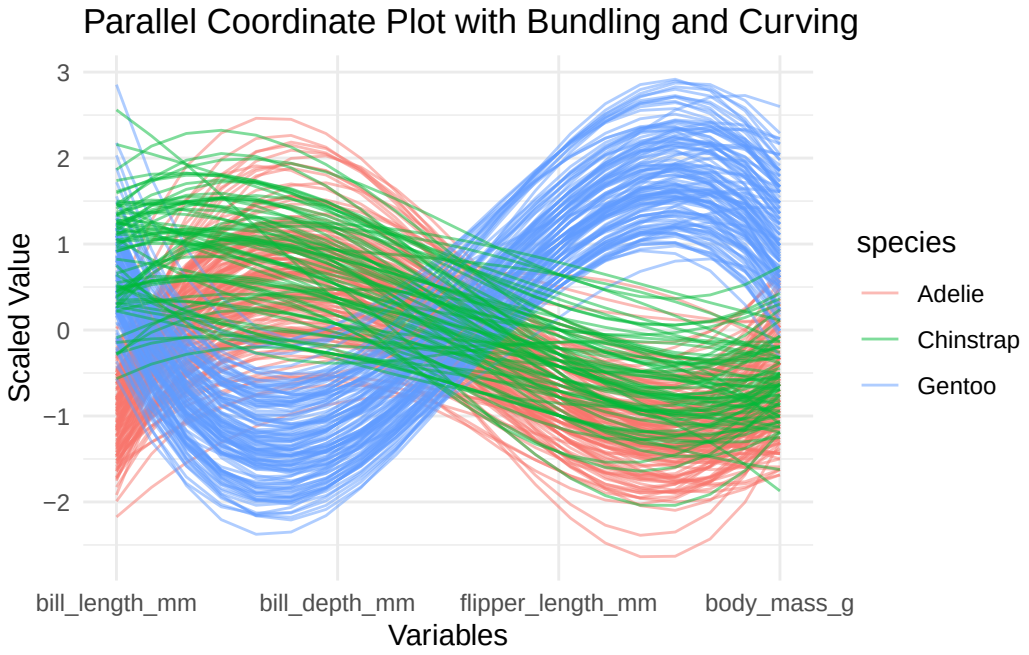

Figure 3: Parallel Coordinate Plot with Bundling and Curving

The new approach uses density estimation methods to transform the PCP image based on polylines into a continuous, smooth depiction of data density. This change shows how to see groups and trends usually hidden in regular PCPs. To make things even better, Moustafa has interactive parts that let users experience different data dimensions in real-time. This makes PCPs even better at analysis. With its interactive nature, people who aren't experts can use this method confidently and successfully.

Moustafa's research indicates that to deal with even larger datasets more efficiently, future efforts should focus on enhancing density estimation methods. These methods can be used in biology and finance to demonstrate their usefulness and adaptability for examining real-world data (Moustafa 2011).

Bundling groups with similar paths may also obscure individual data points or outliers, which can be crucial in some analyses. "The clarity gained in bundle simplification comes at the cost of data precision," warns McDonnell and Mueller (2008), pointing out that outliers or unique data paths might be lost.

Dimension Reduction Techniques

Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are frequently used to improve the readability of PCPs before charting. These strategies help to reduce the number of dimensions while retaining the most informative parts of the data. PCPs can be used to visualize reduced dimensions, providing a better understanding of high-dimensional interactions. Axis ordering is another important feature that has been adjusted. The relationships between variables can be highlighted by sorting the axes according to measurements such as correlation or mutual information. Automatic axis arrangement techniques have been developed to reduce line crossings, making recognizing patterns easier.

The article "Orientation-Enhanced Parallel Coordinate Plots" by Raidou et al. describes a new way to make parallel coordinate plots (PCPs) more accessible to read and understand. The authors suggest a method that uses input about direction to solve the clutter and overlap problems with regular PCPs. This method changes the direction of the plot axes on the fly by applying both automatic orientation and user-interactive adjustments. This method reduces visual noise and makes data patterns and correlations more visible. This will make it easier for users to study and analyze large datasets.

The suggested orientation-enhanced PCPs are designed with the user's needs at the forefront, offering a more informative visual representation. Several preprocessing steps, including principal component analysis (PCA) and multi-dimensional scaling (MDS), were employed to determine the optimal axis orientation. This ensures that data clusters are maximally separated and data lines are minimally overlapped. The interactive features allow users to adjust the orientation according to their preferences and requirements, enhancing the usability and adaptability of the plots for research needs (Raidou et al., 2015).

Dimension reduction has become critical for managing PCPs with high-dimensional data. Techniques such as principal component analysis (PCA) and clustering help to focus on the most significant data features. Reducing dimensions simplifies analysis, enabling analysts to focus on the most important features while avoiding overwhelming visual data. "Dimension reduction helps to retain essential information without drowning the user in less relevant details," claims Wegman and Luo (1997).

Wegman and Luo (1997) applied PCA within PCPs, enabling users to represent key data features while minimizing less relevant dimensions. Guo et al. (2011) utilized clustering and hierarchical dimension reduction techniques to streamline high-dimensional datasets, enhancing usability without sacrificing detail. Yang et al. (2013) demonstrated the integration of non-linear dimension reduction methods within PCPs to capture complex relationships in data more effectively.

Dimension reduction techniques like PCA or clustering can inadvertently remove nuances or minor patterns that might be relevant in specific cases. Guo et al. (2011) caution that, "while beneficial, these methods might lead to information loss, potentially masking relationships only visible in higher dimensions."

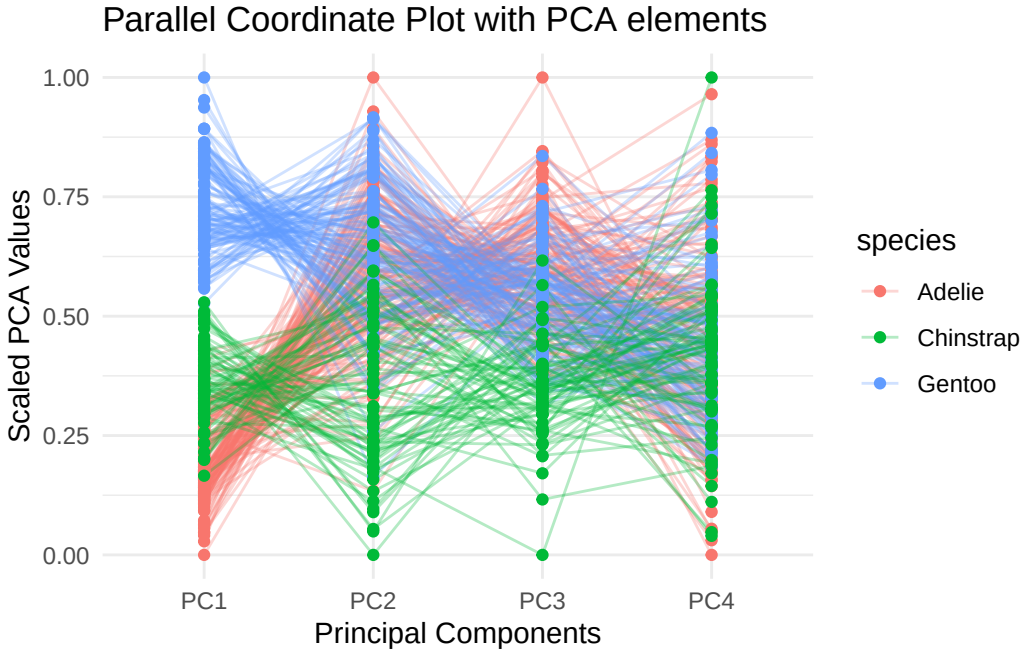

Figure 4: PCA in Parallel Coordinate Plot

Enhanced Color Encoding and Shading

Visual upgrades like color coding, opacity modifications, and changing line thickness have been implemented to solve overplotting concerns, particularly with massive datasets. Significant data trends can be highlighted using different colors or modifying line opacity in response to data density, with less relevant information deemphasized. This approach helps users identify trends and outliers that might otherwise be hidden. Line thickness can be adjusted to reflect other variables, such as data frequency or confidence intervals, adding depth to the visual depiction. Such multi-layered visualization techniques provide an additional layer of information to classic PCPs

In their 2013 work "State of the Art of Parallel Coordinates," Heinrich and Weiskopf give a full overview. They discuss the different methods used to avoid overplotting and improve data interpretation. They discuss how to use alpha blending, which changes the transparency of lines to clear up space, and interactive line manipulation methods that let users explore different data points in real-time. These tips are important for making parallel coordinate plots easier to read, especially when working with big datasets.

In their 2008 study, "Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-Timepoint Data Sets," Blaas et al. stress how important it is to have visual traits that can be changed, such as the colors and opacity of clusters. Their method lets users change these settings in real-time, which makes it easier to focus on specific data features and control line density. In the same way, Raidou et al. (2015), in "Orientation-Enhanced Parallel Coordinate Plots," talk about how different levels of opacity and color can help draw attention to patterns, show line densities, and allow for interactive data study. In their work, they also talk about how to deal with the problem of overplotting in complex datasets by smoothing and averaging polylines.

In their paper "Visual Clustering in Parallel Coordinates," Zhou et al. (2008) make another important addition by suggesting a tool that lets users choose color and opacity to draw attention to clusters in the data. This method helps tell the difference between data groups and changes the shape of curves to make things clearer. In their 1999 paper "Hierarchical Parallel Coordinates for Exploration of Large Datasets," Fua et al. also talk about ways to reduce visual clutter by changing the thickness of lines and using hierarchical data representations. These methods make it possible to effectively visualize data density, which gives you a better look at the trends hidden in big datasets.

These studies show that changing how parallel coordinate plots look, such as color coding, adjusting brightness, and moving lines around, can make them much easier to read and understand. These methods are significant for dealing with thick and overlapped data, making parallel coordinate plots a more helpful tool for studying data in multiple dimensions.

By adding color gradients and shading, PCPs can encode additional data attributes, such as density or frequency, improving the ability to spot trends and correlations. Color gradients and shading effectively encode additional variables, allowing more complex data insights to be

derived visually. "Color encoding adds a new layer of perception, turning a purely structural plot into a multi-dimensional analysis tool," explains Theus and Urbanek (2009).

Theus and Urbanek (2009) introduced color-coded parallel coordinates, which allow users to map additional variables using color and thereby increase interpretive power. Bertini et al. (2005) applied density shading techniques to PCPs to make frequent patterns in large datasets stand out more prominently. Novotny and Hauser (2006) developed opacity and color-blending techniques in PCPs, which help understand overlapping patterns more intuitively.

Heavy reliance on color may lead to visual strain, especially in cases where multiple gradients overlap or in users with color vision deficiencies. Bertini et al. (2005) noted, "When too many colors are applied, the encoding becomes confusing, and can overwhelm rather than enhance the viewer's understanding."

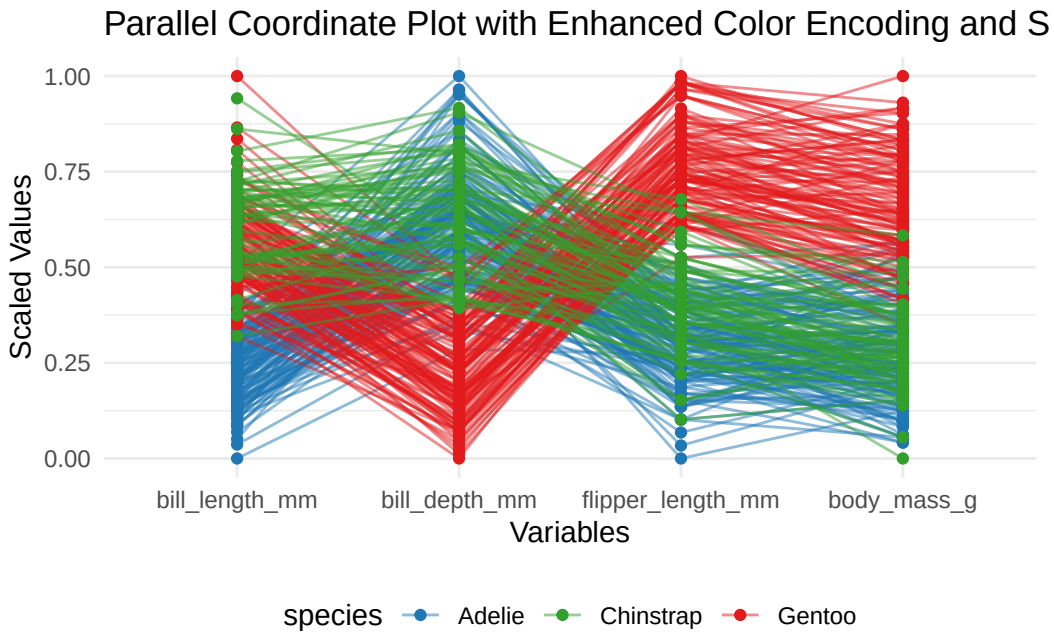

Figure 5: Enhanced Color Encoding and Shading in PCP

Reordering and Axis Flipping

Adaptive reordering and axis flipping based on correlation measures or user-defined parameters can simplify the analysis of multidimensional relationships. Reordering and axis flipping align more relevant dimensions, minimizing intersections and making relationships more interpretable. "Reordering transforms PCPs from a cluttered tangle into a roadmap of relationships," states Inselberg and Dimsdale (1990).

Inselberg and Dimsdale (1990) proposed reordering to enhance interpretability, especially for highly interrelated variables. LeBlanc et al. (1990) introduced correlation-based reordering to position axes, reducing visual clutter by aligning more related dimensions. Peng et al. (2004) enabled automatic reordering algorithms that flip axes according to user-defined weights, giving users more control over PCP readability.

Automatic reordering may reduce control over axis sequence, potentially misaligning dimensions relevant to a specific analytical question. LeBlanc et al. (1990) pointed out that "the gain in interpretability through automated ordering can sometimes sacrifice user intention, misaligning axes crucial to the analysis context."

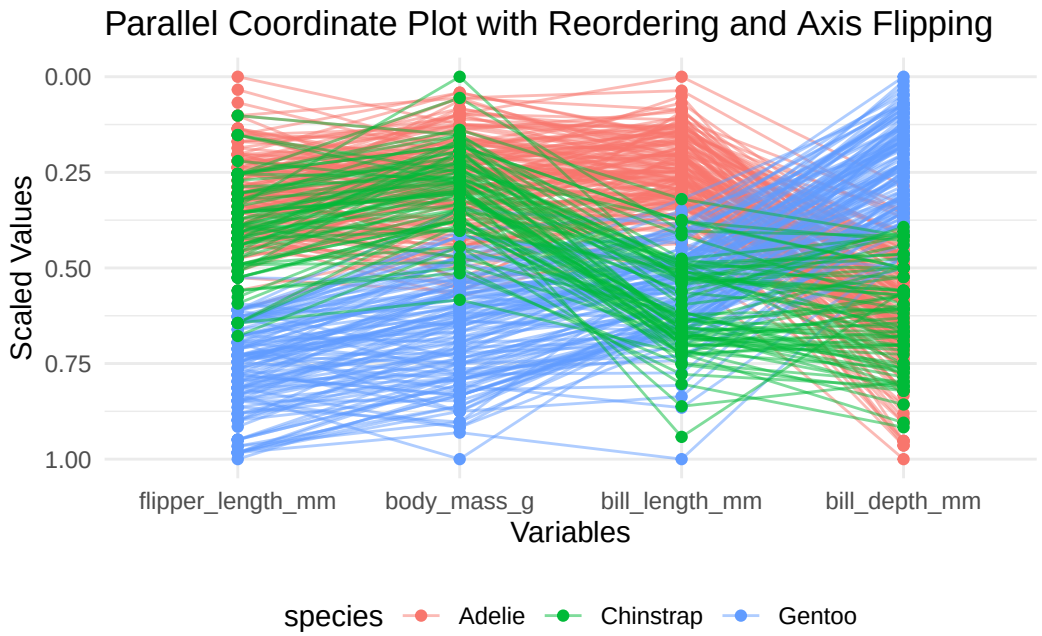

Figure 6: Reordering and Axis Flipping in PCP

Cluster-Based and Hierarchical PCPs

Clustering techniques, such as spectral clustering, have been used in PCPs to group related data points and expose the underlying structure of high-dimensional datasets. Clustering data and viewing the clusters in parallel coordinate plots makes detecting links between data points and cluster features easier. This technique is beneficial for discovering patterns not immediately evident in raw data and providing insights into the natural grouping of data points (Zhao & Kaufman, 2012). Several data reduction procedures, such as hierarchical clustering and principal component analysis, are used before plot generation to improve interpretability. These strategies help to limit the number of dimensions visualized by focusing on the most important components that explain the majority of the variance in the data.

Zhao and Kaufman's (2012) work "Structure Revealing Techniques Based on Parallel Coordinates Plot" talks about how traditional parallel coordinate plots (PCPs) can't always show important patterns in complex, high-dimensional data because of problems like overplotting. To deal with these problems, the writers suggest new ways to sort and cluster data specifically made for PCPs. Using spectrum theory, they develop algorithms that group similar polylines and sort the axes to show hidden trends and correlations. This makes the structure of the data easier to see. A correlation-based sorting method is also added to arrange the axes to make relationships between variables stand out. This makes it easier to see trends across dimensions.

The study also talks about view-range metrics, which use aggregation limits to help visualize data more clearly, even when the datasets are noisy. Results from experiments show that these improvements make it much easier for PCPs to find useful patterns, trends, and correlations, which makes data analysis more efficient. The results show that the suggested approaches improve PCPs' ability to analyze big, complicated datasets by showing important data structures that would be hard to see otherwise.

For multidimensional data, clustering and hierarchical PCPs allow data grouping and organized representation to enable easier cluster comparison. Clustering and hierarchical visualization allow for data grouping, clarifying large datasets and enhancing comparison among groups. "Hierarchical clustering in PCPs offers a clearer, more organized data narrative by showing both the big picture and finer details," argues Fua et al. (1999).

Fua et al. (1999) introduced hierarchical PCPs, providing a zoomable interface to represent data clusters, allowing for detailed subset examination. Geng et al. (2004) applied clustering in PCPs to group similar data points, highlighting clusters and making general patterns more accessible. Etemadpour et al. (2015) extended these techniques by incorporating hierarchical clustering for user-defined levels of granularity.

Although clusters simplify the data landscape, they can obscure individual data points. Geng et al. (2004) observed that "while useful for general patterns, clustering may bury unique or outlier data, potentially hiding significant findings in homogenous groups."

Optimized Layout Algorithms

Layout algorithms for PCPs have been optimized to handle multi-dimensional data while minimizing line crossings, which enhances clarity and allows for smoother navigation. Optimized layouts minimize line crossings, making PCPs more straightforward, particularly with extensive, high-dimensional data. "Effective layout design in PCPs can be the difference between comprehensibility and chaos," Ankerst et al. (1998) claim, emphasizing that thoughtful placement enhances legibility.

Ankerst et al. (1998) proposed a layout algorithm that places axes to minimize line crossings, increasing the legibility of high-dimensional PCPs. Johansson et al. (2005) optimized the

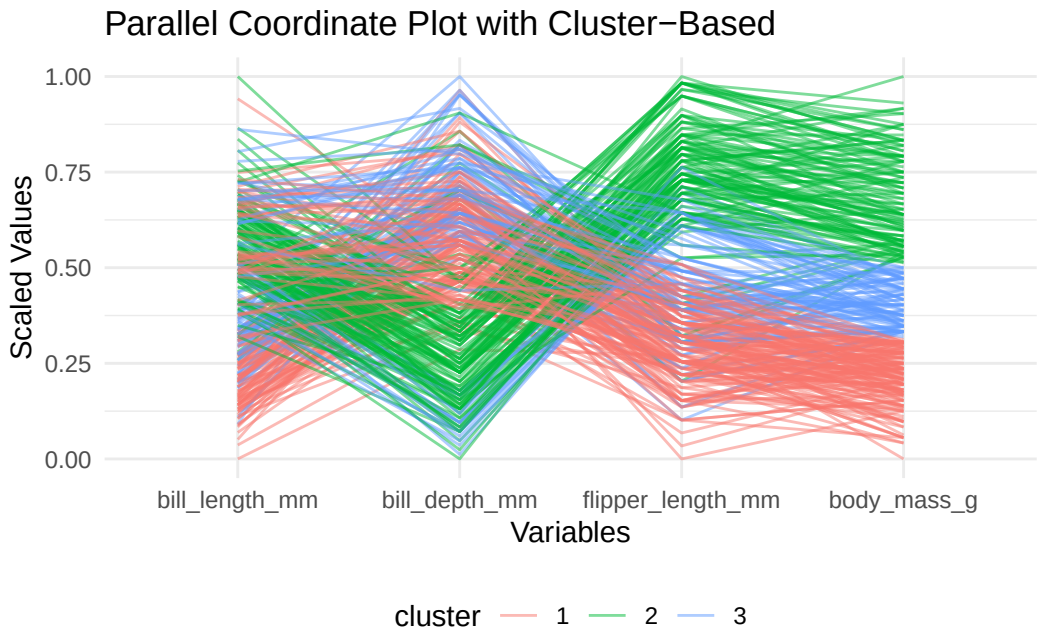

Figure 7: Cluster-Based in PCP

axis positioning algorithm to maximize the space between overlapping lines, improving data distinction. Hao et al. (2007) implemented a randomized layout approach that provides an efficient layout for PCPs, beneficial for large, high-dimensional data.

Layout optimizations sometimes follow heuristic approaches that may not always align with user-defined analysis needs, possibly overlooking preferred layouts. Johansson et al. (2005) state that "algorithmically optimized layouts prioritize minimal intersections over interpretability, occasionally misaligning with the user's focus."

Categorical and Hybrid PCPs

Traditional PCPs were primarily designed to collect continuous data. However, real-world datasets often contain a mix of continuous, ordinal, and categorical data. To solve this, changes have been made to display mixed-type data within the same PCP simultaneously. Categorical data, for example, can be visually differentiated using unique colors, symbols, or segmented lines, although continuous variables are still represented by standard lines linking the axes. Categorical Parallel Coordinate Plots (CPCPs) were created to handle the challenge of representing categorical data in PCPs, which are not suitable for continuous axis representation. Specific strategies, like adjustments to the ends of links, have been suggested to enhance how category and numerical values are connected visually. This strategy enhances the interpretability of mixed data by modifying the plotting criteria for axes representing category variables.

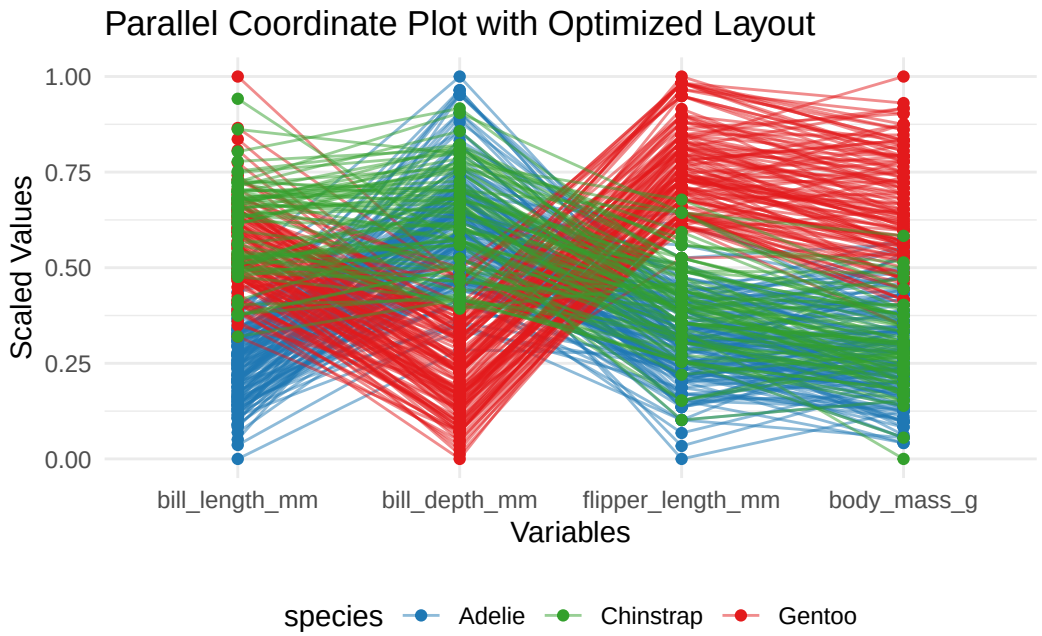

Figure 8: Optimized Layout in PCP

CPCPs use discrete axis segments or unique markers for each category, transforming categorical variables into distinguishable visual elements. Siirtola and Räihä (2006) introduced axis segmentation, using markers and spacings to represent categorical values distinctly, improving categorical data visualization in PCPs. Inselberg (2009) enhanced this representation by introducing specific encodings, which allowed categories to be differentiated visually, preserving the PCP's interpretative power. Johansson and Forsell (2016) explored alternative segmentation methods to minimize visual clutter, enabling effective handling of multiple categorical values on the same axis.

The Generalized Parallel Coordinate Plot (GPCP) is a modified version of the original PCP that introduces nonlinear changes to the axes for more complex data representation. These can show more complicated relationships in the data that might not be visible in a regular PCP. Nonlinear scales such as logarithmic and exponential transformations are possible. In their work in 1997, Wegman and Luo provided a comprehensive discussion of the GPCP idea. This adjustment allows the visualization to display various data connections and handle skewed data distributions.

CPCPs adjust by organizing categorical axes more clearly or using symbols and colors to represent categories for easier interpretation. Wegman et al. (1990) suggested using symbols and color codes to represent categories, improving how categorical data is understood in parallel coordinates. Holten and van Wijk (2009) introduced color gradients for categories to facilitate distinguishing between multiple categorical values, especially in the analysis of intricate datasets. LeBlanc et al. (1990) highlighted the significance of arranging axes in CPCPs, posi-

tioning categories with strong relationships closer together to assist users in making meaningful comparisons.

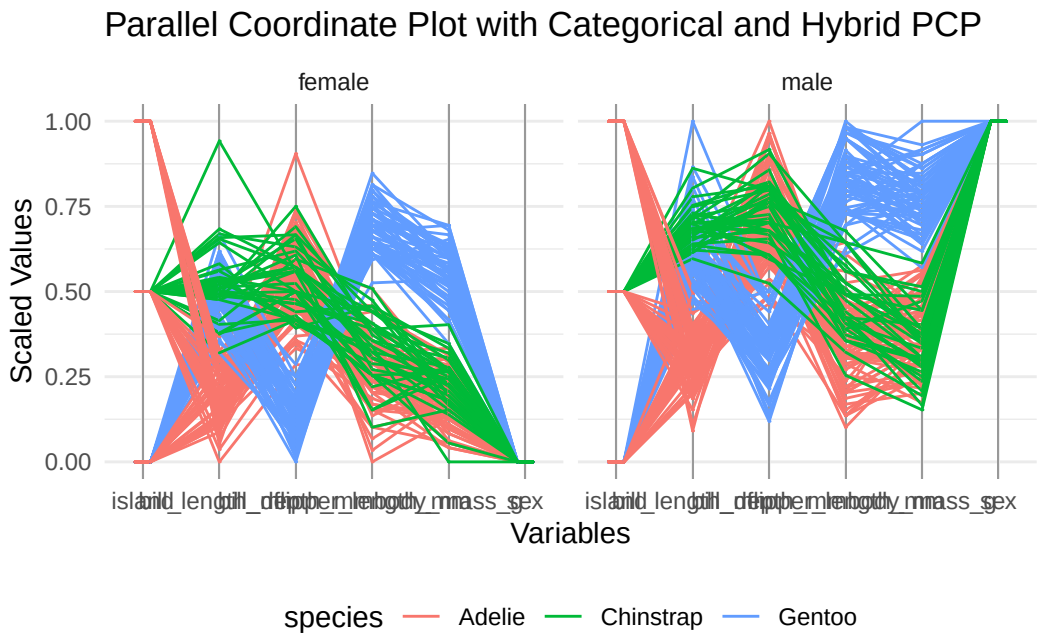

Figure 9: Categorical and Hybrid in PCP

Cognitive Perceptions in Changing of Bins

When histograms have fewer bins, more data points fit into each bin, making the data distribution look less smooth. This method is like using fewer splits to make numerical attributes discrete, which leads to bigger, more general groups. This kind of generalization could hide important differences in the data, making it too easy to understand the trends beneath. Wilkinson's research (1999) shows that this rough representation can give the wrong impression that the data is all the same, hiding important trends that could be important for making decisions.

On the other hand, adding more bins to a histogram is like using more granular splits, which separate continuous data into smaller periods. In the same way, this method lets the parallel coordinate plots tell the difference between different numbers more precisely, and it finds more detailed trends in the data. According to Talbot et al. (2010), this can show small patterns and give more details, but it can also add noise or make random changes stand out more. In the same way, too much discretization can cause overfitting, where the model picks up noise in the data instead of trends that can be used in real life.

According to Tufte (2001), too many bins can confuse the user, making it hard to find helpful information in the data. When there are fewer splits, the output can become easier to

understand and prone to overfitting because it may focus too much on specific cases of the data instead of overall trends. Finding the right mix between granularity and generalization is important in histogram binning and decision tree discretization.

Aesthetic biases in data display can also change how discretization methods are seen. For example, people are more likely to like histograms that look balanced and symmetrical, which can change how they understand the data. This is similar to how parallel coordinate plots with evenly spread breaks across feature values might be seen as more accurate or reliable, even if the data doesn't support this. Kirk (2016) says these visual preferences can change how the data is interpreted, leading to a wrong view.

Finally, choosing the right bin width in a histogram is similar to choosing the right number of intervals when discretizing numerical features for parallel coordinate plots from a statistics point of view. To avoid giving false impressions, both methods need to be carefully chosen. Narrow bins in a histogram, like too many breaks in a numerical variable, can make small changes stand out more, leading to overfitting and wrong conclusions. Freedman and Diaconis (1981) say that picking the correct bin width (or number of splits) is important to avoid these problems and ensure the discretization method finds meaningful patterns without complicating the visualization.

In conclusion, the cognitive theories that explain how we see changing histogram bins easily apply to separating numbers in parallel coordinate plots. Both require giving up some details to keep things simple, significantly affecting how the data is interpreted and how well the visualization works. It's important to find the right mix between underfitting and overfitting, and the discretization method needs to match the type of data and the analysis that will be done.

Cleveland, William S, and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–54.

Heer, Jeffrey, and Michael Bostock. 2010. "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–12.

Heinrich, J, and D Weiskopf. 2009. "Continuous Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 1531–38. https://doi.org/10.1109/TVCG.2009.131.

Holten, Danny, and Jarke J Van Wijk. 2009. "Force-Directed Edge Bundling for Graph Visualization." In *Computer Graphics Forum*, 28:983–90. 3. Wiley Online Library.

Inselberg, Alfred. 1985. "The plane with parallel coordinates." *The Visual Computer* 1 (2): 69–91. https://doi.org/10.1007/BF01898350.

Moustafa, Rida E. 2011. "Parallel Coordinate and Parallel Coordinate Density Plots." *Wiley Interdisciplinary Reviews: Computational Statistics* 3 (2): 134–48.

Simkin, David, and Reid Hastie. 1987. "An Information-Processing Analysis of Graph Perception." *Journal of the American Statistical Association* 82 (398): 454–65.

Spence, Ian. 1990. "Visual Psychophysics of Simple Graphical Elements." *Journal of Experimental Psychology: Human Perception and Performance* 16 (4): 683.

Wegman, Edward J. 1990. "Hyperdimensional data analysis using parallel coordinates." *Journal of the American Statistical Assoiation* 85: 664–75.