Lit Review for handling numerical ties in PCPs

Denise Bradford

Introduction

I think the thing to do in the introduction is describe the problem of both large n and large p data; then state that your focus will primarily be on ways to work with data that has $p \geq 4$ and $n \geq 50$ to enhance our ability to visually detect patterns across multiple dimensions. You can introduce PCPs as a way to track a single observation across multiple dimensions, and then you can narrow down your focus into the problem of ties and how they disrupt the ability to keep track of a single observation.

In modern data analysis, the complexity of datasets with large n (number of observations) and large p (number of variables) presents unique challenges. High-dimensional data, where $p \geq 4$ and $n \geq 50$, is increasingly common in genomics, finance, and social sciences domains. Analyzing such data often requires methods that enhance interpretability while preserving the intricate relationships within the data. However, the scale of observations and variables can obscure meaningful patterns and impede traditional visualization techniques.

Large p datasets are particularly challenging because visualizing relationships across multiple dimensions often leads to information overload, occlusion, and cognitive strain (Keim 2002). Visual representations struggle to maintain clarity when $p \geq 4$, as patterns become increasingly difficult to discern in higher-dimensional spaces (**alber2020?**). Similarly, large n datasets exacerbate these issues by introducing visual clutter, making it harder to track individual observations (Heer and Bostock 2010). These challenges require innovative visualization techniques that simplify complex data while preserving its structural integrity.

~~Our data-driven modern civilization [A bit trite...].svp now depends heavily on the ability to precisely evaluate and exploit [exploit?? Maybe a different word?].svp huge and complex datasets. A common way [Not sure PCPs are actually all that common, but they are a good solution].svp to visually show high-dimensional data is with parallel coordinate plots (PCPs).~~

Parallel Coordinate Plots (PCPs) have become a powerful way to examine data with more than one dimension. PCPs show each observation as a polyline on parallel axes, letting analysts follow specific data points in multiple directions (Inselberg 1985). By setting factors up as parallel vertical axes, PCPs make it easier to find patterns, trends, and outliers in large

datasets. 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. I don't know that we need this paragraph here, at the moment. Don't get rid of it entirely, though, it's not bad – just not quite on topic here

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 (Inselberg 1997). 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 1984 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 and 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. I don't know that we need this paragraph here, at the moment. Don't get rid of it entirely, though, it's not bad – just not quite on topic here

Parallel Coordinate Plots (PCPs)

Parallel coordinate plots (PCPs) leverage a projective coordinate system, instead of a 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 represented as a line that crosses each parallel axis representing a variable; clusters, then, appear as a group of lines which have similar paths.

I think it might be best to customize the ggpairs plot so that it's only a scatterplot matrix – that makes for a simpler comparison than the multiplot hybrid beast that's the default ggpairs plot. Scatterplot matrices are the most direct comparison to PCPs because they show Cartesian pairs of variables

Figure 1 is a generalized pairs plot, containing histograms, density plots, scatterplots, 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 provide univariate distribution information to supplement the bivariate scatterplots for numeric variables (bill length, bill depth, flipper length, and body mass). 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 the bivariate relationship between pairs of numerical variables. While some of the information is obscured due to overplotting, the overall relationship between variables is relatively clear and is reinforced by the numerical values displayed in the corresponding pair of variables across the diagonal. Overall, the matrix-style plot does an excellent job of showing the links within and between variables. The major downside to scatterplot matrices is that it is not possible to easily connect a point in one scatterplot to a corresponding point in another; the data representation makes it difficult to get a sense of the multivariate relationships beyond any combination of $p = 2$ variables.

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)

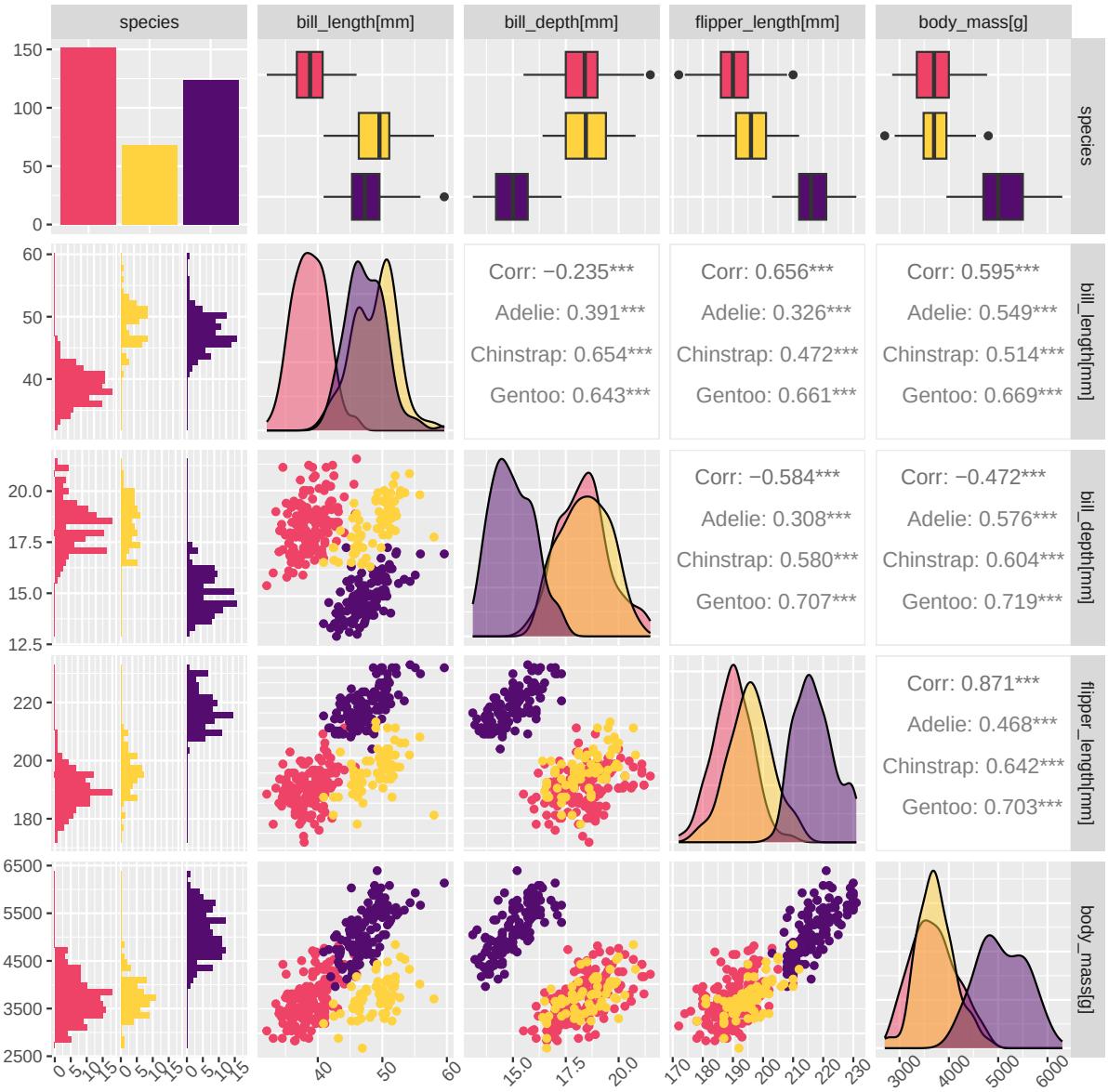

Figure 1: A generalized pairs plot

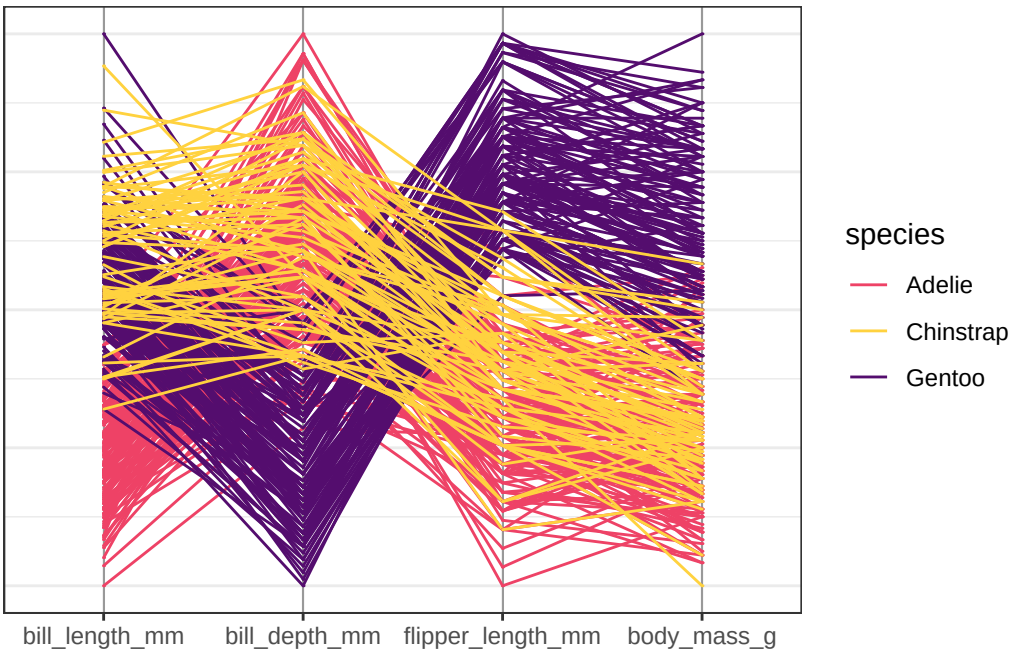

Figure 2: A parallel coordinate plot

tends to have lower values for these measurements. There is an overall negative relationship between flipper length and bill depth, indicated by the "X" shape of the lines, but within each species, the overall relationship between flipper length and bill depth is positive, as indicated by the largely parallel block of lines of each color. Even in projection space, we can see Simpson's paradox at work; the middle plot in the fourth row of Figure 1 shows the same basic information. While there is a significant amount of information obscured due to overplotting, it is easier to connect observations across $p > 2$ vertical axes, providing a greater ability to visualize more than two dimensions. At each vertical axis, the plot resembles a rug plot, which can provide some information about density, but is less direct than a continuous density plot, such as those shown on the diagonal of Figure 1. Cleveland (1993)

As originally proposed, PCPs could be difficult to interpret, in part because PCPs were initially defined only for numeric variables; extensions which treated categorical variables as numeric suffered from overplotting , as different lines converge on a single point and then diverge again, destroying the ability to trace a single observation through the categorical-turned-numerical axis. The `ggpcp` package (VanderPlas et al. 2023) introduced a new way to handle categorical variables, dividing the axis up into "boxes" and ordering observations within those boxes; for relatively small n, this preserves the ability to follow single observations across the plot, and for larger n, a series of lines moving together converge to form an approximate hammock plot . A demonstration in Figure 3 replicates Figure 2 with the addition of a categorical species axis on each side of the plot. In the InfoVis community, other modifications to parallel coordinate plots have been proposed: smoothed lines, density-based PCPs(J. Heinrich and Weiskopf 2009),

bundling of similar points, and other modifications such as interactivity (Johansson and Forsell 2015) may support identification of clusters and outliers in multidimensional space.

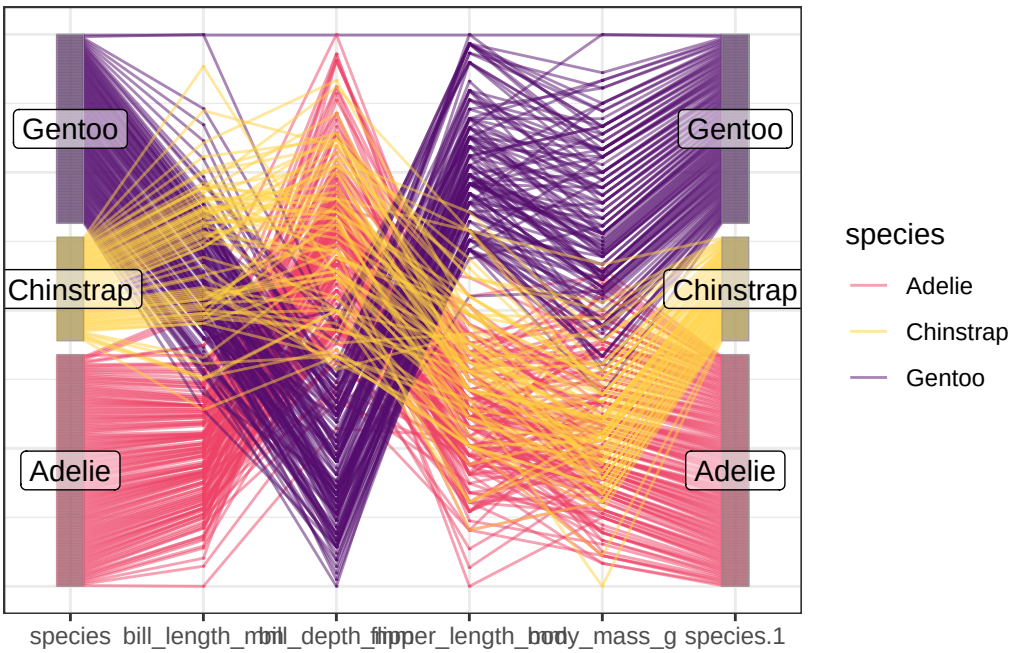

Figure 3: A generalized parallel coordinate plot, with species on the left and right of the plot; observations are ordered on the right side based on the value of body_mass_g, and on the left side based on the value of bill_length_mm. Translucent lines reduce the impact of overplotting and allow perception of the Adelie lines (which are plotted first) even as the Gentoo and Chinstrap lines are plotted on top. With this treatment, we can see that the strong positive relationship between bill depth and flipper length in Gentoo penguins is much less pronounced in Adelie and Chinstrap penguins; there are many more line crossings in both species, which suggests a more moderate relationship.

Why?

Plots using Cartesian coordinates are straightforward to understand because they are commonly encountered and taught in grade school. Unfortunately, plots using Cartesian coordinates, such as scatterplots, are limited to two variables displayed using spatial dimensions, which are perceived more accurately than other aesthetic mappings such as color and shape (Cleveland and McGill 1984). Projections of three-dimensional scatterplots can be created, but these charts can be difficult to read and interpret; interactive 3D scatterplots still suffer from the loss of information inherent to 2D projection on a screen, but allow for some sense of the full shape of the data. Data sets with more than 3 numerical dimensions are

extremely common, but cannot be easily shown using Cartesian coordinates; analysts must resort to strategies like tours (Wickham et al. 2011), dimension reduction (van der Maaten and Hinton 2008) XXX PCA XXX, or plotting bivariate relationships between variables in order to visualize these data sets. 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.

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. repeatedly?? Not sure what you mean by this PCPs are great for drawing attention to trends, clusters, and outliers and efficiently show many variables in a small area. are great for is very informal language – can you rephrase this to be more formal? e.g. PCPs effectively show bivariate relationships, clusters of similar data points across multiple variables, and outliers, as shown in ...., and then provide a figure showing how you can see those qualities within a PCP? However, as with their Cartesian equivalents, PCPs are vulnerable to overplotting, becoming difficult to read and interpret when N is large. In addition, PCPs are a much less familiar form of visualization than scatterplots; there is an initial adjustment period required in order to understand which features of a PCP correspond to familiar features of a scatterplot, as shown in **?@fig-features**. This lack of familiarity may be the biggest drawback of PCPs; however, there are ... studies which explore this issue. XXX are there studies that explore how quickly people acclimate to PCPs? If not, we should consider running one!

Modifications to Parallel Coordinate Plots

Since Inselberg (1985) 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. In this section, we examine some of the modifications proposed to enhance PCPs after their original introduction.

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 (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 (2015) explored alternative segmentation methods to minimize visual clutter, enabling effective handling of multiple categorical values on the same axis.

CPCPs adjust by organizing categorical axes more clearly or using symbols and colors to represent categories for easier interpretation. Is this actually a CPCP development, or did it just develop alongside the idea of CPCPs? It seems like it would apply to all PCP-variants. Wegman et al. (1990) Use the citation – (**XXX?**) suggested using symbols and color codes to represent categories, improving how categorical data is understood in parallel coordinates. Holten and van Wijk (2009) Use the citation – (**XXX?**) introduced color gradients for categories to facilitate distinguishing between multiple categorical values, especially in the analysis of intricate datasets. LeBlanc et al. (1990) Use the citation – (**XXX?**) highlighted the significance of arranging axes in CPCPs, positioning categories with strong relationships closer together to assist users in making meaningful comparisons.

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. Wegman and Luo (1997) provide a comprehensive discussion of the GPCP idea. This doesn't really add to the discussion - and the next sentence does not follow from this one. Rewrite to make your point clearer, and cite ideas, not people. This adjustment allows the visualization to display various data connections and handle skewed data distributions.

Hammock Plots: Categorical Variants of PCPs

Hammock plots are similar to parallel coordinate plots, but were designed to show categorical data. Symanzik, Friendly, and Onder (2018) showed the Titanic dataset using hammock plots,

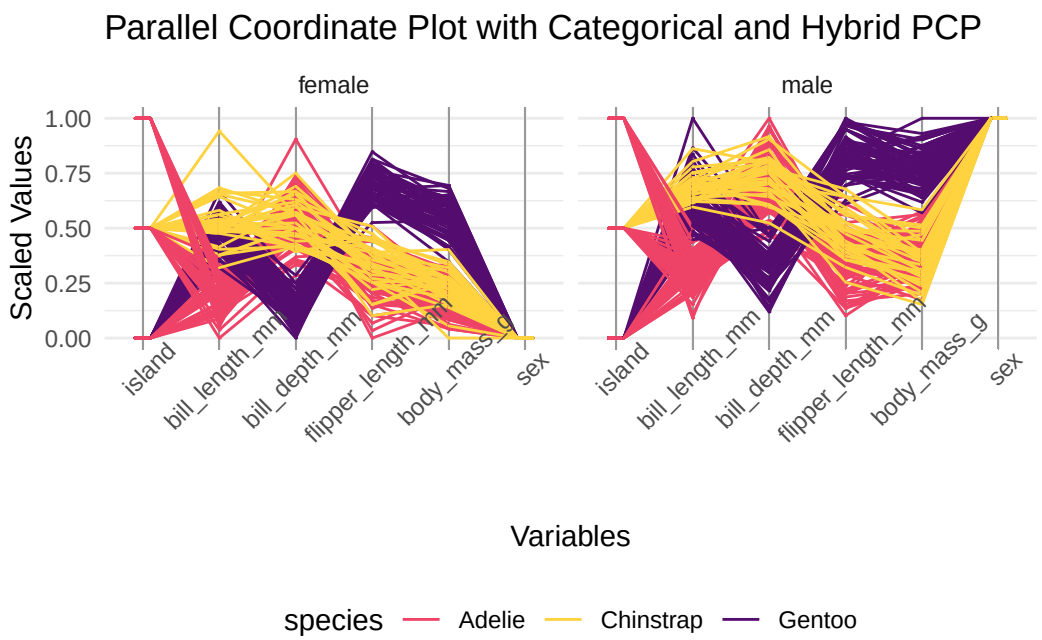

Figure 4: Categorical and Hybrid in PCP

arguing that "Hammock plots allow for line alignment representing similar or identical values." This makes understanding the data in parallel coordinate space easier and gives you a better sense of the different data types and their groups. They do say, though, that Hammock plots "may add visual clutter in cases lacking categorical data distinctions," which is meant to stress the problems that might arise when they are used with continuous variables.

Kavvadias et al. (1996) use a decomposition method called "Hammock-on-ears" to work on bigger graph-theoretic problems, with the goal of enhancing the ability to find ties in large datasets. This shows that Hammock plots can handle large amounts of complex data. However, they warn that "the method may require significant computational resources," which means it might not work for simpler or smaller-scale tasks. Actually, it means the opposite - if you don't have the computer power, you can't use the method for big datasets, but it probably works for smaller ones

Pilhöfer, Gribov, and Unwin (2012) show that Hammock plots "manage clusters and resolve data ties" well, especially when comparing data clusters. Show how? Is there an experiment? Or is this just a claim based on intuition? This is especially helpful in datasets that are hard to see Why are they hard to see? because Hammock plots make it easier to separate the data. Still, Pilhöfer et al. points out a problem: "The focus on clustering may introduce unnecessary complexity in dispersed data." They suggest that Hammock plots work best for datasets with clear clusters Do they provide a metric for identifying such datasets? How clear is clear?.

Ge and Hofmann (2020) improved the framework for visualizing data by adding Hammock plots to a grammar-based system. What? No, that's the first ggpcp paper draft – it didn't

talk about Hammock plots. You want to look into the research Heike did with Vendettuoli circa 2012-2015 for hammock plots and variations This makes them better suited for data that has both category and numerical relationships. They say this grammar-based adaptation "retains the established behaviors of parallel coordinates" while giving you more options for dealing with edge cases like categorical ties. This makes Hammock plots more flexible for use with large datasets. Please don't conflate PCPs and hammock plots – they're similar, but not the same thing Also, please don't cite arxiv papers when a peer-reviewed version exists – e.g. Penguins go Parallel

VanderPlas et al. (2023) investigate more deeply how useful Hammock plots are for dealing with numerical ties in categorical datasets. They show that "Hammock plots reduce overlap in tied data," which makes it easier to understand parallel coordinate plots. I can't find this quote anywhere in the paper – are you actually copying quotes from the paper when you quote things??? Please go back and review all quotes and make sure you can put a page number that's accurate – otherwise, remove the quotes and rephrase – but make sure you're accurate to the meaning!!. However, they also admit that these plots are "less effective for continuous data without ties or when categorical values are minimal," This phrase does not exist in the paper! suggesting that Hammock plots offer limited benefits in uniformly distributed datasets.

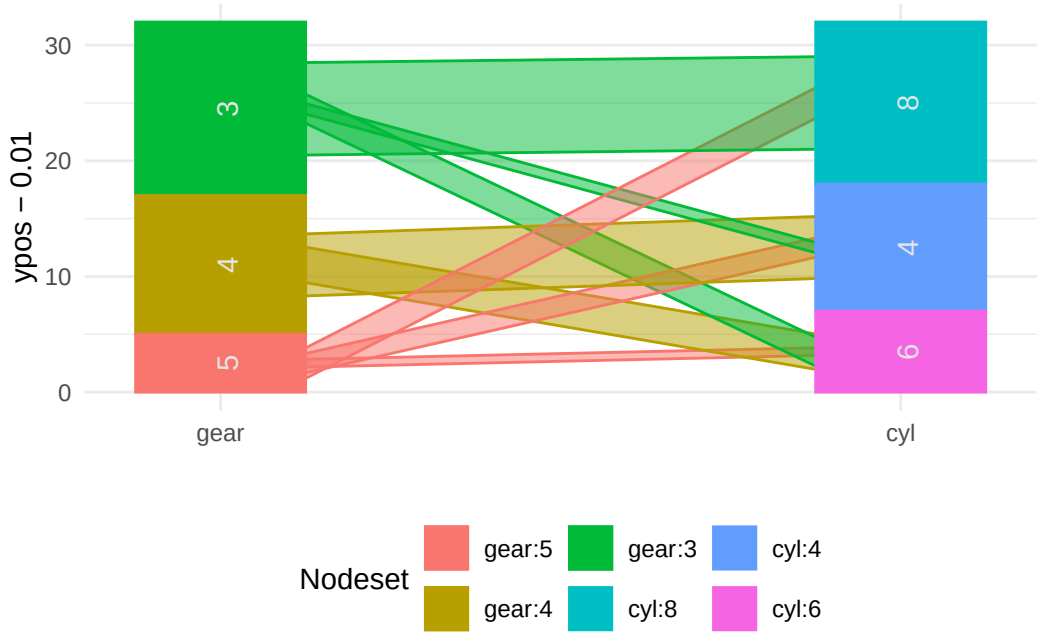

Figure 5: Hammock

Figure 5 is a Hammock plot that shows the relationship between two categorical variables, gear and cylinders (cyl). Each color-coded band represents a different category within these variables. The plot provides frequency but lacks precise counts or numeric values. Overplotting is evident, making it difficult to trace individual pathways. The plot shows the general distri-

bution and relationships between gear and cylinder categories, but lacks numeric relationships or exact proportions.

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 (2009) note that "bundling offers a compelling solution to mitigate the chaos often found in high-dimensional visualizations".

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).

In Figure 6, the PCP modification demonstrates adding 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.

Rida E. Moustafa (2011) 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.

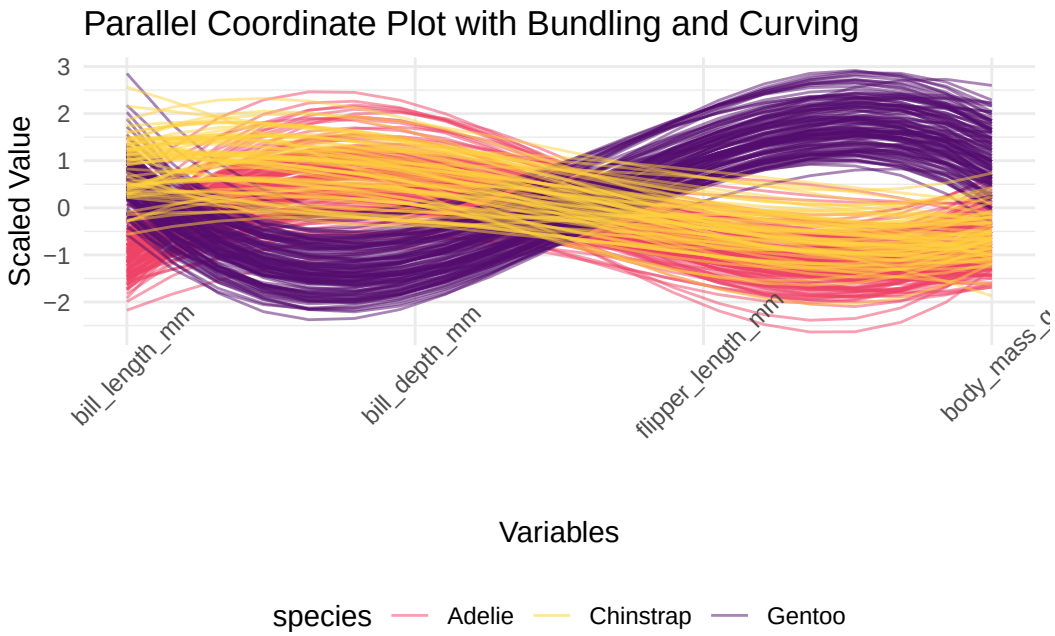

Figure 6: A Bundling and Curving Parallel Coordinate Plot

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. Rida E. Moustafa (2011) 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.

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).

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."

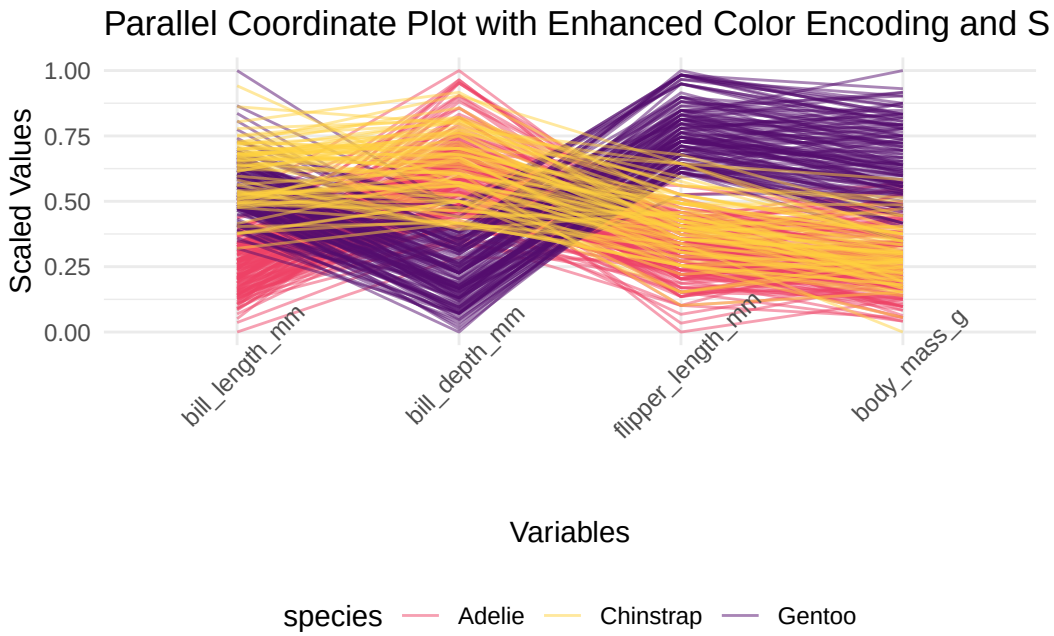

Figure 7: Enhanced Color Encoding and Shading in PCP

In Figure 7, 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 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.

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 applied PCA within PCPs, enabling users to represent key data features while minimizing less relevant dimensions (Wegman and Luo 1997). Guo et al. 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. 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. caution that, "while

beneficial, these methods might lead to information loss, potentially masking relationships only visible in higher dimensions," (Guo, Xiao, and Yuan 2012).

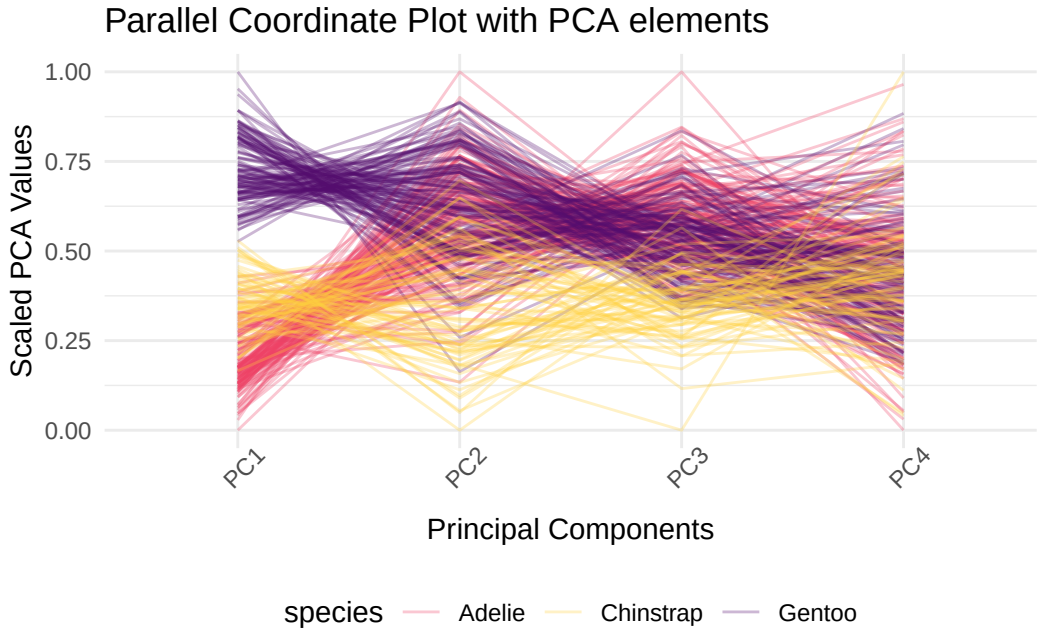

Figure 8: A PCA in Parallel Coordinate Plot

In Figure 8, 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 implied through color-coded clusters and trends, suggesting some separation between species along specific principal components.

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).

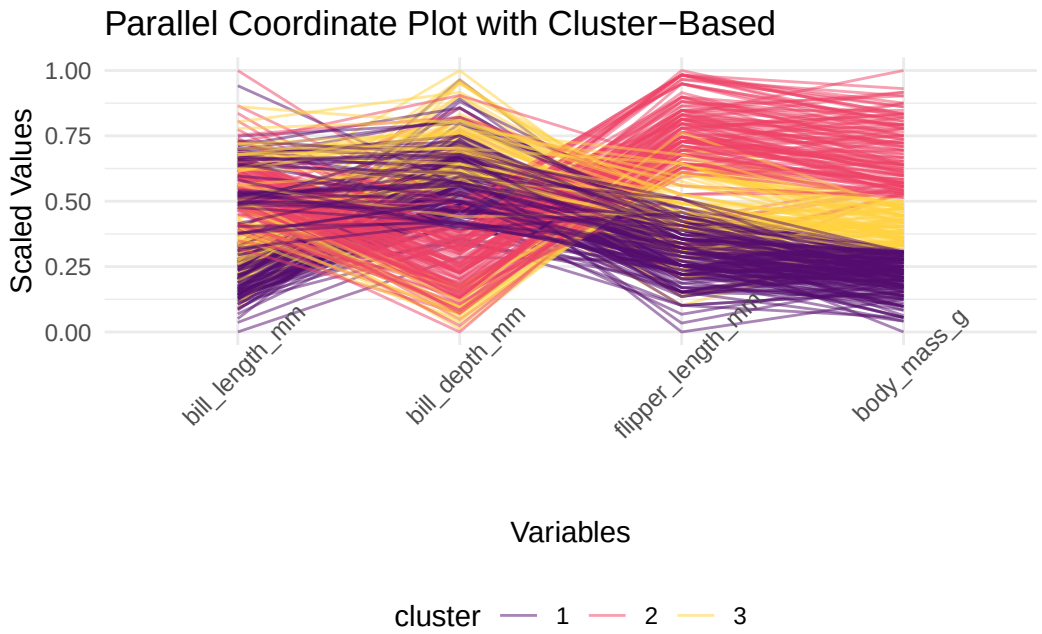

Figure 9: Cluster-Based PCP

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, Berchtold, and Keim 1998) Is there a user study showing that this makes PCPs easier to read?

Automatic, heuristic PCP layout algorithms may produce results that do not align with user-defined analysis needs (Johansson et al. 2005). Instead, Johansson et al. (2005) optimized the axis positioning algorithm to maximize the space between overlapping lines, improving data distinction. What does data distinction mean here? Are there experimental studies showing this?

Hao et al. (2007) implemented a randomized layout approach that provides an efficient layout for PCPs, beneficial for large, high-dimensional data.

You need to show these different options and comment on the differences! Provide more detail about whether experimental evidence supports the approach. How does the approach in Ankerst, Berchtold, and Keim (1998) compare to the ggpcp ordering approach to minimize line crossings in one direction or another for categorical variables? I suspect that these algorithms change axis positioning, where ggpcp changes ordering within a category – but you need to explicitly make that comparison.

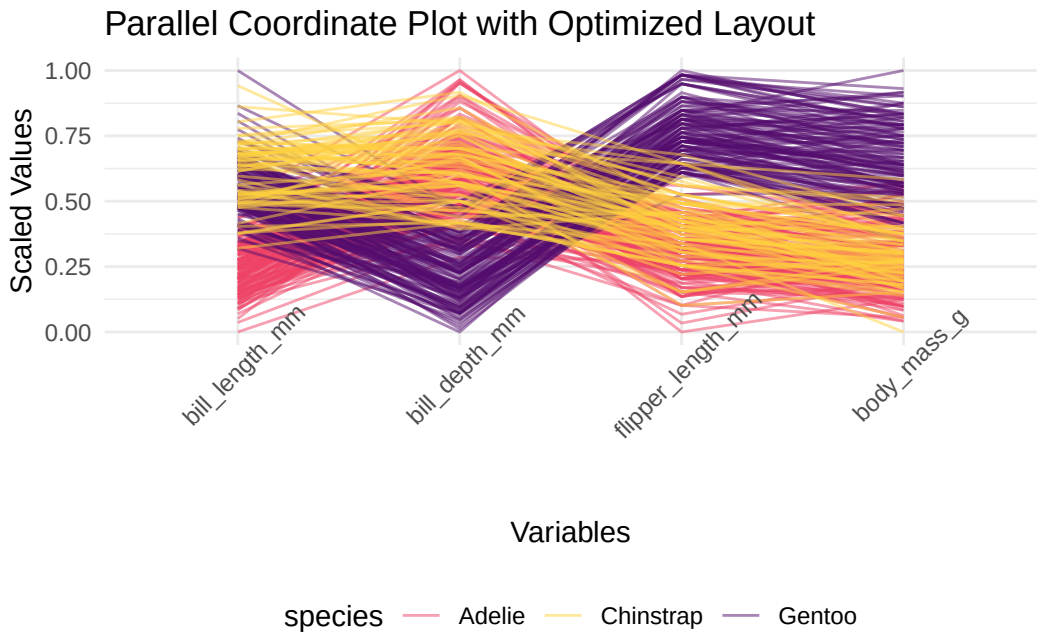

Figure 10: Algorithm Optimized Layout in PCP. There needs to be a lot more information here – what do you see from the picture?

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 live parallel coordinates plot on Plotly's ggplot2 page has many dynamic features that make looking at many different data types easy. Users can interact with the plot by moving their mouse across parallel axes to filter and separate specific data ranges. This lets them analyze only certain factors. When you move your mouse over a line, a tooltip appears with all of that line's numbers for each dimension. This makes it easier to understand large datasets. You can change the order of the axes by dragging them horizontally. This lets users adjust the dimensions to find hidden patterns and connections more easily.

Additionally, brush choices change the visualization in real-time, drawing more attention to relevant data paths and lessening the importance of others. This makes it easier to spot trends and outliers. These interactive parts work together to create an easy-to-use and adaptable way to see and analyze high-dimensional data in a parallel coordinates environment, ("Parallel Coordinates Plot in Ggplot2," n.d.).

Interactive Parallel Coordinate Plot

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. Citation needed 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 Citation needed 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.

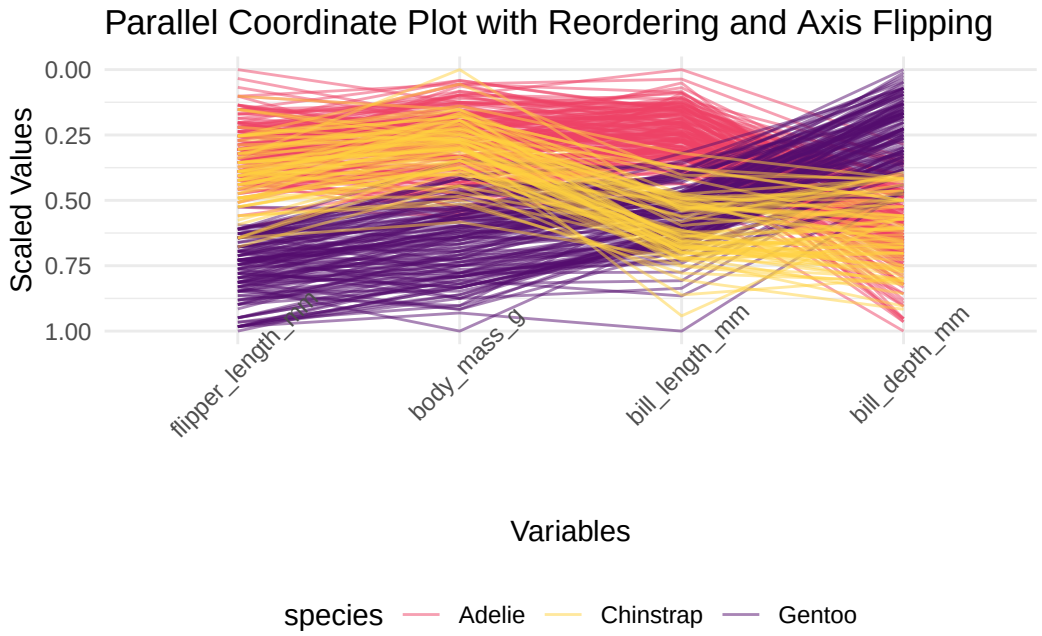

Figure 11: Reordering and Axis Flipping in PCP

In Figure 11, 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.

Handling Numerical Ties

Getting rid of ties in statistical data analysis has been hard for a long time for experts who want to make visualizations more clear. Previously, statistical methods focused on rank-based approaches and binary data handling. They often needed help with tied observations, which threw off distributions or made them more complex. Earlier methods for dealing with ties depended on changing ranking systems or adding rules to break ties, which could make results less accurate or easier to understand. As graphics data analysis grew, pioneers like Tukey, Chambers, and Cleveland created new ways to deal with ties visually. There are now several ways to deal with ties, such as slowly increasing the space between values. This makes the visualization more accurate and easier to understand without changing the structure of the data itself.

Tufte's work, particularly in "The Visual Display of Quantitative Information", is grounded in empirical evidence that shows how visual noise—any element that complicates or crowds the visual field—can impede the interpretation of data. He advocates for "graphical excellence," a design philosophy that presents data most clearly and efficiently. According to Tufte, separating tied values can reduce overlapping lines and decrease visual noise, making it easier for users to see patterns and accurately compare variables.

In complicated graphs like parallel coordinate plots, tied numbers often cause lines to overlap and hide other data points. Tufte calls this "chartjunk." He says that even minor cuts in visual noise can make it much easier for people to process information quickly and correctly. For instance, if a vast dataset has many similar or similar values, separating the tied values can help the viewer see trends and relationships without having to think too hard (Tufte 2001).

Random Jittering of Data Points

Random jittering of data points is a standard method that adds a small random value to each tied observation. This is especially helpful in scatter plots and dot plots to keep them from being too dense. This approach, discussed in "Graphical Methods for Data Analysis" by Chambers et al., is easy to use and keeps the overall distribution, making things clearer without changing the data (Chambers 1983). One problem with random jittering is that it can hide exact numbers, making it hard to understand if the jittering distance is too big. Experiments

show that jittering makes visualizations much clearer when there are small overlaps, especially when used rarely.

In random jittering, we add a small random value (usually drawn from a uniform or normal distribution) to each tied observation to differentiate them.

For a set of tied values $x_1, x_2, ..., x_n$, we define:

$$x_i' = x_i + \epsilon_i$$

where:

- x_i' is the jittered value of x_i,
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$,
- σ is the standard deviation for the normal jitter.

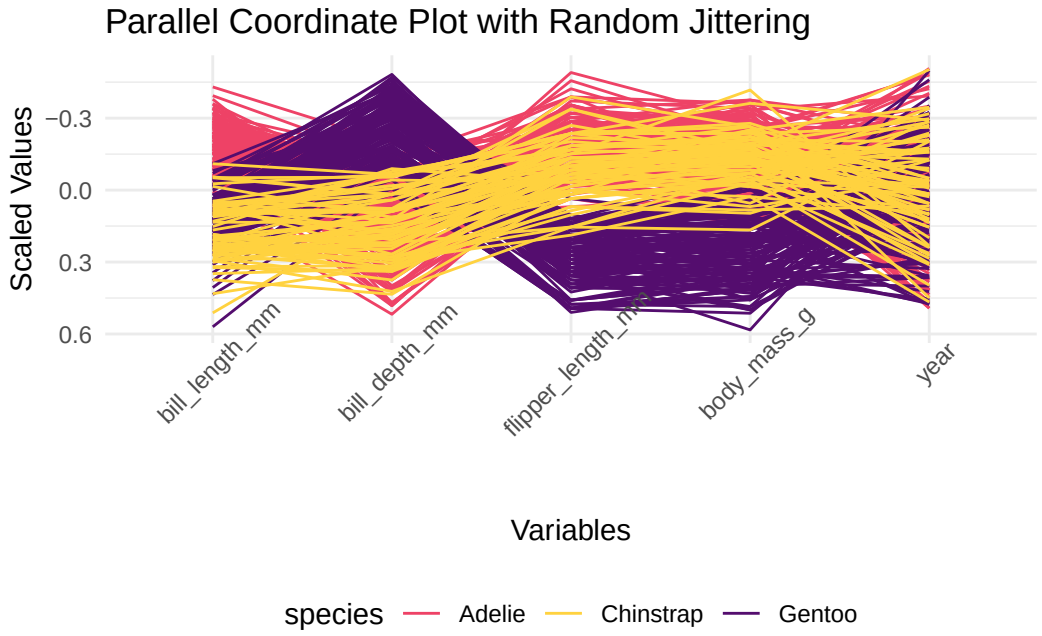

Figure 12: Random Jittering of Data Points in PCP

Rank Jittering (Rank Adjustment)

Rank jittering is another useful method for non-parametric analyses. It changes the ranks of tied values instead of the real data values. This makes it easier to see the ranks in non-parametric tests like the Wilcoxon signed-rank test. Conover's "Practical Nonparametric Statistics" shows that rank jittering keeps the ordinal form of data while not changing the

distribution much (Conover 1999). The best thing about rank jittering is that it keeps rank-based readings while making ties less noticeable. However, the method can be hard to run on computers when dealing with big datasets, and making too many rank changes can accidentally change the test significance levels. For these reasons, it is important to be careful when using rank jittering.

In rank jittering, we adjust the ranks of tied observations by adding a small random or systematic jitter to each tied rank. This is usually applied to data that will undergo rank-based testing.

for tied ranks $R_1, R_2, ..., R_n$:

$$R_i' = R_i i + \delta_i$$

where:

- R_i' is the jittered rank of R_i,
- $\delta_i \sim \mathcal{N}(0, \sigma^2_{rank})$
- σ_{rank} are small values ensuring that the adjustment is minor.

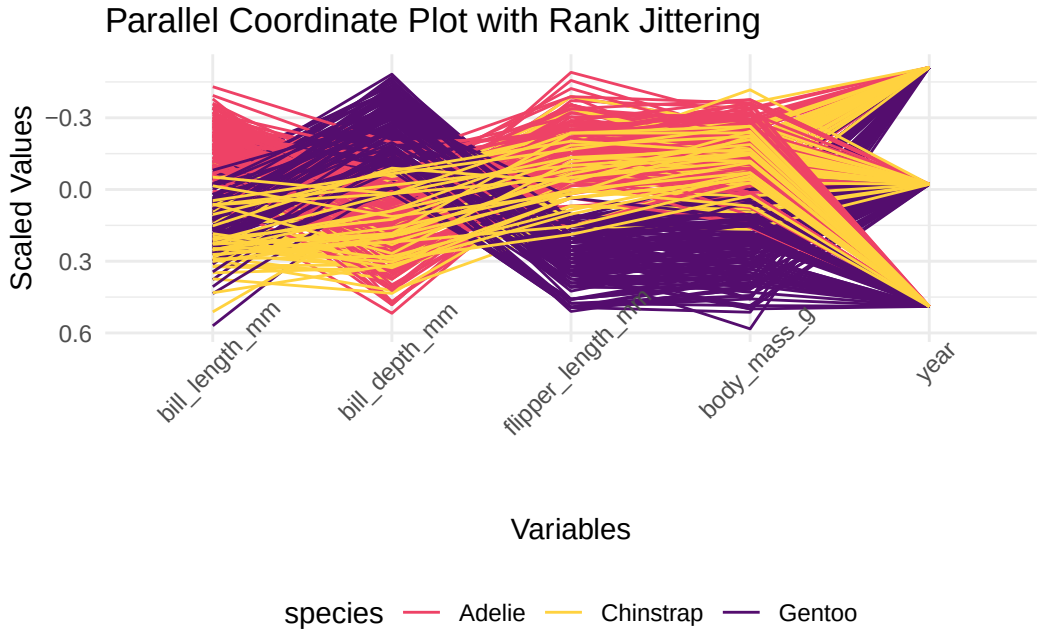

Figure 13: Rank Jittering of Data Points in PCP

Deterministic Jittering (Fixed Perturbation)

When you need to be clear, deterministic jitter gives you a structured way to do it by adding a small fixed value (epsilon) to each tied observation instead of depending on chance. This method explained in more detail in Tukey's "Exploratory Data Analysis", is useful when random noise is not wanted because it ensures that each observation is moved in a way that can be predicted and repeated (David and Tukey 1977). Deterministic jitter is easier to understand than random jitter, but it needs to be done carefully so that it doesn't introduce fake patterns, especially when scaling is involved. Deterministic jittering is a good choice for random noise that doesn't affect the integrity of the distribution when used carefully.

In deterministic jittering, a small, fixed value ϵ is added to each tied value to systematically spread them apart without introducing randomness.

For tied values $x_1, x_2, ..., x_n$, we use:

$$x_i' = x_i + i\epsilon$$

where:

- x_i' is the adjusted value of x_i,
- i is the indeex of the tied observation (e.g., $i = 1, 2, ..., n$),
- ϵ is a small fixed distance chosen based on the data scale.

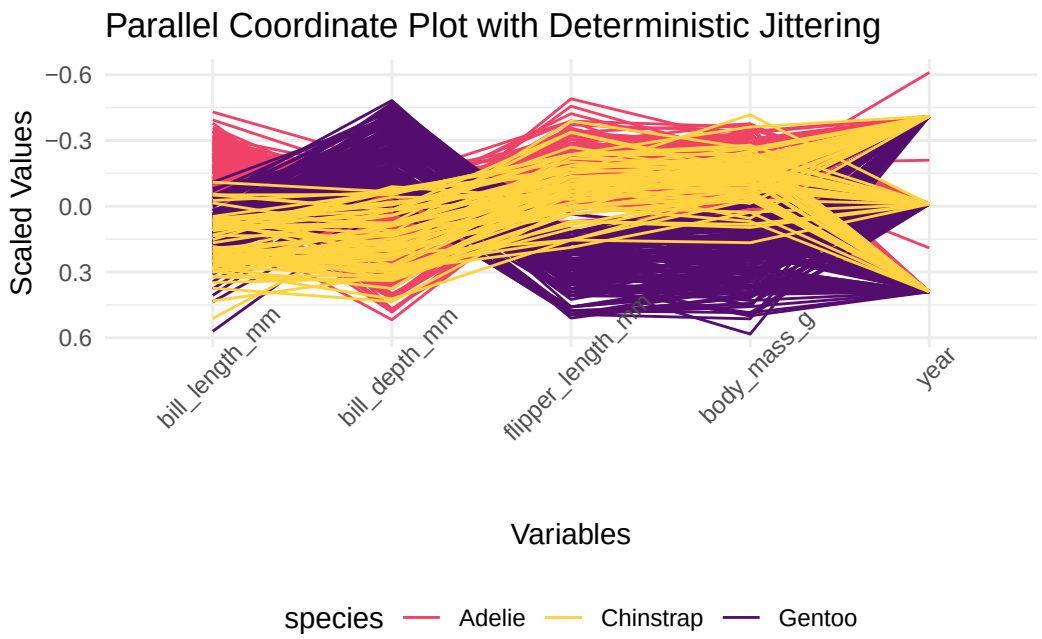

Figure 14: Deterministic Jittering of Data Points in PCP

Mean or Median Splitting

One more popular method is mean or median splitting, which gives values tied to their mean or median values. This method works especially well for symmetric distributions and is perfect for images that need to show things centrally, like box plots. Wilcox's "Modern Statistics for the Social and Behavioral Sciences" shows that splitting the mean or median gives a fair view, especially for boxplot data (Wilcox 2017). This method works best with datasets with few ties because it can hide the range and variability within tied numbers if used too much.

In mean or median splitting, we replace each tied value with the mean or median of the tied group. This technique centers tied observations on their central tendency.

For set of tied values $x_1, x_2, ..., x_n$ let:

$$x_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } x_{median} = median(x_1, x_2, ..., x_n)$$

Then:

$$x_i' = x_{mean} \text{ or } x_i' = x_{median}$$

This replaces each tied value x_i with the computed mean or median, clustering them at a central point.

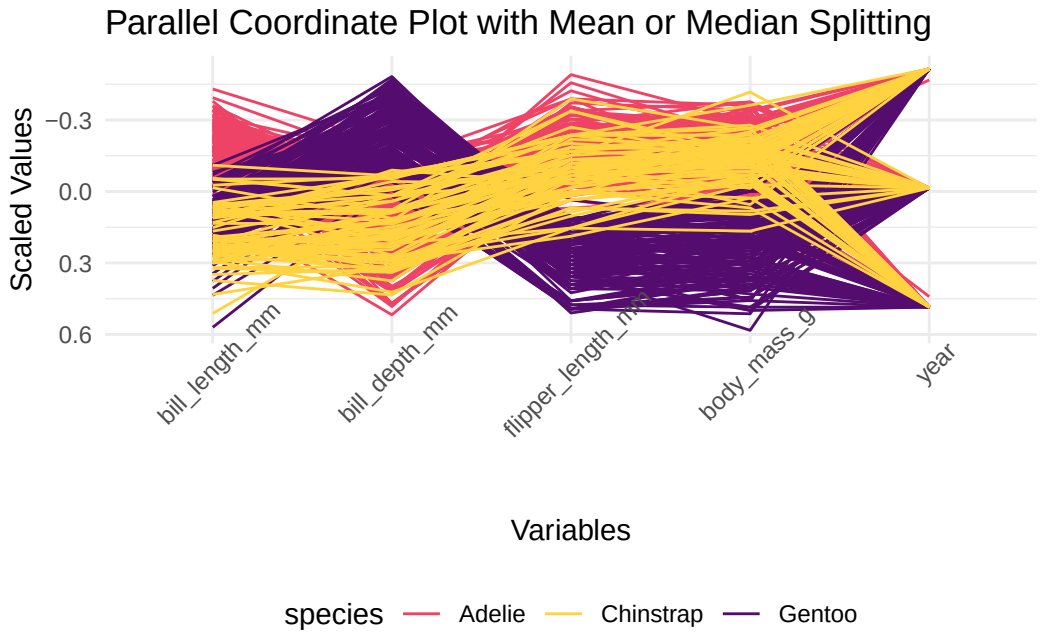

Figure 15: Mean or Median Splitting of Data Points in PCP

Kernel Density Estimation (KDE) with Bandwidth Adjustment

Kernel density estimation (KDE) with bandwidth adjustment is a powerful method for density visualizations. It spreads out overlapping observations by giving each tied value a slightly different weight. This method works especially well in density plots because it evens data distributions without changing the original numbers. Silverman's work in "Density Estimation for Statistics and Data Analysis" shows how KDE with changed bandwidths correctly shows multimodal distributions, reducing errors caused by tied values. KDE works, but you must be careful when choosing the bandwidth because making the wrong choice can lead to misleading visual representations, especially when files are very concentrated (Silverman 2018).

For KDE with bandwidth adjustment, each data point is used to estimate a probability density function (PDF) by smoothing over a specified bandwidth h, which can be adjusted to account for ties.

The KDE for a set of observations $x_1, x_2, ..., x_n$ is:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

where:

- $\hat{f}(x)$ is the estimiated density at x,
- $K(\cdot)$ is a kernel function
- h is the bandwidth, chosen to spread tied values slightly by selecting a larger h for clusters of tied observations.

The choice of bandwidth h controls the amount of smoothing, which helps visually differentiate tied values.

Each way to show data with ties has unique benefits depending on the type of display and the type of data. Random and predictable jittering, for example, works well in scatterplots because they stop the plot from getting too big and show hidden patterns. KDE and rank-based methods, on the other hand, work well in density plots and ordinal visualizations. Researchers can improve visual clarity by using only some of these methods, ensuring the representation is correct and easy to understand. However, when parallel coordinate plots are used, where multiple directions are plotted simultaneously, the old ways of dealing with ties only sometimes work. In this case, adding the distance between values tied across multiple axes could show patterns and connections that are hard to see because the lines overlap. By creating a new method for parallel coordinate plots that adds space between ties in a planned way, we could see data trends and links across dimensions more clearly while keeping the dataset's integrity. This new idea would make it easier to understand complex, multidimensional data.

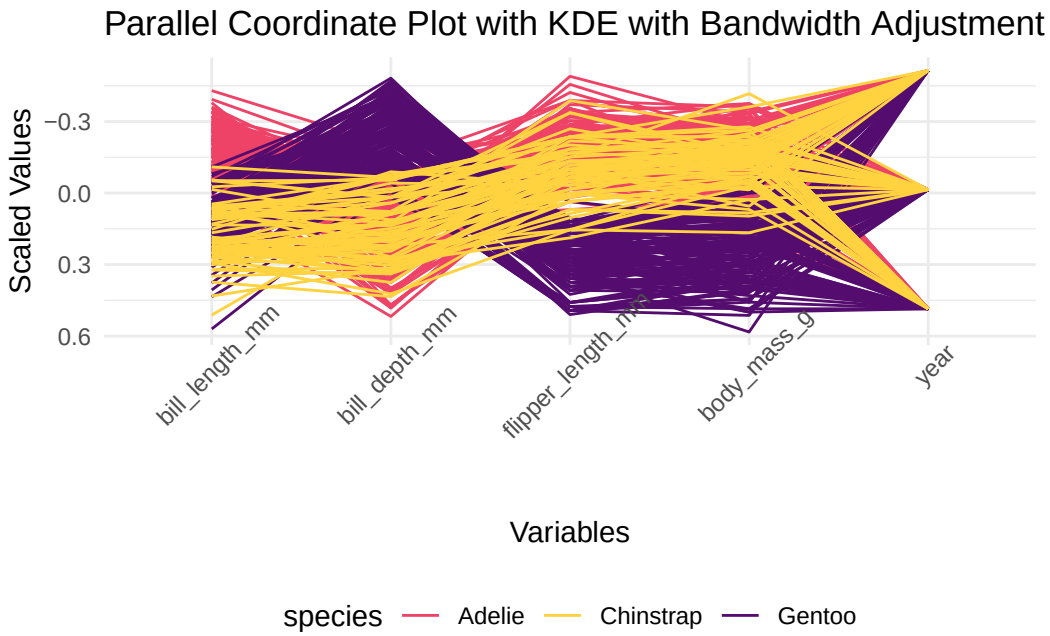

Figure 16: KDE with Bandwidth Adjustment of Data Points in PCP

Handling Ties in PCPs

The idea of adding distance to visual ties has changed over time as methods for visualizing data have improved to find a balance between detail and clarity in large datasets. In the early days of visualization, especially in network analysis and multidimensional scaling, closeness and alignment were the main ways to show patterns and ties between things. However, as data grew, visual clutter became a big problem. This was especially true in parallel coordinates plots (PCPs), where lines and intersections that overlapped made it hard to see the underlying data structure. To fix this problem, researchers started looking into ways to give representational parts of these graphs more "visual distance." They did this using methods like line bundling and changing the spacing between lines to avoid overplotting. This shows how separating visual elements can help with cognitive grouping without breaking up the continuity of the data. In PCPs, early studies showed that visual distance could make intersections clearer and improve the overall shape of individual paths. This theory has shaped later developments, showing visual distance as a spatial tool and a key cognitive help in data interpretation needed to get around today's more complex visualizations.

Axis Reordering

As mentioned in the section on modifications, axis reordering changes the order of dimensions to reduce the number of times lines cross between variables in PCPs, a type of data display.

This method makes clusters easier to see by lining up visually linked variables, cutting down on clutter, and making patterns stand out more without overlapping. One benefit is that it works well for datasets where certain factors are strongly related, making clusters and trends easier to read without changing the data. Changing the variable order for clarity rather than data continuity could make it hard to tell if the data is in order. Blumenschein et al. say that reordering based on correlation minimizes unnecessary line intersections, allowing natural clusters to emerge. This shows how important changing the arrangement can be for seeing clearly. Studies in this area have shown that changing the order of dimensions makes it much easier to find patterns in datasets where dimensions are linked. However, it only works well on datasets with dimensions tied together (Blumenschein et al. 2020).

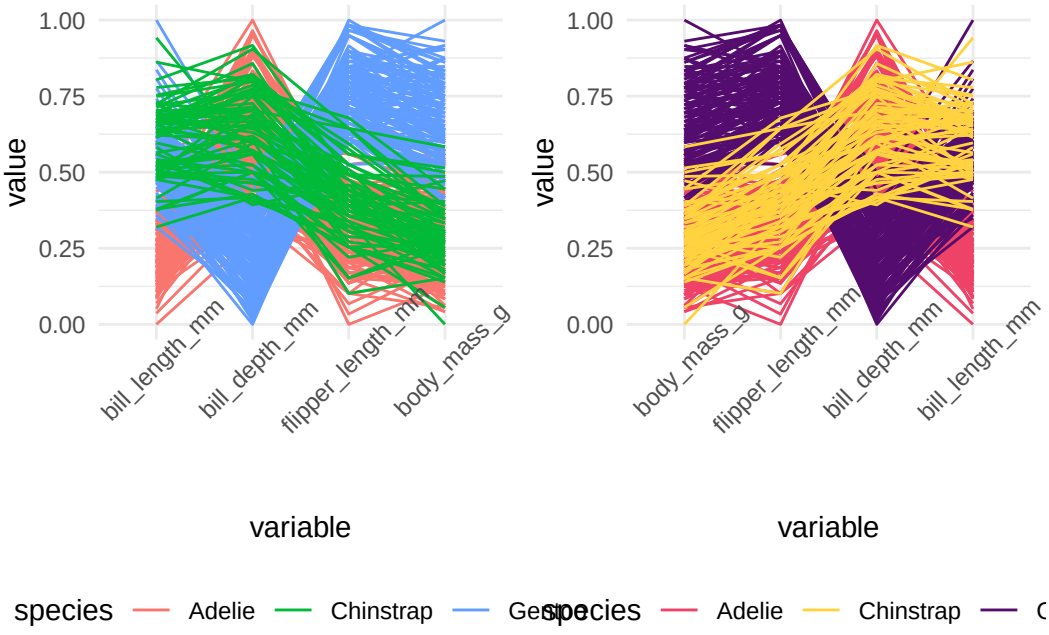

Figure 17: PCP Without Axis Reordering

Line Bundling

By putting similar lines into "bundles," line bundling cuts down on unnecessary crossings and clutter. This method works well in dense PCPs, where many lines cross and make it hard to see individual patterns. Heinrich et al. say that "bundling lines based on value similarity reduces cognitive strain and enhances visual coherence in complex datasets." This shows how bundling can make complex data easier to understand without losing larger patterns. Experiments have shown that bundling makes large datasets easier to read, which makes it easier to find general trends even when the data is heavily packed (Julian Heinrich et al. 2011). It makes things easier to understand by collecting similar data lines; this works especially well for numerical

ties where values are close but not exactly the same. Bundling, on the other hand, can hide individual lines within a group, making it hard to tell one data point from another.

Quantization and Aggregation

Quantization, also known as Quantized Generalized Parallel Coordinate Plots or QGPCP, groups close-together values into bins. This makes it easier to see patterns in tightly packed data. This method cuts down on the number of lines that cross each other, which makes it great for big datasets. Moustafa states, "By quantizing close data points, we can minimize overplotting and improve discernment of overall trends." This shows that removing some details can make data easier to read. Quantization can keep important patterns while lowering noise, as tests on real-world data show. However, it may hide small but important differences in the data (Rida EA Moustafa 2009). It makes the image easier to understand by removing small details. This is helpful for high-dimensional datasets with a lot of line overlap. However, detail can be lost during aggregation, making this method less useful for situations where exact data representation is needed.

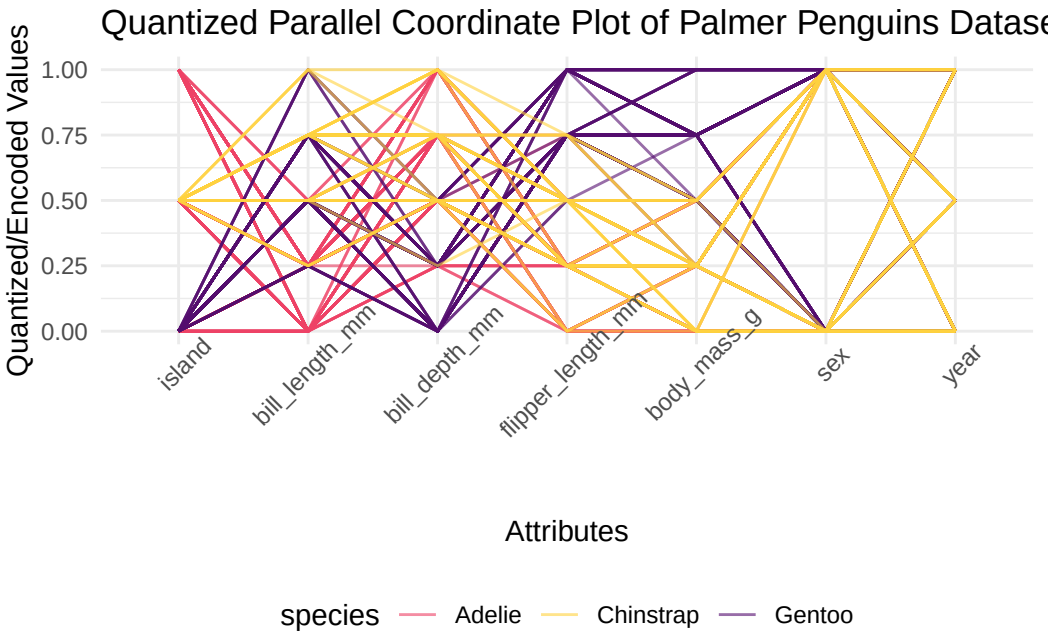

Figure 18: Quantized Generalized Parallel Coordinate Plots

In Figure 18, the overlapping of the Quantized Parallel Coordinate Plot makes it harder to find specific individuals. Quantization makes the visualization even easier by putting numbers into groups. However, this needs to improve some of the accuracy of the original data. The plot can help you see big patterns, like how some species may follow similar paths across many traits, but it doesn't show exactly how they are related or where they live. Using jittering or

interactive plots to cut down on overplotting could make things clearer and make figuring out how factors relate easier.

Color-Coding and Transparency

By changing the color and transparency of lines, you can tell them apart based on density, categories, or values. This makes it easier to tell the difference between PCP lines that meet. Firat et al. say, "Line brushing and transparency adjustments improve the gestalt of the data clusters." This means that changes in how we see things help us group data cognitively without changing the structure. Firat et al. say that user tests have shown that PCPs with color-coded and clear lines are easier to understand and recognize patterns, especially for differentiating categorical data (Firat, Swallow, and Laramee 2023). This tool can be used to handle both numerical and categorical ties. The color and transparency options make it easy to visually separate clusters without changing where the lines are placed. However, this method might only work for plots with little information because even changes in color can't fully fix heavy overplotting.

Edge Bundling for Categorical Data

Edge bundling, which comes from network visualization, groups paths for category values in PCPs. This method works best with categorical data because it lets you visually break up ties without changing other factors. Palmas et al. state that "bundling lines with similar orientations prevents overplotting, allowing clearer identification of dominant trends within complex data." This means bundling can help find trends even in spaces with many dimensions. Bundling made finding categorical clusters in complex PCPs easier and decreased users' work (Palmas et al. 2014). Edge bundling is an excellent way to clear up clutter and make categorical groups easier to see. It lets users see categorical differences without messing up numerical relationships. On the other hand, each bundle may hide small details, making them less useful for datasets that need a lot of precision.

Stacked Histograms and Density Plots

This change adds histograms or density plots on top of PCPs to show how category data is distributed in terms of frequency. This lets users see how distributions change without adding extra line clutter. According to Bok et al., adding histogram overlays gives categorical context, making variable distributions easier to see without affecting line continuity. This means that frequency-based cues can help you understand categorical distributions without adding more visual clutter. Clear frequency-based differentiation in categorical data without too much line overlap improves the visual understanding of distributions when used with PCPs. Real-world category dataset tests showed that histogram overlays made it easier to differentiate between data sets and see what they meant (Bok, Kim, and Seo 2020). Adding histograms makes

things look more complicated, which could be too much for some people if used with other PCP changes

Perceptual Cues (Line Thickness, Texture, etc.)

By changing the lines' width, texture, or other visual properties, you can tell them apart from overlapping lines. This helps people see patterns and tell values apart without separating them physically. According to Chang et al., "Perceptual variations in lines not only differentiate data but also enhance recognition of broader patterns." These cues help with cognitive grouping and pattern differentiation, even in complicated datasets. They work well in areas with a lot of visual information because they let users see trends without changing the structure or order of the data. They are helpful for both numerical and categorical ties. Studies with users showed that changes in how people saw things made finding patterns and trends in large, complex datasets easier. This made it easier for PCPs to look around (Chang, Dwyer, and Marriott 2018). In areas with many people, it could be useful because overlapped lines can make it harder to see.

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.

Blumenschein, Michael, Xuan Zhang, David Pomerenke, Daniel A Keim, and Johannes Fuchs. 2020. "Evaluating Reordering Strategies for Cluster Identification in Parallel Coordinates." In *Computer Graphics Forum*, 39:537–49. 3. Wiley Online Library.

Bok, Jinwook, Bohyoung Kim, and Jinwook Seo. 2020. "Augmenting Parallel Coordinates Plots with Color-Coded Stacked Histograms." *IEEE Transactions on Visualization and Computer Graphics* 28 (7): 2563–76.

Carswell, C Melody. 1992. "Choosing Specifiers: An Evaluation of the Basic Tasks Model of Graphical Perception." *Human Factors* 34 (5): 535–54.

Chambers, John M. 1983. *Graphical Methods for Data Analysis*. Chapman; Hall/CRC.

Chang, Chunlei, Tim Dwyer, and Kim Marriott. 2018. "An Evaluation of Perceptually Complementary Views for Multivariate Data." In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, 195–204. IEEE.

Cleveland, William S. 1993. *Visualizing Data*. Hobart press.

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.

Conover, William Jay. 1999. *Practical Nonparametric Statistics*. Vol. 350. john wiley & sons.

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

David, FN, and JW Tukey. 1977. "Exploratory Data Analysis." *Biometrics* 33 (4): 768.

Firat, Elif E, Ben Swallow, and Robert S Laramee. 2023. "Pcp-Ed: Parallel Coordinate Plots for Ensemble Data." *Visual Informatics* 7 (1): 56–65.

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

Ge, Yawei, and Heike Hofmann. 2020. "A Grammar of Graphics Framework for Generalized Parallel Coordinate Plots." *arXiv Preprint arXiv:2009.12933*.

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, Yuan Luo, Arthur E Kirkpatrick, Hao Zhang, and Daniel Weiskopf. 2011. "Evaluation of a Bundling Technique for Parallel Coordinates." *arXiv Preprint arXiv:1109.6073*.

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.

———. 1997. "Multidimensional Detective." In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, 100–107. IEEE.

———. 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."

Kavvadias, Dimitris J, Grammati E Pantziou, Paul G Spirakis, and Christos D Zaroliagis. 1996. "Hammock-on-Ears Decomposition: A Technique for the Efficient Parallel Solution of Shortest Paths and Other Problems." *Theoretical Computer Science* 168 (1): 121–54.

Keim, Daniel A. 2002. "Information Visualization and Visual Data Mining." *IEEE Transactions on Visualization and Computer Graphics* 8 (1): 1–8.

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.

Moustafa, Rida EA. 2009. "QGPCP: Quantized Generalized Parallel Coordinate Plots for Large Multivariate Data Visualization." *Journal of Computational and Graphical Statistics* 18 (1): 32–51.

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.

Palmas, Gregorio, Myroslav Bachynskyi, Antti Oulasvirta, Hans Peter Seidel, and Tino Weinkauf. 2014. "An Edge-Bundling Layout for Interactive Parallel Coordinates." In *2014 IEEE Pacific Visualization Symposium*, 57–64. IEEE.

"Parallel Coordinates Plot in Ggplot2." n.d. https://plotly.com/ggplot2/parallel-coordinates-plot/.

Pilhöfer, Alexander, Alexander Gribov, and Antony Unwin. 2012. "Comparing Clusterings Using Bertin's Idea." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2506–15.

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.

Shah, Priti, and Eric G Freedman. 2011. "Bar and Line Graph Comprehension: An Interaction of Top-down and Bottom-up Processes." *Topics in Cognitive Science* 3 (3): 560–78.

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

Silverman, Bernard W. 2018. *Density Estimation for Statistics and Data Analysis*. Routledge.

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.

Symanzik, Jürgen, Michael Friendly, and Ortac Onder. 2018. "The Unsinkable Titanic Data."

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.

van der Maaten, Laurens, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (86): 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html.

VanderPlas, Susan, Yawei Ge, Antony Unwin, and Heike Hofmann. 2023. "Penguins Go Parallel: A Grammar of Graphics Framework for Generalized Parallel Coordinate Plots." *Journal of Computational and Graphical Statistics* 32 (4): 1572–87.

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, Dianne Cook, Heike Hofmann, and Andreas Buja. 2011. "tourr: An R Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software, Articles* 40 (2): 1–18. https://doi.org/10.18637/jss.v040.i02.

Wilcox, Rand. 2017. *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. Chapman; Hall/CRC.

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.