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. A common way to visually show high-dimensional data is with parallel coordinate plots (PCPs). These plots are especially good at showing relationships and trends between multiple factors. However, seeing ties and overlapping data points can be hard when many numbers are the same or very close. This can cause important information to be lost in the research. This study fixes the problems that regular PCPs have with finding numerical ties by adding a new method that makes it easier to tell the difference between data sets that overlap. We want to create a methodology that makes it easier to tell the difference between ties by using a standard delta difference approach. Our suggested methods should make it easier to understand data, leading to more accurate research and decision-making in large, complex 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.

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. These improvements in clarity and usability of PCPs utilizing data pattern highlighting and visual clutter reduction help (J. Heinrich and Weiskopf 2009), for example, present several approaches to bundle comparable trajectories in PCPs, so minimizing overlap and improving pattern identification; Johannsen et al. (2012) address the advantages of dynamic axis reordering to highlight different data characteristics.

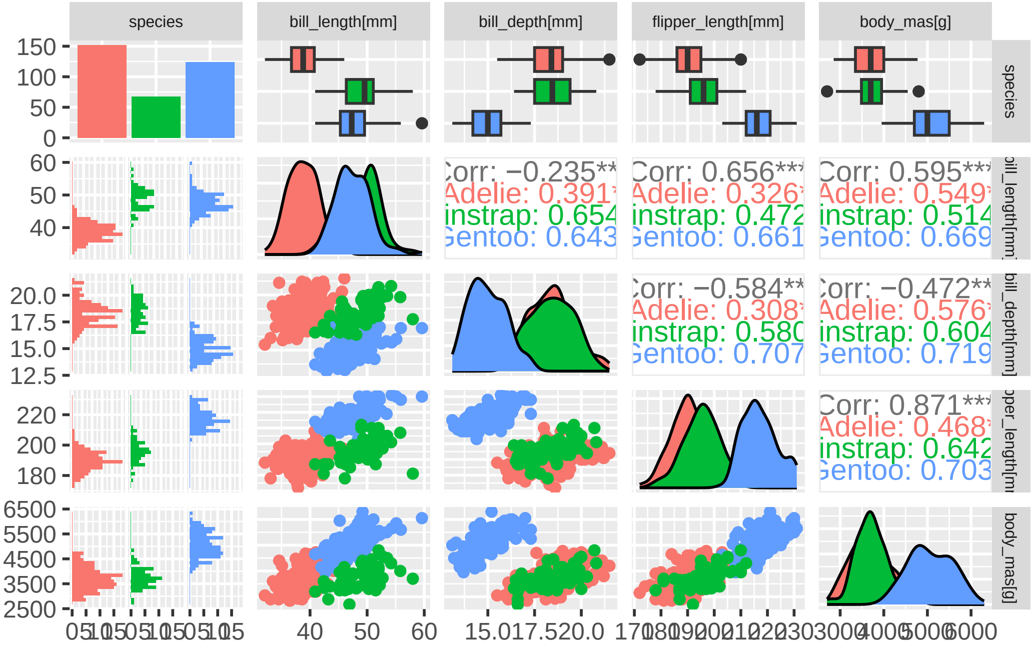

Figure 1: A generalized pairs plot

Figure 1 is a generalized pairs plot, containing histograms, density plots, scatter plots, and correlation matrices. It shows how the variables for three kinds of penguins (Adelie, Chinstrap, and Gentoo) are distributed and how they are related. Histograms and density plots make it easy to see how each variable (bill length, bill depth, flipper length, and body mass) is spread across the different species. The density plots show clear trends for each species, regarding the length of the bills and flippers, where Gentoo penguins tend to have higher values. In the boxplots that show body mass and flipper length, outliers are clear because they show numbers very different from the rest of the data.

The scatter plots show how two variables are related to each other. There is a clear trend for overplotting in denser areas, especially for bill depth vs. bill length and flipper length vs. bill length. The correlation values show more information about these connections. They show strong positive and negative correlations, with some species having different patterns. For example, in Gentoo penguins, there is a strong positive correlation between the length of their flippers and their body mass. Other species have different association patterns, which means that the ways that traits are related can be different for each species. Overall, the matrix-style plot does an excellent job of showing the links within and between variables. However, it is hard to understand precisely what overlapping areas mean in scatter plots because individual points are hard to see.

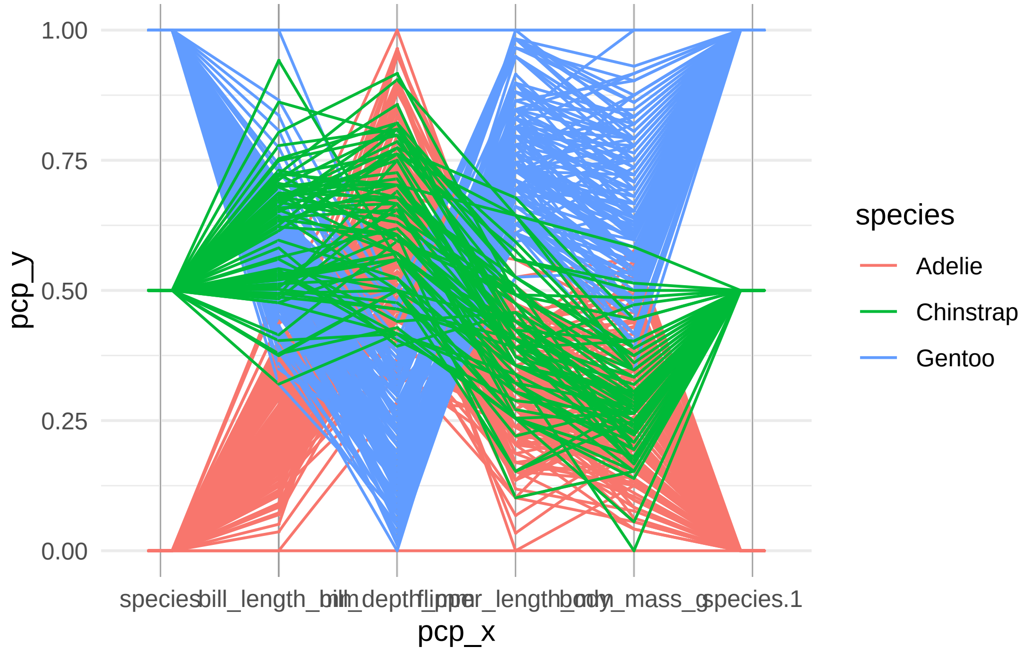

Figure 2: A parallel coordinate plot

In Figure 2, each line shows a different penguin, and the colors of the lines show the species. The Gentoo (blue) tends to have higher body mass and flipper length values, while the Adelie (red) tends to have lower values for these measurements. However, the large number of lines makes it hard to see smaller details, and it's hard to pick out individual spots or clear outliers in this visualization.

It's hard to directly find outliers because the merging lines make it hard to see individual differences. The plot doesn't show specific distribution shapes like a histogram or density plot, but how the lines are grouped at specific value ranges shows how each species generally behaves. The slope and intersection patterns of the lines suggest that the variables are related. Note that the plot demonstrates that if the length of a bird's bill goes up, other variables may go up or down similarly, based on the species. Overall, the plot works well for seeing broad patterns and trends across species, but it could be better at picking out specific data points or weak connections because it's too crowded on the screen.

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 (Julian Heinrich and Weiskopf 2013b).

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, modifications 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 developed interactive features like brushing and linking to assist in dimensional analysis, making comparing attributes within large datasets easier (Siirtola and Räihä 2006). Inselberg expanded these features by enabling dynamic filtering and axis rearrangement to tailor the display to user needs (Inselberg 2009).

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

The D3 Graph Gallery has a live parallel coordinates graph with dynamic features, making exploring and analyzing data easier. Users can move their mouse over lines to see specific data tracks. This interactivity feature makes finding and separating different data types or trends in the dataset easier. The graph also lets the user “brush” on any vertical axes, allowing users to choose and filter data points based on specific ranges or criteria. This interactivity feature changes the visualization on the fly to show the filtered group. Tooltips show up when you move their mouse over lines or axis names. They give the user more information about the data points, which helps the user understand them better. The user can change the order of the axes by dragging them horizontally. This interactivity feature lets the user change how dimensions are compared, which makes it easier to find patterns and correlations between different factors. The parallel coordinates plot is a powerful way to see and understand multidimensional data in a way that is easy to understand and use because it has these dynamic parts.

Interactive Parallel Coordinate Plot

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 furthered this approach by introducing curvilinear PCPs, where curved lines help distinguish intersecting paths for clearer visual analysis (McDonnell and Mueller 2008). Johansson and Forsell evaluated the effectiveness of bundling and curving in PCPs, establishing criteria for when these modifications enhance interpretability (Johansson and Forsell 2015).

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. Curving in PCPs visually separates overlapping lines, making data relationships more discernible. It improves line separation by reducing intersections, making it easier to differentiate between trajectories. Curved lines are more aesthetically pleasing and reduce cognitive load, enhancing readability. They simulate a more three-dimensional space within a two-dimensional PCP, making it easier to perceive data points' relative positioning. Research shows curving lines reduces visual clutter and improves data interpretation Qu et al. (2007). This strategy has improved PCPs' ability to visualize complicated datasets with overlapping points (Kachhway 2013).

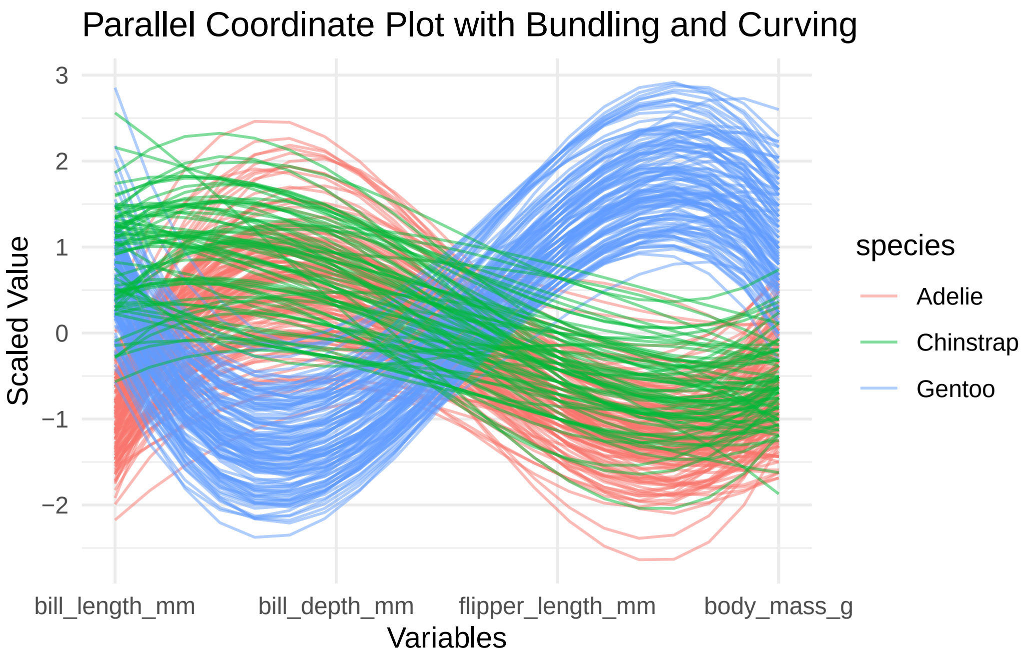

Figure 3: A Bundling and Curving Parallel Coordinate Plot

In Figure 3, the parallel coordinate plot adds bundling and curving, which makes the lines smoother, less distracting, and better able to tell the patterns between the species apart.

It's easier to see general trends within each species when the data is bundled, but it still doesn't show specific distribution shapes like a histogram might. Bundling makes it harder to see outliers in this image because the extreme values are "pulled" into the main flow, making it harder to see individual deviations. The curved lines make the relationships between the variables stand out, indicating that specific measurements may go up or down together depending on the species. Still, they make it harder to see small changes between individual points.

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 (Moustafa 2011).

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, pointing out that outliers or unique data paths might be lost (McDonnell and Mueller 2008).

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, (Raidou et al. 2015).

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," (Wegman and Luo 1997).

Wegman and Luo (1997) applied PCA within PCPs, enabling users to represent key data features while minimizing less relevant dimensions (Wegman and Luo 1997). Guo et al. (2011) utilized clustering and hierarchical dimension reduction techniques to streamline high-dimensional datasets, enhancing usability without sacrificing detail (Guo, Xiao, and Yuan 2012). Yang et al. (2013) demonstrated the integration of non-linear dimension reduction methods within PCPs to capture complex relationships in data more effectively (Yang et al. 2017).

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."

In Figure 4, the parallel coordinate plot incorporates principal component analysis (PCA) components (PC1 to PC4) on the x-axis, with the scaled PCA values on the y-axis, color-coded by species (Adelie in red, Chinstrap in green, and Gentoo in blue). PCA reduces dimensionality, meaning that each PC represents a combination of the original variables, simplifying complex relationships. The lines between components show how each penguin transitions across these derived features, making it easier to observe general trends for each species in the reduced space. Due to the numerous overlapping lines, overplotting remains an issue, though species-level patterns are visible, with distinct separations in specific PCs. Outliers are challenging to detect due to the bundled nature of the lines. While distribution shapes aren't explicitly visible, the clustered points at each PC axis indicate each species' general concentration of values. Though individual variability remains obscured, relationships between PCs and species are

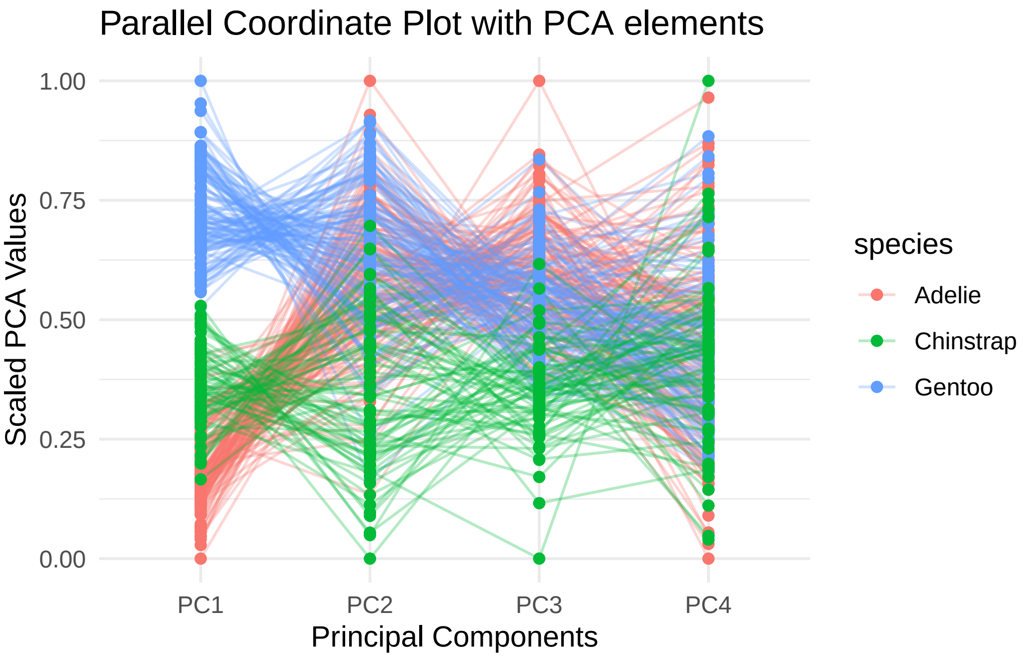

Figure 4: A PCA in Parallel Coordinate Plot

implied through color-coded clusters and trends, suggesting some separation between species along specific principal components.

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, (Julian Heinrich and Weiskopf 2013b).

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 (Blaas, Botha, and Post 2008). In the same way, Raidou et al., 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 (Raidou et al. 2015). 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. 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 (Zhou et al. 2008). In their 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 (Fua, Ward, and Rundensteiner 1999). 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,” (Theus and Urbanek 2008).

Theus and Urbanek introduced color-coded parallel coordinates, which allow users to map additional variables using color and thereby increase interpretive power (Theus and Urbanek 2008). Bertini et al. applied density shading techniques to PCPs to make frequent patterns in large datasets stand out more prominently (Bertini, Dell’Aquila, and Santucci 2005). Novotny and Hauser developed opacity and color-blending techniques in PCPs, which help understand overlapping patterns more intuitively (Novotny and Hauser 2006).

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. noted, “When too many colors are applied, the encoding becomes confusing, and can overwhelm rather than enhance the viewer’s understanding.”

In Figure 5, the parallel coordinate plot uses enhanced color coding. It might have changed the line density, which makes the differences between species (Adelie is red, Chinstrap is green, and Gentoo is blue) more obvious across the variables (bill length, bill depth, flipper length, and body mass). The plot does a good job of showing general trends for each species. For

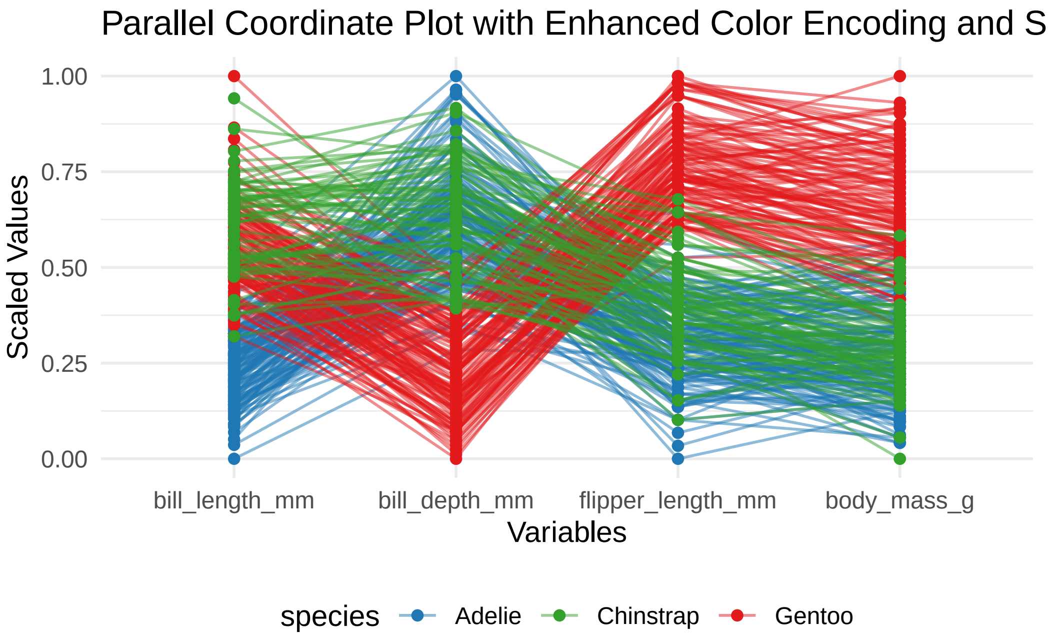

Figure 5: Enhanced Color Encoding and Shading in PCP

example, the Gentoo tends to have higher flipper length and body mass values, while the Adelie tends to have lower values. Overplotting is still a problem, especially near the middle ranges of each axis, which hides small differences and makes it hard to find outliers directly. The plot doesn't show individual distribution shapes, but how the lines are grouped on each axis suggests that each species has central tendencies. The general trends of each species' lines suggest relationships between numerical factors, but it still needs to be easier to figure out individual correlations because they overlap. This graph gives a broad picture of trends across species, but it could be better at picking out specific data points or rare observations.

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," (Inselberg and Dimsdale 1990).

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

Changing the order of the axes in parallel coordinate plots can make them easier to read by putting variables that are strongly correlated or connected by theme next to each other. Moving the axes around reduces the amount of visual noise caused by crossings between variables that aren't strongly linked to each other. This makes patterns and clusters stand out more. The method of minimizing visual noise helps show hidden patterns in the data that would not be seen otherwise. Researchers Johansson et al. (2005) and Wegman and Luo (1997) all found that reordering is an important way to find connections in PCPs. If you use well-thought-out reordering algorithms, you can see groups, trends, and outliers more clearly and avoid the feeling of overplotting. This method works incredibly well for grouping data with many dimensions, making it easier to see.

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

Axis flipping is a way to eliminate unnecessary line crossings and bring out trends in data by flipping the scale of one or more axes. It works well when two PCP axes next to each other have negatively correlated factors. This change allows different patterns to show up without too many crossings, giving a clearer picture of how the data is related. Axis flipping is an important tool for dealing with negative correlations because it clarifies the connection between dimensions by reducing visual interference. It also shows patterns of correlation that regular PCPs might miss, especially when working with datasets that have many factors that are unrelated to each other or are related in the opposite way. Automated axis flipping methods can change plots on the fly based on data, which lowers the chance of mistakes when users interpret PCPs. Heinrich and Dasgupta's studies both agree that axis flipping makes PCPs easier to understand by turning bad relationships into a more familiar shape Julian Heinrich and Weiskopf (2013a).

Reordering and flipping axes based on correlation or user-defined parameters simplify the interpretation of complex relationships. "Dynamic reordering optimizes the alignment of correlated variables, making it easier to identify associations and trends across dimensions" (Yuan et al. 2009). This adaptation increases flexibility, allowing the plot to respond to specific analytical needs and user preferences.

Reordering and axis shifting work well together to fix the issue of overplotting in PCPs. Rearranging linked variables makes them easier to see while flipping the axes reduces the number of times lines cross for negatively correlated variables. Combining these two tools makes looking for patterns and exploring large datasets easier.

By using these methods, users can better understand how multiple variables are connected, which helps them make better decisions based on visual information. Research shows that these methods make parallel coordinate plots easier to understand, more accurate, and better able to find groups and trends, all of which are very important for data analysts.

Continuous reordering can create inconsistency, as users may find it challenging to track relationships when axes are frequently altered. “Constant reconfiguration of the axes can disrupt the analytical flow, making it difficult to build a stable mental model of the data structure” (Julian Heinrich and Weiskopf 2013a). Additionally, for datasets with minimal correlation, this method may offer limited value, as reordering may not result in clearer insights.

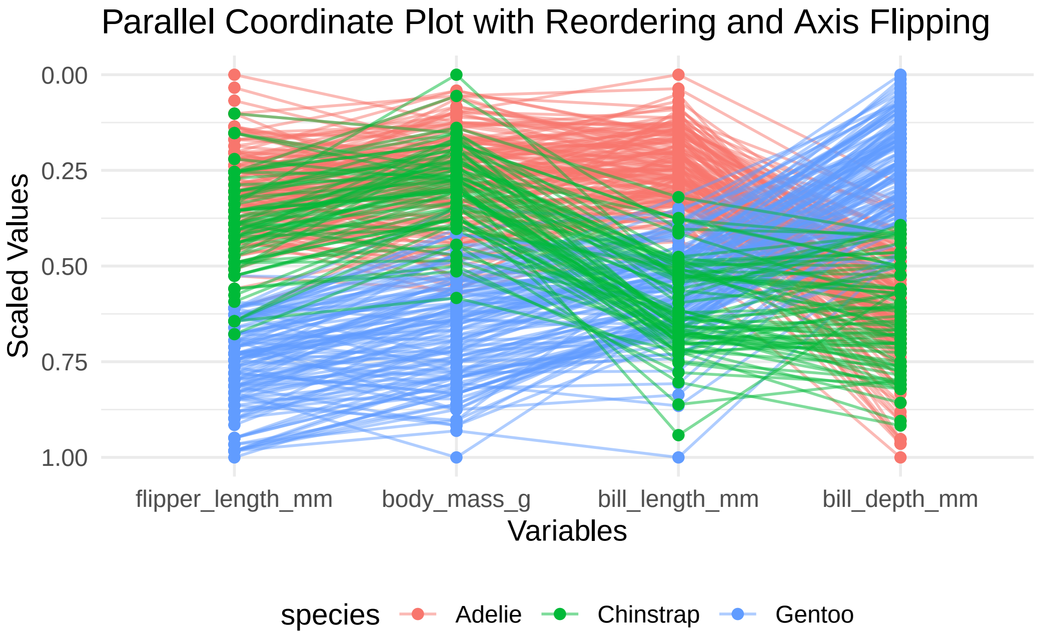

Figure 6: Reordering and Axis Flipping in PCP

In Figure 6, the parallel coordinate map changes the variables by rearranging them and flipping the axes to show patterns and connections between the penguin species more clearly. By flipping some axes and putting related factors closer together (like body mass and flipper length), the plot makes species-based trends stand out more: The Gentoo (blue) tends to have a shallower bill but a bigger body and longer flippers, while the Adelie (red) tends to have the opposite pattern. Overplotting is still a problem, especially when values are close together in the middle ranges, making it hard to tell the difference between lines or peaks. You can’t see the distributions’ exact forms, but the lines’ density and spread along each axis give you a good idea of the range of values for each species. The new order makes relationships between variables stand out more because the changes between the axes make species-based correlations in measurement trends more apparent. However, individual relationships are still hard to see because lines cross.

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 and 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 (Zhao and Kaufman 2012).

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. (Fua, Ward, and Rundensteiner 1999).

Fua et al. introduced hierarchical PCPs, providing a zoomable interface to represent data clusters, allowing for detailed subset examination (Fua, Ward, and Rundensteiner 1999). Geng et al. applied clustering in PCPs to group similar data points, highlighting clusters and making general patterns more accessible (Geng, Deng, and Ali 2005). Poco et al. extended these techniques by incorporating hierarchical clustering for user-defined levels of granularity (Poco et al. 2011).

Although clusters simplify the data landscape, they can obscure individual data points. Geng et al. observed that "while useful for general patterns, clustering may bury unique or outlier data, potentially hiding significant findings in homogenous groups," (Geng, Deng, and Ali 2005).

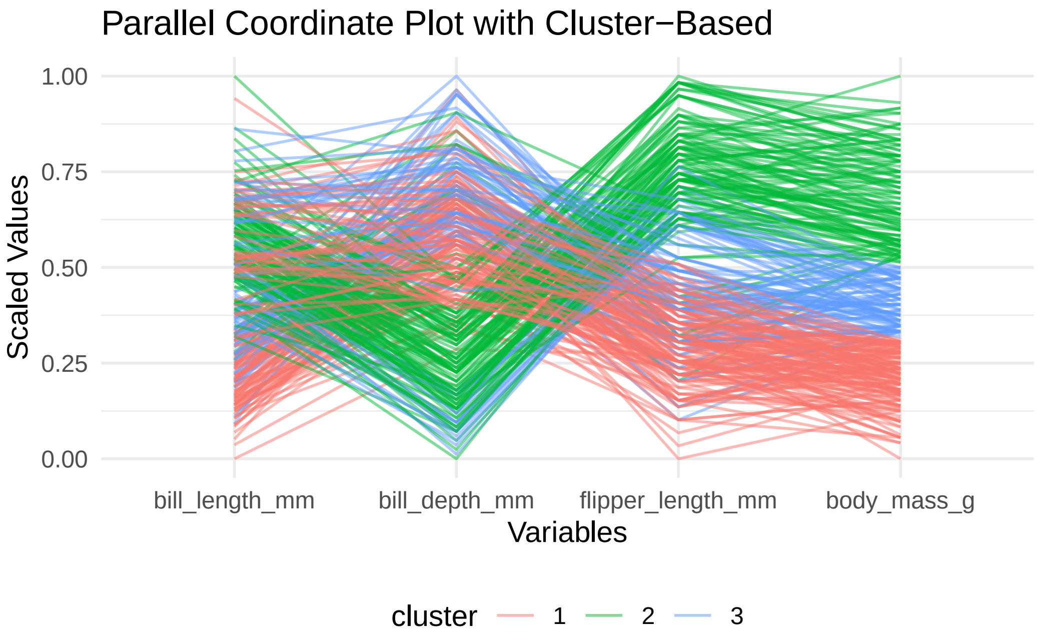

Figure 7: Cluster-Based PCP

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. claim, emphasizing that thoughtful placement enhances legibility (Ankerst, Berchtold, and Keim 1998).

Ankerst et al. proposed a layout algorithm that places axes to minimize line crossings, increasing the legibility of high-dimensional PCPs. Johansson et al. optimized the axis positioning algorithm to maximize the space between overlapping lines, improving data distinction (Johansson et al. 2005). 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. state that "algorithmically optimized layouts prioritize minimal intersections over interpretability, occasionally misaligning with the user's focus.," (Johansson et al. 2005).

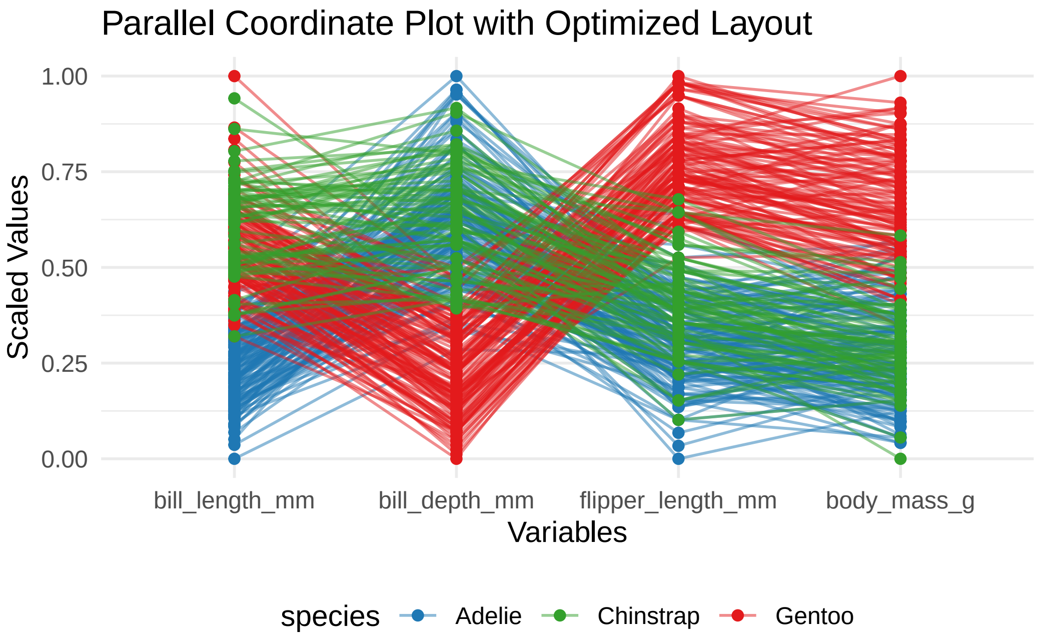

Figure 8: Algorithm Optimized Layout in PCP

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.

CPCPs use discrete axis segments or unique markers for each category, transforming categorical variables into distinguishable visual elements. Siirtola and Räihä introduced axis segmentation, using markers and spacings to represent categorical values distinctly, improving categorical data visualization in PCPs, (Siirtola and Räihä 2006). Inselberg enhanced this representation by introducing specific encodings, which allowed categories to be differentiated visually, preserving the PCP's interpretative power (Inselberg 2009). Johansson and Forsell explored alternative segmentation methods to minimize visual clutter, enabling effective handling of multiple categorical values on the same axis, (Johansson and Forsell 2015).

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 (Wegman and Luo 1997).

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

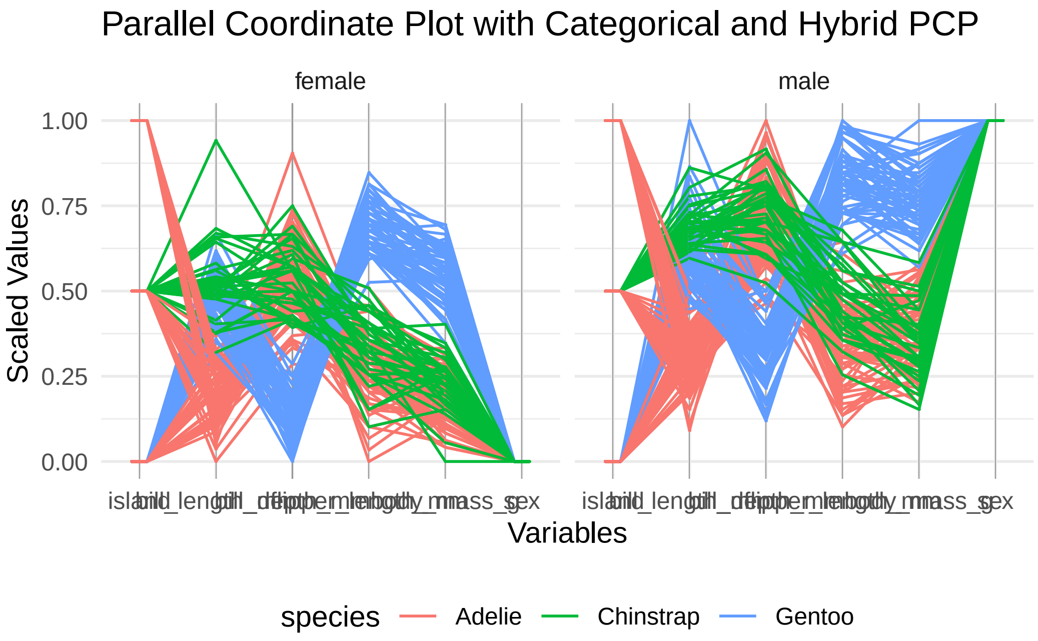

Figure 9: Categorical and Hybrid in PCP

Cognitive Perceptions in Changing of Bins (???)

In order to accurately capture the nuances of the data distribution, it is important to carefully consider the number of bins used in a histogram. Increasing the number of bins in a histogram enhances data categorization by providing a more precise representation of patterns and trends. This method prevents oversimplification and ensures essential differences are not missed. Wilkinson's research 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 (Wilkinson 2012).

On the other hand, adding more bins to a histogram is like using more granular splits, which separate continuous data into smaller periods. Similarly, this technique helps parallel coordinate plots distinguish between numbers more accurately, revealing detailed data trends. According to Talbot et al., this method can reveal subtle patterns and provide additional details, but it may also introduce noise or highlight random changes more prominently (Talbot, Gerth, and Hanrahan 2012). 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.

Tufte suggests that an excessive number of bins can confuse users, making it challenging to locate valuable information in the data (Tufte 2001). Tufte says that visual noise can make a histogram useless when bins are too small, or histograms are stacked on top of each other because viewers may be unable to tell the difference between important parts of the data distribution. Tufte stresses how important clarity is in data visualization by talking about how overplotting can be bad. He supports simple designs that put reading important data insights ahead of dense representations. His criticism aligns with empirical studies in graphical perception that show overplotting can confuse viewers, making them less likely to correctly understand distribution shapes or spot outliers when the display is too much.

Ultimately, the process of selecting the optimal bin width in a histogram mirrors the task of choosing the correct number of intervals when discretizing numerical features for parallel coordinate plots, highlighting the statistical importance of these decisions in data visualization and analysis. 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 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 (Freedman and Diaconis 1981).

Wickham's paper "Bin-summarise-smooth: a framework for visualizing large data" discusses finding the right mix between data compression and order preservation when binning numerical data for visualization (Wickham 2013). Wickham's research highlights the necessity of binning methods to take into account the initial order of the data. This is important for big datasets where distances between data points, like flight times, are crucial for understanding the data's meaning. Wickham's method makes visualizations smoother while showing how the data is distributed by carefully grouping data points into bins.

Wickham's framework makes visualizing very big datasets with little computing power easy. It handled millions of records from U.S. airlines by turning raw data into summarized files, which used much less memory. Then, these bins were smoothed out so that the visualization could be understood without needing a lot of computing power. The system is good for large-scale data analytics because it is efficient.

On the other hand, Wickham's system needs initial binning, which can cause information loss when data points are close together, like minute-by-minute temperature readings over a day. If the weather data for a day were broken up into hourly chunks, changes within an hour might be missed. This rough binning might make it hard for users to see small changes or trends that could be important in fields that need to be precise about time, like stock trading or watching the environment.

If the original bins don't closely match the distribution, the method might give a wrong picture of the data. Wickham's binning method might need to be carefully tweaked in situations where the data isn't normal, like when most of the numbers in an income distribution are close together at the lower end, and a few are very high. If these changes aren't made, the bins may not give a true picture of the distribution, highlighting or under-highlighting certain parts of the data. To fix this problem, users may need to carefully change the sizes of the bins or use different binning methods to ensure that the distributions are shown correctly.

Hauskrecht and Xue examine how to train classification models and the importance of bins for keeping data in order while sorting values. They use equal-distance binning to divide number ranges without messing up the order too much. This method effectively sorts things into groups while keeping the basic order. It does this without losing the interpretive value of number distances, which is important for tasks that depend on how close two data points are to each other (Xue and Hauskrecht 2017).

Assigning data points to bins of similar size in equal-distance binning helps eliminate small differences or noise within each bin, making outliers or extreme values less important. This noise reduction can help classification models generalize better by letting them focus on bigger trends in the data instead of small ones that don't fit. For example, when using income data to guess how a customer will act, equal-distance binning could lessen the effect of odd income numbers that could mess up the model otherwise. Getting rid of noise makes the model more stable and less likely to overfit on extreme values, which leads to better prediction and stability.

However, using fixed bin sizes can make data too simple in places with many people, so there needs to be more detailed information in these gaps. For example, when sensors record fast temperature changes, equal-distance binning may group several small changes into a single bin, hiding smaller but possibly important changes. When you oversimplify, important changes in dense data areas may not be picked up as well. This makes it harder for models to find trends that depend on these small differences. To get around this problem, you might need more processing power or a custom binning approach to keep important insights in areas with a lot of data.

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. Finding the correct balance between underfitting and overfitting is crucial, and the discretization method should align with the data type and analysis requirements.

References

Ankerst, Mihael, Stefan Berchtold, and Daniel A Keim. 1998. "Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data." In *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*, 52–60. IEEE.

Bertini, Enrico, Luigi Dell'Aquila, and Giuseppe Santucci. 2005. "Springview: Cooperation of Radviz and Parallel Coordinates for View Optimization and Clutter Reduction." In *Coordinated and Multiple Views in Exploratory Visualization (CMV'05)*, 22–29. IEEE.

Blaas, Jorik, Charl Botha, and Frits Post. 2008. "Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-Timepoint Data Sets." *IEEE Transactions on Visualization and Computer Graphics* 14 (6): 1436–51.

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.

Dasgupta, Aritra, and Robert Kosara. 2010. "Pargnostics: Screen-Space Metrics for Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 16 (6): 1017–26.

Freedman, David, and Persi Diaconis. 1981. "On the Histogram as a Density Estimator: L 2 Theory." *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 57 (4): 453–76.

Fua, Ying-Huey, Matthew O Ward, and Elke A Rundensteiner. 1999. *Hierarchical Parallel Coordinates for Exploration of Large Datasets*. IEEE.

Geng, Huimin, Xutao Deng, and Hesham Ali. 2005. "A New Clustering Algorithm Using Message Passing and Its Applications in Analyzing Microarray Data." In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, 6–pp. IEEE.

Guo, Hanqi, He Xiao, and Xiaoru Yuan. 2012. "Scalable Multivariate Volume Visualization and Analysis Based on Dimension Projection and Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 18 (9): 1397–1410.

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, Julian, and Daniel Weiskopf. 2013a. "State of the Art of Parallel Coordinates." In *Eurographics 2013 - State of the Art Reports*, edited by M. Sbert and L. Szirmay-Kalos. The Eurographics Association. https://doi.org/10.2312/conf/EG2013/stars/095-116.

———. 2013b. "State of the Art of Parallel Coordinates." *Eurographics (State of the Art Reports)*, 95–116.

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.

———. 2009. "Parallel Coordinates: Interactive Visualisation for High Dimensions." *Trends in Interactive Visualization: State-of-the-Art Survey*, 49–78.

Inselberg, Alfred, and Bernard Dimsdale. 1990. "Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry." In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, 361–78. IEEE.

Johansson, Jimmy, and Camilla Forsell. 2015. "Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 579–88.

Johansson, Jimmy, Patric Ljung, Mikael Jern, and Matthew Cooper. 2005. "Revealing Structure Within Clustered Parallel Coordinates Displays." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 125–32. IEEE.

Kachhway, Inder Singh. 2013. "Enhancement in Visualization of Parallel Coordinates Using Curves."

LeBlanc, Thomas J, John M Mellor-Crummey, and Robert J Fowler. 1990. "Analyzing Parallel Program Executions Using Multiple Views." *Journal of Parallel and Distributed Computing* 9 (2): 203–17.

McDonnell, Kevin T, and Klaus Mueller. 2008. "Illustrative Parallel Coordinates." In *Computer Graphics Forum*, 27:1031–38. 3. Wiley Online Library.

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

Novotny, Matej, and Helwig Hauser. 2006. "Outlier-Preserving Focus+ Context Visualization in Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 12 (5): 893–900.

Poco, Jorge, Ronak Etemadpour, Fernando Vieira Paulovich, TV Long, Paul Rosenthal, Maria Cristina Ferreira de Oliveira, Lars Linsen, and Rosane Minghim. 2011. "A Framework for Exploring Multidimensional Data with 3d Projections." In *Computer Graphics Forum*, 30:1111–20. 3. Wiley Online Library.

Qu, Huamin, Wing-Yi Chan, Anbang Xu, Kai-Lun Chung, Kai-Hon Lau, and Ping Guo. 2007. "Visual Analysis of the Air Pollution Problem in Hong Kong." *IEEE Transactions on Visualization and Computer Graphics* 13 (6): 1408–15.

Raidou, Renata Georgia, Martin Eisemann, Marcel Breeuwer, Elmar Eisemann, and Anna Vilanova. 2015. "Orientation-Enhanced Parallel Coordinate Plots." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 589–98.

Siirtola, Harri, and Kari-Jouko Räihä. 2006. "Interacting with Parallel Coordinates." *Interacting with Computers* 18 (6): 1278–1309.

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.

Talbot, Justin, John Gerth, and Pat Hanrahan. 2012. "An Empirical Model of Slope Ratio Comparisons." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2613–20.

Theus, Martin, and Simon Urbanek. 2008. *Interactive Graphics for Data Analysis: Principles and Examples*. CRC Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information (2nd Edition)*. USA: Graphics Press.

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

Wegman, Edward J, and Qiang Luo. 1997. "High Dimensional Clustering Using Parallel Coordinates and the Grand Tour." In *Classification and Knowledge Organization: Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation eV, University of Freiburg, March 6–8, 1996*, 93–101. Springer.

Wickham, Hadley. 2013. "Bin-Summarise-Smooth: A Framework for Visualising Large Data." *Had. Co. Nz, Tech. Rep.*

Wilkinson, Leland. 2012. *The Grammar of Graphics*. Springer.

Xue, Yanbing, and Milos Hauskrecht. 2017. "Efficient Learning of Classification Models from Soft-Label Information by Binning and Ranking." In *The Thirtieth International Flairs Conference*.

Yang, Vincent, Harrison Nguyen, Norman Matloff, and Yingkang Xie. 2017. "Top-Frequency Parallel Coordinates Plots." *arXiv Preprint arXiv:1709.00665*.

Yuan, Xiaoru, Peihong Guo, He Xiao, Hong Zhou, and Huamin Qu. 2009. "Scattering Points in Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 1001–8.

Zhao, Xin, and Arie Kaufman. 2012. "Structure Revealing Techniques Based on Parallel Coordinates Plot." *The Visual Computer* 28: 541–51.

Zhou, Hong, Xiaoru Yuan, Huamin Qu, Weiwei Cui, and Baoquan Chen. 2008. "Visual Clustering in Parallel Coordinates." In *Computer Graphics Forum*, 27:1047–54. 3. Wiley Online Library.