Lit Review for handling numerical ties in PCPs

Denise Bradford

Introduction

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 (Velagala et al. 2020). 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.

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.

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 $p \geq 4$ dimensions of data. Here, we examine parallel coordinate plots (PCPs) as a solution to $p \geq 4$ 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 $N \geq 4$ dimensional data across different magnitudes of N (observations) and p (variables).

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.

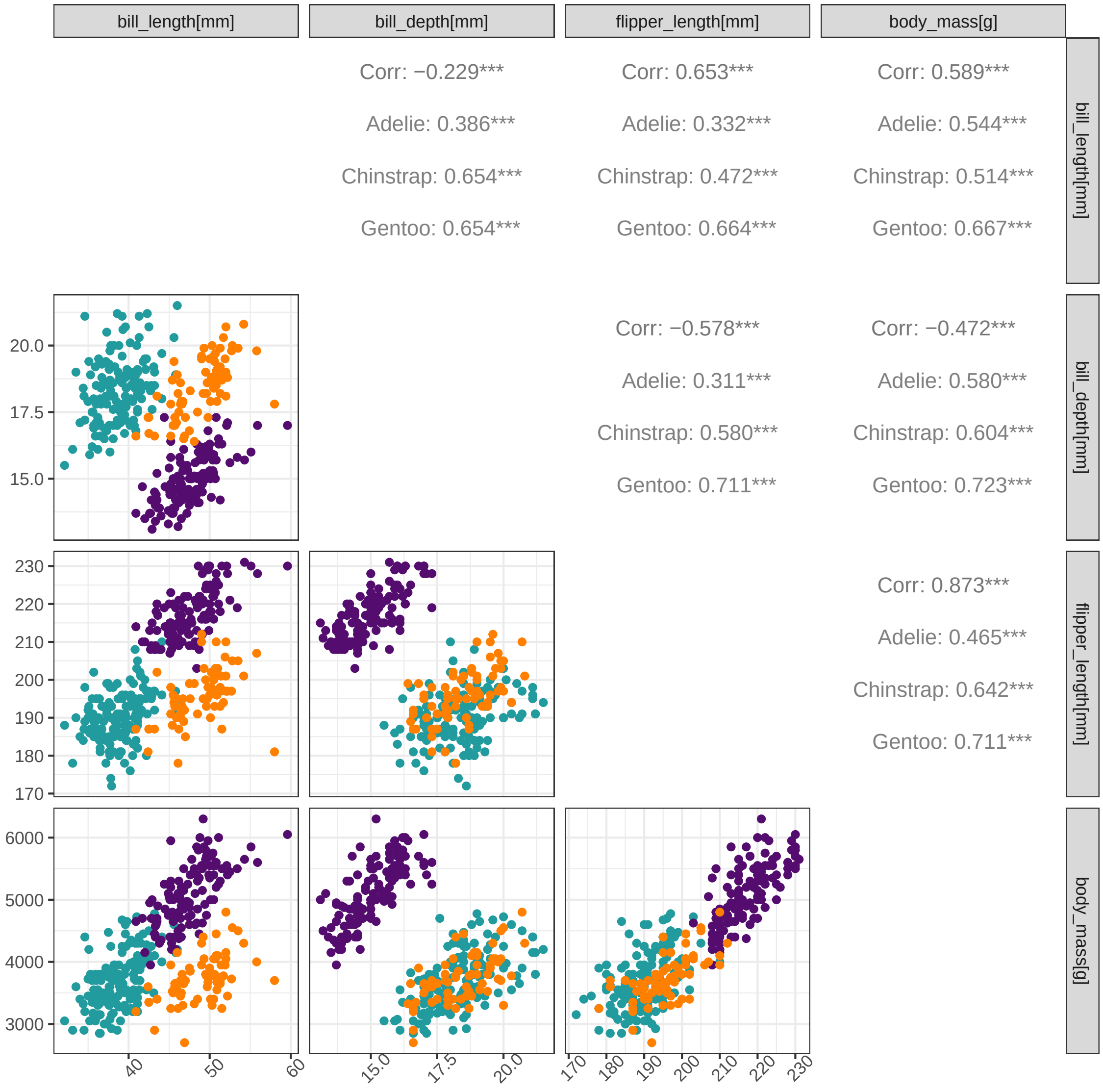

Figure 1: A generalized pairs plot

Figure 1 is a **scatterplot matrix**, which shows pairwise relationships between variables in Cartesian space. It shows how the variables for three kinds of penguins (Adelie, Chinstrap, and Gentoo) are distributed and how they are related.

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.

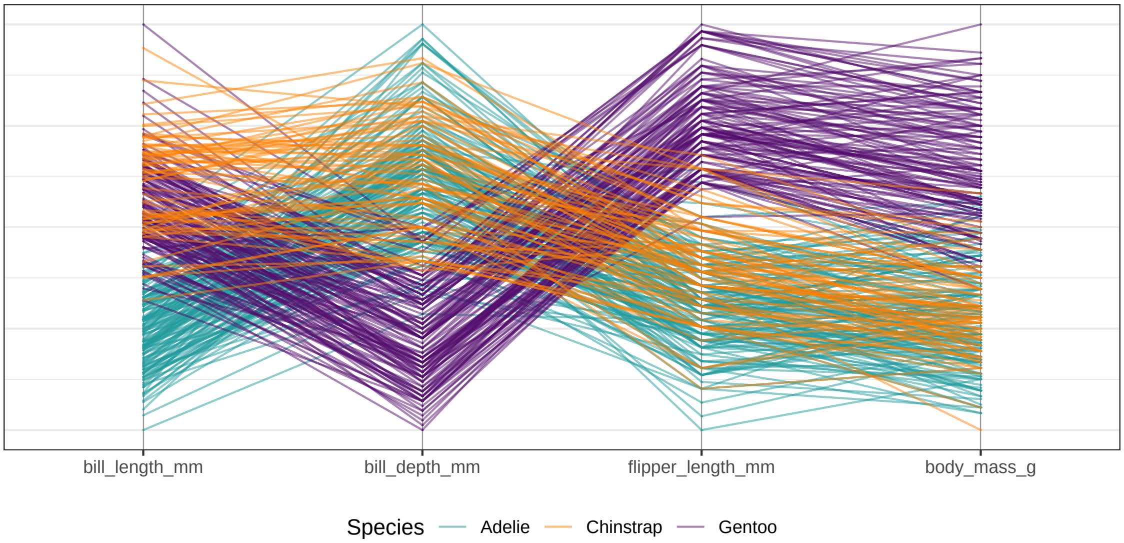

Figure 2: A parallel coordinate plot

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. 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 second plot in the third 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,Cleveland (1993) which can provide some information about the distribution of the variable, but is less direct than a continuous density plot.

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.

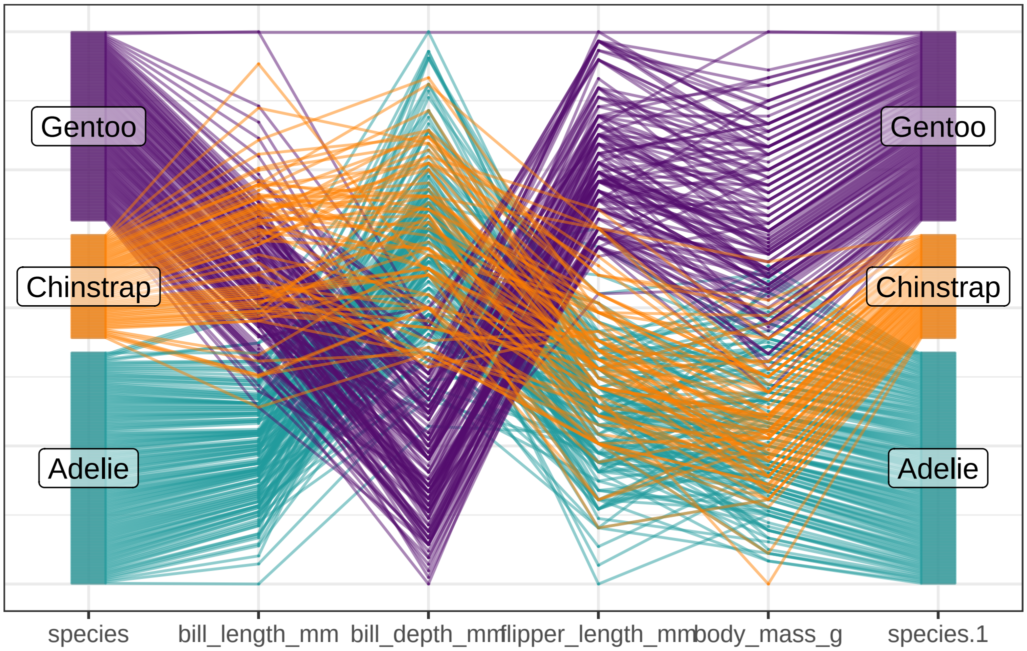

Figure 3: A generalized parallel coordinate plot, with species on the left and right of the plot; observations are ordered on the right side based on the value of body_mass_g, and on the left side based on the value of bill_length_mm. Translucent lines reduce the impact of overplotting and allow perception 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, in part 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 three 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; Abdi and Williams 2010), 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. PCPs are useful 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, 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**. Comparisons show that whilst PCPs are more suited for exploratory research when dimensions surpass reasonable limits for scatterplot matrices (Munzner 2014; Inselberg 2009), scatterplot matrices clearly show a depiction of interdimensional interactions for small to medium-scale data.

Although PCPs are sometimes criticized for overplotting and difficulty recognizing patterns, such as clusters or non-linear correlations, which scatterplots more efficiently disclose, PCPs provide a compact depiction of high-dimensional data (Johansson et al. 2005; Qu et al. 2007). Studies show, however, that users adjust to PCPs relatively fast following appropriate introduction since interactive techniques and hierarchical approaches significantly increase their usability (Fua, Ward, and Rundensteiner 1999; Claessen and Van Wijk 2011). When combined with user training and interactive visualization techniques, PCPs remain beneficial for exploratory analysis of higher-dimensional data, even if scatterplots provide instantaneous interpretability and clarity—especially for two-dimensional relationships (Inselberg 2009; Julian Heinrich and Weiskopf 2013a).

Variations on 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, as well as variations on PCPs which evolved in parallel for e.g. categorical variables.

Categorical Parallel Axis Plots

Traditional PCPs were primarily designed to display continuous data. However, real-world datasets often contain a mix of continuous, ordinal, and categorical data. In parallel to the development of numerical PCPs, a number of categorical plots with similar goals developed. Over time, the different approaches to categorical and numerical variables converged, leading to several different types of mixed-variable, parallel axis plots. We will first consider categorical plots with parallel vertical axes, and then examine mixed-type variants.

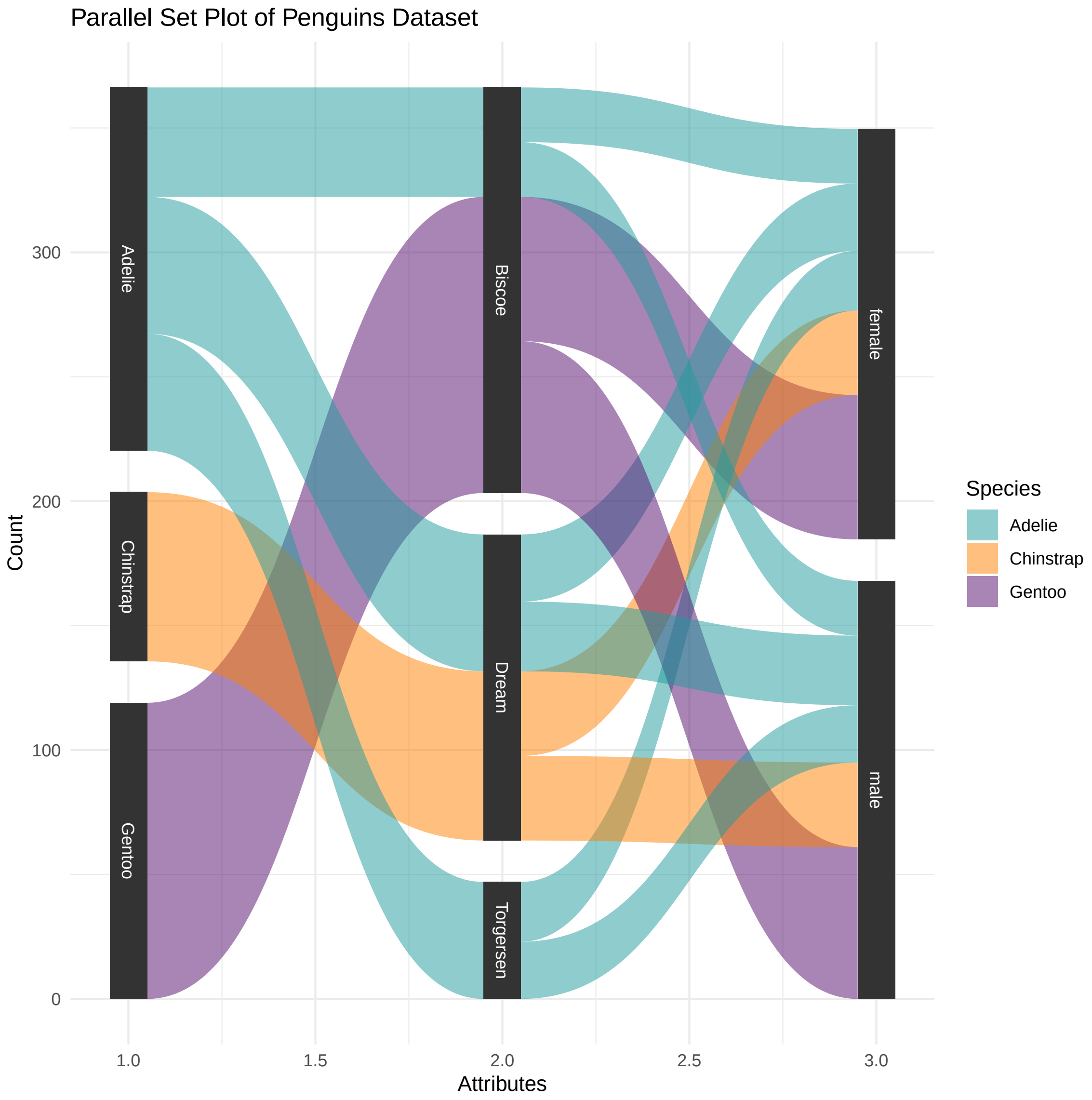

Figure 4: A Parallel Sets Plot

Parallel Sets Plots

Designed to examine multidimensional numerical data, parallel sets originated from parallel coordinate charts, first proposed by Alfred Inselberg in the 1950s. Led by Daniel Keim and others, the adaptation of categorical data first surfaced in the late 1990s. Kosara, Bendix, and Hauser (2006) further refined these methods by emphasizing the design and use of parallel plots to make complicated relationships more understandable.

Dynamic, interactive visualizations were made possible by developments in computational tools such as D3.js and Matplotlib in the 2010s. Temporal zooming and filtering improved the analytical depth. Therefore, these tools have become more critical for historical study (Bostock, Ogievetsky, and Heer 2011).

Parallel sets present multidimensional relationships, enabling access to complicated information, trends, and anomaly-revealing power. They may oversimplify continuous data and, in poorly crafted designs, provide cluttered, difficult-to-interpret visuals.

Parallel sets provide an intense way to find trends in multidimensional datasets and help to close the gap between quantitative and qualitative historical studies. These tools allow for predictive modeling of historical trends as computer capability increases and machine learning interacts with data visualization, changing our understanding of the past and its consequences for the present. Basic advances like Kosara, Bendix, and Hauser (2006) and the ongoing improvement of parallel sets plots promise a future of more affluent, more perceptive historical research.

For frequency-based representations, parallel sets replace individual data points, abstracting data to emphasize linkages. The SET-STAT-MAP system is flexible for mixed-data situations like environmental and policy studies because it expands parallel sets to include numerical and spatial data, Wang et al. (2022). Quality measures, including overlap, ribbon width variance, and mutual information, optimize dimension and category ordering and improve visualizations' interpretability.

Parallel sets depend on interactivity to allow users to investigate data relationships more naturally. These tools let users create aesthetically pleasing layouts that best fit their application. Across many fields, including meteorology, consumer behavior study, marketing, and resource management, parallel sets find application.

Dennig et al. (2021) introduced "ParSetgnostics," a set of quality metrics to reduce visual clutter in parallel sets. Metrics like overlap, ribbon width variance, and mutual information optimize dimension and category ordering, significantly improving visualizations' interpretability; for instance, applying these metrics reduced clutter by up to 81% in test cases. Parallel sets are also useful for instruction and analysis since their combination of statistical summaries and category flow analysis reveals a deeper understanding. Combining categories and automating axis selections guarantees that parallel sets stay scalable, therefore matching the aim of reducing the number of possible combinations produced on display without compromising data integrity.

Hammock Plots

Hammock plots are similar to parallel coordinate plots, but were designed to show categorical data (Schonlau 2003).

Fig 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

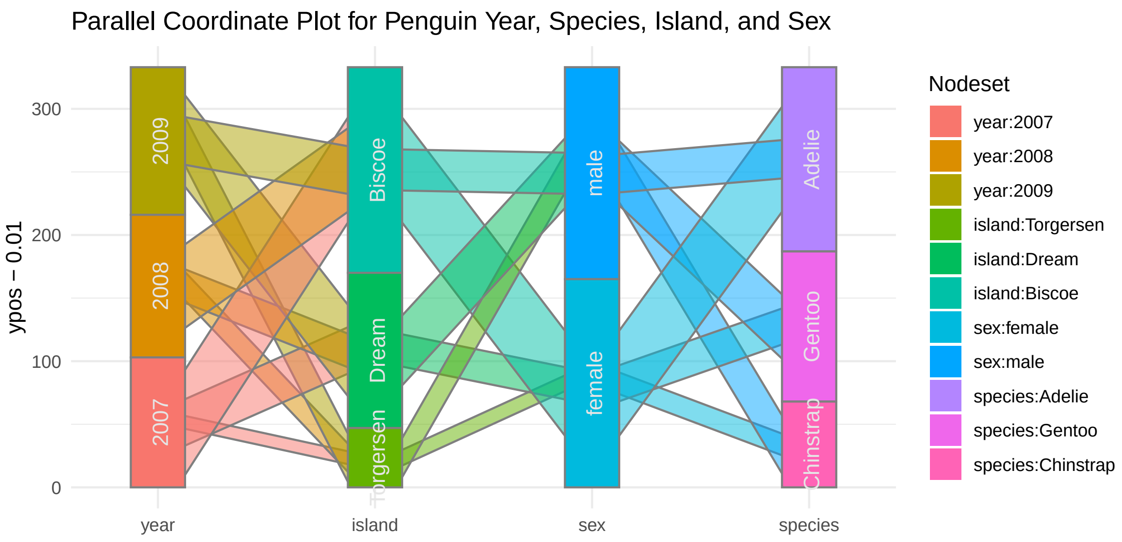

Figure 5: Hammock Plot (Schonlau 2024)

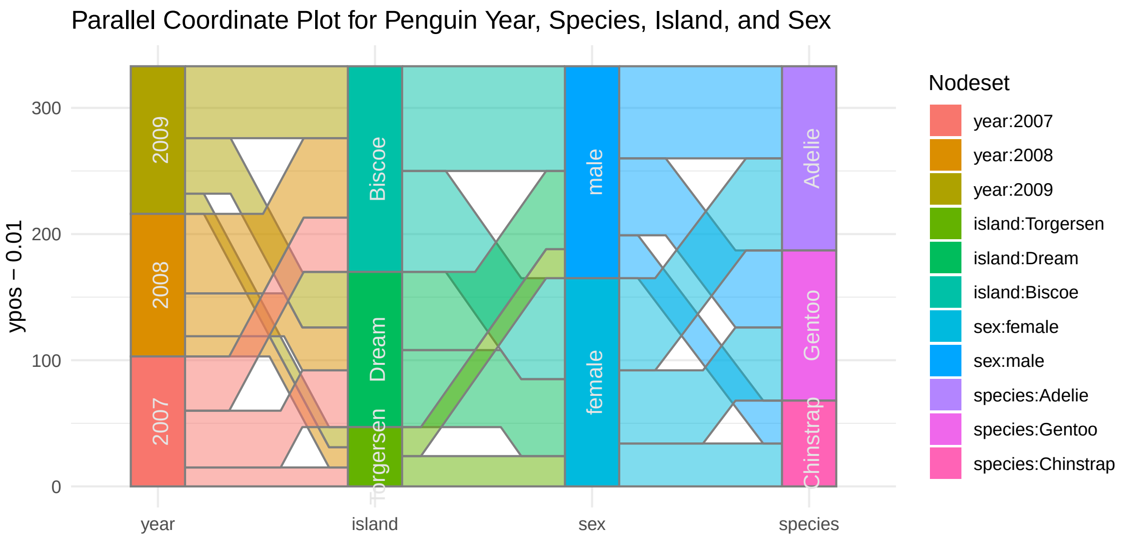

Figure 6: Common Angle Plot (Hofmann and Vendettuoli 2013)

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

While allowing mixed data types, they are meant to show relationships between categorical variables visually. Especially in health science datasets, key characteristics of Hammock plots include their capacity to simplify complicated relationships and lower visual clutter by using "bandwidth" representations to highlight links between categories.

When it comes to multidimensional categorical data, hammock plots show a more compact representation than mosaic plots, lowering the cognitive load on the observer and surpassing conventional tools like bar charts. These benefits are absolutely vital in disciplines including health research, where datasets can comprise mixed numerical and categorical variables needing simultaneous display.

Symanzik, Friendly, and Onder (2018) showed the Titanic dataset using hammock plots, 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 small amounts of complex data. However, they warn that "the method may require significant computational resources," which means it might not work for complex or larger-scale tasks.

Pilhöfer, Gribov, and Unwin (2012a) show that Hammock plots "manage clusters and resolve data ties" well, especially when comparing data clusters. Key strategies comprise clustering visualization, which employs algorithms to reorder rows and columns, obtain a pseudo-diagonal form, and resolve data connections via optimal sorting algorithms like the Weighted Bertin Classification Criteria (WBCC). In graphical displays, these techniques minimize crossing lines or misalignments and maximize concordance. Optimized displays of clustering results, using the "Ecotest dataset," show how ordering methods enhance cluster identification, Pilhöfer, Gribov, and Unwin (2012b). The paper also offers an evidence-based method, using measures such as the Bertin Classification Criteria (BCC) to evaluate graphical order, supported by computational tests. Both theoretical considerations and actual data support the assertions regarding Hammock plots controlling clusters and tying off conflicts.

Still, Pilhöfer, Gribov, and Unwin (2012a) 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. Both visual and computational methods help to find definite clustering in the content. The main technique is pseudo-diagonalization, in which rows and columns of a data matrix are rearranged to visually show clusters by combining similar entries. This technique is compared by a top-down partitioning method, which clusters the data using measurements such as Kendall's τ or the Bertin Classification Criteria (BCC), determining ideal cut spots in rows and columns. The BCC, which computes the difference between conforming and non-conforming pairings to gauge the alignment of matrix entries with a pseudo-diagonal form, is one of the metrics to evaluate clustering clarity. The Weighted Bertin Classification Criteria (WBCC) emphasizes the proximity of data points and improves this by using distance-based weights; the Bertin Classification Index (BCI) scales the BCC to gauge the strength of the link between clusters. BCI provides a consistent approach to evaluating datasets, with values ranging from 0 (perfectly aligned clusters) to 1 (indicating no clustering). The main evaluation criterion for clarity is the degree to which the matrix resembles a block-diagonal structure; low BCI values imply unambiguous clusters, while high values indicate indistinct or overlapping groupings. The paper also underlines the need for visual inspection, in which visually distinct and well-separated blocks in matrix plots—such as those shown in the Ecotest dataset examples—characterize clear grouping. Still, clarity is very arbitrary since it depends on thresholds selected for BCI and the interpretability of the produced visuals. This method detects and assesses clusters by combining empirical visualization with quantitative measurements.

The `ggparallel` package Hofmann and Vendettuoli (2013) provided a grammar-based interface for visualizing multivariate categorical data, implementing Hammock plots (Schonlau 2024). This makes them better suited for data that has both category and numerical relationships. This grammar-based adaptation retains the established behaviors of parallel categorical displays while providing additional options to handle edge cases like categorical ties.

Hofmann and Vendettuoli (2013) worked on Common Angle Plots, and Schonlau added Hammock plots to improve categorical data display. However, they approached the problem in different ways. Schonlau (2024) Hammock Plots: Visualizing Categorical and Numerical Variables generalizes parallel coordinate plots by swapping lines with box elements. The width of the boxes shows the number of observations, so they can be used for both categorical and mixed data. By changing traditional Hammock plot designs, Schonlau deals with perceptual problems like the reverse line width illusion. Hofmann and Vendettuoli (2013) also addresses the line width illusion, proposing Common Angle plots that use a constant angle to mitigate the line width illusion while maintaining visual continuity of the categories. Hofmann's Common Angle Plots are unique because they use consistent angular links to fix problems when line slopes or widths are different. This ensures that the visualizations are perceived accurately, preserving the marginal proportions while still connecting observations across parallel axes. Schonlau stresses the importance of being flexible with different data types, while Hofmann uses a consistent angular structure to clarify interpretations. Together, these approaches show a shared desire to make categorical and mixed data displays more accurate and useful, using different but complementary methods.

Alluvial plots & Sankey diagrams

Alluvial Plots

Alluvial plots are a powerful tool for visualizing flows and transitions in categorical data, capturing changes over time or across dimensions. Their conceptual inspiration stems from stratified data representations, evoking imagery of sedimentation patterns in geology. Popularized in the early 2000s, these plots have found applications in diverse fields, such as sociology and genomics, where understanding relationships and transitions within datasets is critical(Rosvall and Bergstrom 2008).

The structure of alluvial plots centers around parallel axes that represent categorical dimensions, with bands connecting related elements across these axes. The width of the bands may indicate magnitude or relative importance, but the primary focus is on qualitative transitions. For instance, in bibliometrics, alluvial plots can illustrate the evolution of research topics, while in sociology, they depict migration patterns or

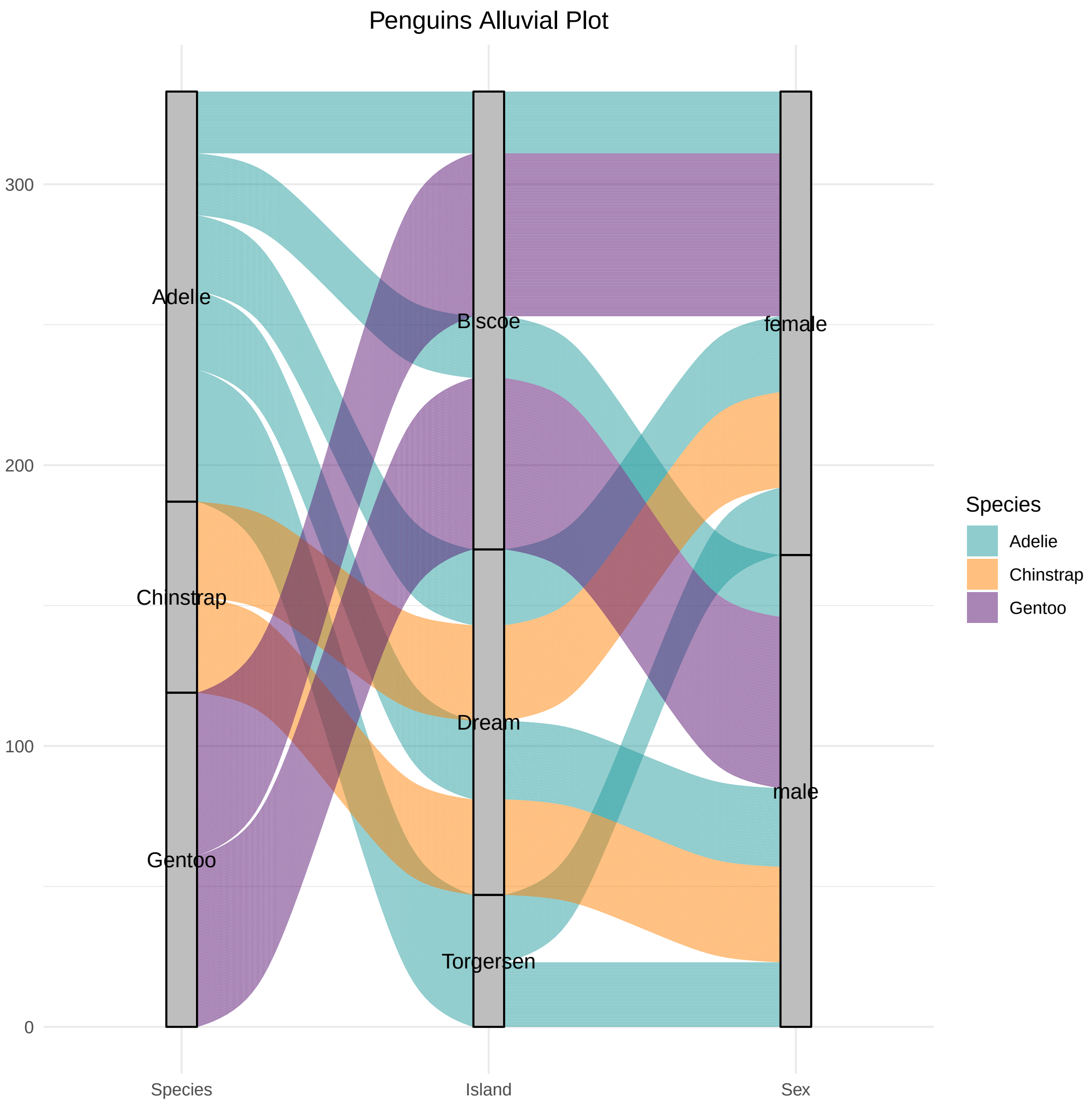

Figure 7: A Alluvial Plots

demographic shifts (**bastiani2004?**). These visualizations are particularly valued for intuitively conveying complex categorical relationships.

Sankey Diagrams

Sankey diagrams, introduced in 1898 by M.H.P.R. Sankey during his study of steam engine efficiency, were initially devised to illustrate energy flows. Over time, their scope has broadened to encompass resource allocations, material transfers, and financial systems, finding relevance in fields like environmental science, logistics, and economics (**schmidt1925?**).

The defining feature of a Sankey diagram is its proportional representation of flow magnitude. Nodes, representing entities or processes, are connected by links whose widths correspond to the transferred quantity. This emphasis on proportionality makes Sankey diagrams ideal for visualizing resource distributions, identifying inefficiencies, and analyzing systems with significant quantitative components. For example, they are often used to track energy inputs and losses in power plants or to represent financial data (**tufte1983?**).

Sankey diagrams and alluvial plots are data visualization techniques that differ in purpose and design. Sankey diagrams focus on quantitative data flows, such as energy or resource transfers, whereas alluvial plots focus on categorical transitions and linkages. They use nodes and proportional linkages to represent flow magnitude, whereas alluvial charts use parallel axes connected by bands to show categorical correlations. Sankey diagrams are more suited to systems with numerical flows, such as energy analysis or financial tracking. In contrast, alluvial plots are better suited to investigating trends in categorical data across dimensions. Both visualizations are cross-disciplinary, scalable to various datasets, and provide considerable customization possibilities.

Then move into mixed type plots -

Hurley (2004) end link algorithm – how does this relate?

CPCPs use discrete axis segments or unique markers for each category, transforming categorical variables into distinguishable visual elements. 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.

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.

The Generalized Parallel Coordinate Plot (GPCP) is a modified version of the original PCP that introduces changes to the axes to allow for more complex data representation. Nonlinear scales such as logarithmic and exponential transformations are possible.

Mixed Variable Parallel Representations

Bundling and Curving

Bundling and curving techniques have emerged as practical solutions for managing visual clutter in high-dimensional data visualizations. These methods group similar paths, reduce overplotting and make overarching patterns more visible. By organizing the chaotic

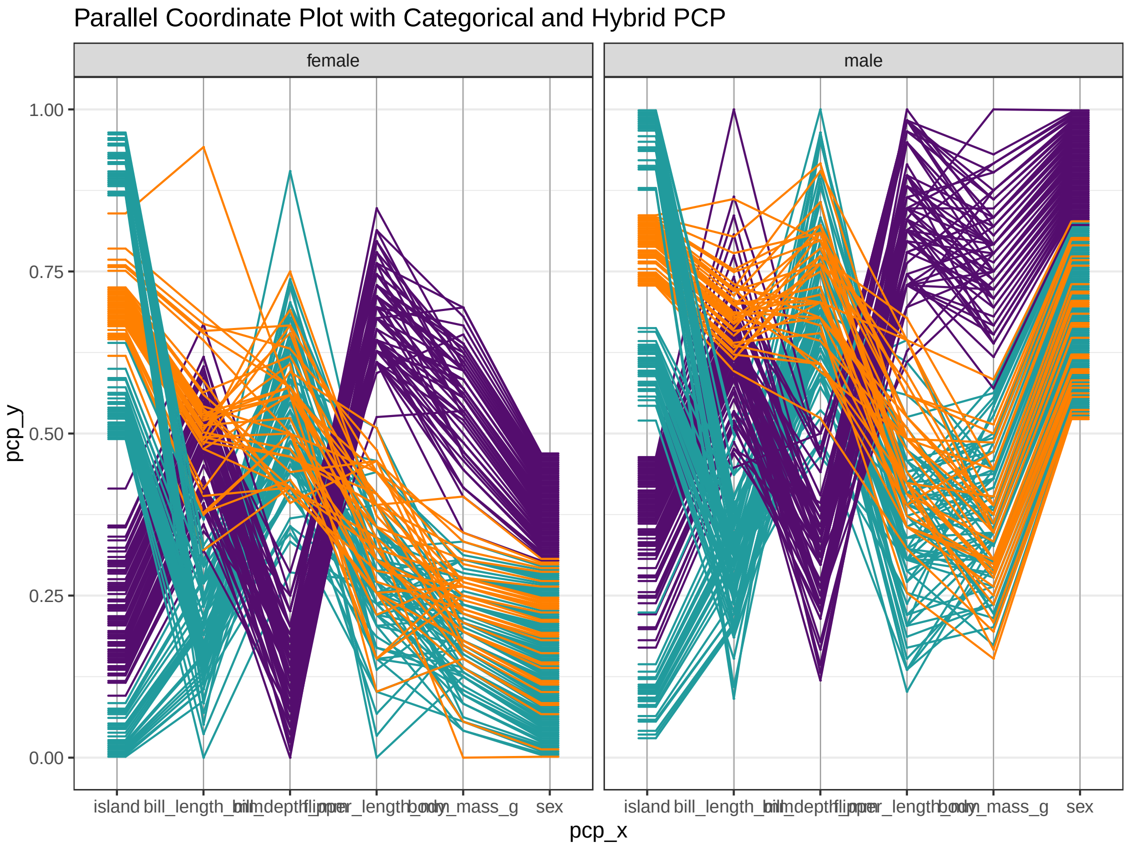

Figure 8: Categorical and Hybrid in PCP

lines often seen in parallel coordinate plots (PCPs), bundling and curving enhance interpretability. As noted by Holten and Van Wijk (2009), bundling provides a compelling strategy to mitigate the visual complexity frequently encountered in these datasets.

Building on this concept, McDonnell and Mueller (2008) introduced curvilinear PCPs, leveraging curved lines to distinguish intersecting paths and improve visual clarity. Subsequently, Johansson and Forsell (2015) assessed the effectiveness of bundling and curving in PCPs, offering criteria for when these modifications are most beneficial for interpretability.

Curving lines in PCPs can be adapted based on data attributes, enabling better visual separation and pattern recognition. This is particularly advantageous when working with strongly correlated dimensions, where overlapping paths obscure relationships. Curved lines visually separate trajectories by reducing intersections, making data relationships easier to discern. Additionally, they enhance aesthetic appeal and reduce cognitive load, improving the overall readability of visualizations. Curved lines offer a more intuitive perception of relative data positions by simulating a three-dimensional effect within a two-dimensional space. Research supports these benefits, with studies showing that curving lines reduces clutter and improves data interpretation Qu et al. (2007). This approach has proven effective in handling complex datasets with overlapping points (Kachhway 2013).

In Figure 9, a PCP modified with bundling and curving demonstrates how these techniques smooth lines, reduce distractions and highlight patterns across different species. Bundling reveals general trends within each species, but it lacks the detail to show specific distribution shapes, such as those visible in histograms. A notable trade-off of bundling is its tendency to obscure outliers, as extreme values are pulled into the primary flow,

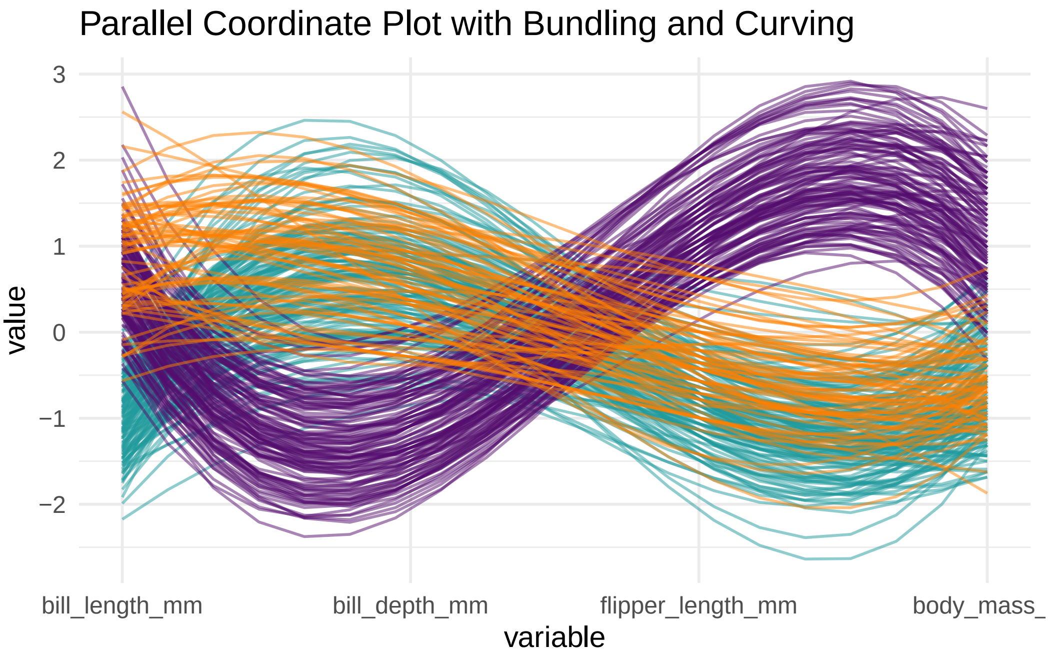

Figure 9: A Bundling and Curving Parallel Coordinate Plot

making deviations harder to detect. However, curved lines make variable relationships more apparent, indicating that specific measurements rise or fall together depending on the species. Despite this, minor differences between individual points may be harder to discern.

Rida E. Moustafa (2011) expanded on PCPs by integrating them with parallel coordinate density plots (PCDPs) to effectively manage large datasets. By transforming PCP images with density estimation methods, Moustafa's approach converts polylines into continuous, smooth depictions of data density. This innovation highlights groupings and trends often hidden in traditional PCPs. Moreover, the interactive elements of Moustafa's methodology allow users to explore different dimensions in real time, making the visualization accessible even to non-experts. This interactivity significantly enhances the analytical capabilities of PCPs.

While bundling simplifies complex datasets and reduces visual clutter, it has limitations. Grouping similar paths can obscure individual data points and outliers, which may be critical in specific analyses. McDonnell and Mueller caution that the clarity achieved through bundling may compromise data precision, as unique data paths or outliers are often lost in the simplification process (McDonnell and Mueller 2008).

Enhanced Color Encoding and Shading

Visual enhancements such as color coding, opacity modifications, and line thickness adjustments have been implemented to address overplotting, particularly in large datasets. These upgrades allow significant data trends to stand out while less relevant information is minimized. For instance, varying colors or adjusting line opacity based on data density helps users identify trends and outliers that might otherwise remain hidden. Similarly, modifying line thickness to represent variables such as data frequency or confidence intervals adds depth to the visualization. These multi-layered techniques enrich traditional parallel coordinate plots (PCPs) by encoding additional dimensions of information.

In their comprehensive review, State of the Art of Parallel Coordinates, Julian Heinrich and Weiskopf (2013a) discusses numerous methods to reduce overplotting and improve data interpretation. Techniques such as alpha blending, which adjusts line transparency to minimize visual clutter, and interactive line manipulation, which allows users to explore data in real time, are highlighted as essential tools for enhancing PCP readability in large datasets.

Building on these foundations, Blaas, Botha, and Post (2008) emphasized the importance of dynamic visual traits like adjustable colors and opacity in their study Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-Timepoint Data Sets. Their approach enables real-time modification of visual settings, allowing users to focus on specific data features and control line density. Similarly, Raidou et al. (2015), in Orientation-Enhanced Parallel Coordinate Plots, demonstrated how varying opacity levels and color gradients can draw attention to patterns, show line densities, and facilitate interactive exploration. These strategies effectively tackle overplotting challenges in complex datasets by employing techniques such as smoothing and averaging polylines.

The work of Zhou et al. (2008) in Visual Clustering in Parallel Coordinates introduced a user-driven approach to highlight data clusters using customizable color and opacity settings. This technique differentiates data groups and reshapes curves to enhance clarity. Likewise, Fua, Ward, and Rundensteiner (1999)'s Hierarchical Parallel Coordinates for Exploration of Large Datasets proposed reducing visual clutter through hierarchical representations and variable line thickness. By effectively visualizing data density, these methods reveal trends otherwise obscured in large datasets.

Color gradients and shading have emerged as powerful tools to encode additional data attributes, such as density or frequency, in PCPs. Using these visual encodings, users can identify trends and correlations more effectively. For example, color gradients transform PCPs from purely structural plots into tools for multi-dimensional analysis. As Theus and Urbanek (2008) demonstrated with color-coded PCPs, mapping additional variables through color significantly enhances interpretive power. This approach was further refined by Bertini, Dell'Aquila, and Santucci (2005), who applied density shading techniques to make frequent patterns in large datasets more visible. Novotny and Hauser (2006) expanded on this work with opacity and color-blending techniques, making overlapping patterns easier to interpret intuitively.

However, reliance on color-based encoding poses challenges. Overusing colors or gradients can lead to visual strain and reduce clarity, especially for users with color vision deficiencies. Bertini et al. noted that excessive color application risks overwhelming the viewer, making the visualization more confusing rather than more insightful.

These studies underscore the importance of adapting visual attributes in PCPs to address overplotting and improve interpretability. By incorporating techniques like color coding, opacity adjustments, and line thickness modifications, PCPs become more effective tools for analyzing multi-dimensional datasets. However, balancing these enhancements to maintain their usability is essential, particularly for diverse audiences. Thoughtful application of these techniques ensures that PCPs remain accessible, clear, and insightful for exploring complex datasets.

The parallel coordinate plot illustrated in Figure 10 employs enhanced color coding to distinguish between species effectively. Each species is represented by a distinct color: red for Adelie, green for Chinstrap, and blue for Gentoo. Adjustments to line density further highlight the differences between species across key variables such as bill length, bill depth, flipper length, and body mass. This approach successfully conveys general

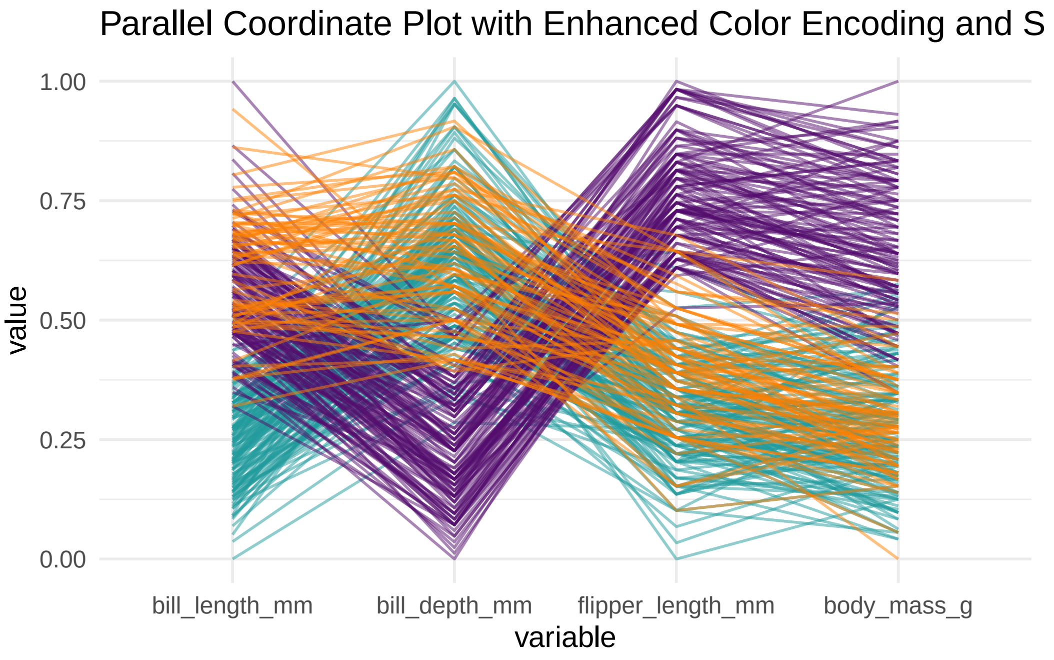

Figure 10: Enhanced Color Encoding and Shading in PCP

trends for each species, such as the Gentoo's higher flipper length and body mass values than the Adelie, which typically exhibits lower values.

Despite these enhancements, overplotting remains an issue, particularly around the middle ranges of each axis. This visual congestion obscures slight differences and makes identifying outliers challenging. While the plot does not explicitly show the distribution shapes of individual variables, the grouping of lines along each axis hints at central tendencies for each species. For instance, clustered lines suggest shared characteristics within species, though these clusters lack precision in depicting variability.

The general trends revealed in the plot suggest relationships between numerical factors, but the overlapping lines make it difficult to isolate individual correlations. While this visualization provides a broad overview of trends across species, its ability to pinpoint specific data points or detect rare observations is limited. To enhance its utility, further modifications—such as interactive features or additional encoding dimensions—could help overcome these limitations and improve the analysis of individual patterns within the data.

Dimension Reduction Techniques

Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), are widely used to improve the readability of parallel coordinate plots (PCPs) before visualization. These methods reduce the number of dimensions while retaining the most significant aspects of the data, simplifying analysis without losing critical information. PCPs of reduced dimensions precisely depict high-dimensional interactions, making patterns and relationships easier to discern.

In addition to dimensionality reduction, axis ordering is crucial in enhancing PCPs. The relationships between variables become more apparent by arranging axes based on correlation, mutual information, or other metrics. Automated axis arrangement techniques have been developed to reduce line crossings, significantly improving the identification of patterns. These optimizations make PCPs a more effective tool for exploring complex datasets.

The Orientation-Enhanced Parallel Coordinate Plots technique, introduced by Raidou et al. (2015), represents a significant innovation in making PCPs more accessible and informative. This method addresses common challenges such as clutter and overlapping lines by dynamically adjusting plot axis orientations. By applying both automatic and user-interactive adjustments, this approach reduces visual noise and highlights meaningful data patterns and correlations. Modifying axis orientation on the fly offers users greater control over the visualization, enabling more precise analysis of large datasets.

Orientation-enhanced PCPs incorporate preprocessing steps like PCA and multidimensional scaling (MDS) to optimize axis arrangements. These techniques ensure that data clusters are well-separated and that overlapping lines are minimized. Interactive features further empower users to customize the orientation based on their specific research needs, making these plots adaptable to various analytical scenarios (Raidou et al. 2015).

Reducing dimensions has become essential for managing the complexity of PCPs when visualizing high-dimensional data. Techniques such as PCA and clustering allow analysts to focus on the most critical features of the data while reducing visual overload. PCA, for example, condenses the data into components that capture the majority of variance, ensuring essential information is retained. Similarly, clustering simplifies the dataset by grouping similar features, making trends more discernible. Wegman and Luo (1997) demonstrated the value of PCA within PCPs by enabling users to represent key data features while discarding less relevant dimensions. Further advancements by H. Guo, Xiao, and Yuan (2012) introduced hierarchical dimension reduction and clustering techniques, streamlining the analysis of high-dimensional datasets while preserving usability. As explored by Yang et al. (2017), non-linear methods further enhanced PCPs by capturing complex, non-linear relationships that traditional techniques might overlook.

While dimension reduction offers significant advantages, it has trade-offs. Techniques like PCA or clustering can inadvertently remove subtle nuances or minor patterns, potentially masking relationships that exist only in higher-dimensional spaces. H. Guo, Xiao, and Yuan (2012) cautioned that these beneficial methods might lead to information loss, particularly for niche analyses where fine-grained details are crucial.

By balancing the need for simplification with careful consideration of the data's nuances, PCPs can effectively represent high-dimensional datasets while maintaining their interpretive power. Combining dimensionality reduction, axis optimization, and user-interactive features ensures that PCPs remain robust tools for exploring and analyzing complex data.

In Figure 11, the parallel coordinate plot leverages principal component analysis (PCA) to simplify high-dimensional data. The x-axis represents PCA components (PC1 to PC4), while the y-axis displays scaled PCA values, color-coded by species: Adelie (red), Chinstrap (green), and Gentoo (purple). PCA encapsulates complex relationships into fewer dimensions by reducing the original variables into principal components, allowing for a more precise depiction of overall trends.

The lines connecting the components highlight how each penguin transitions across these derived features, making species-level patterns more discernible in the reduced dimensional space. Despite this, overplotting persists due to the numerous overlapping lines, particularly in densely populated regions. However, distinct separations between species

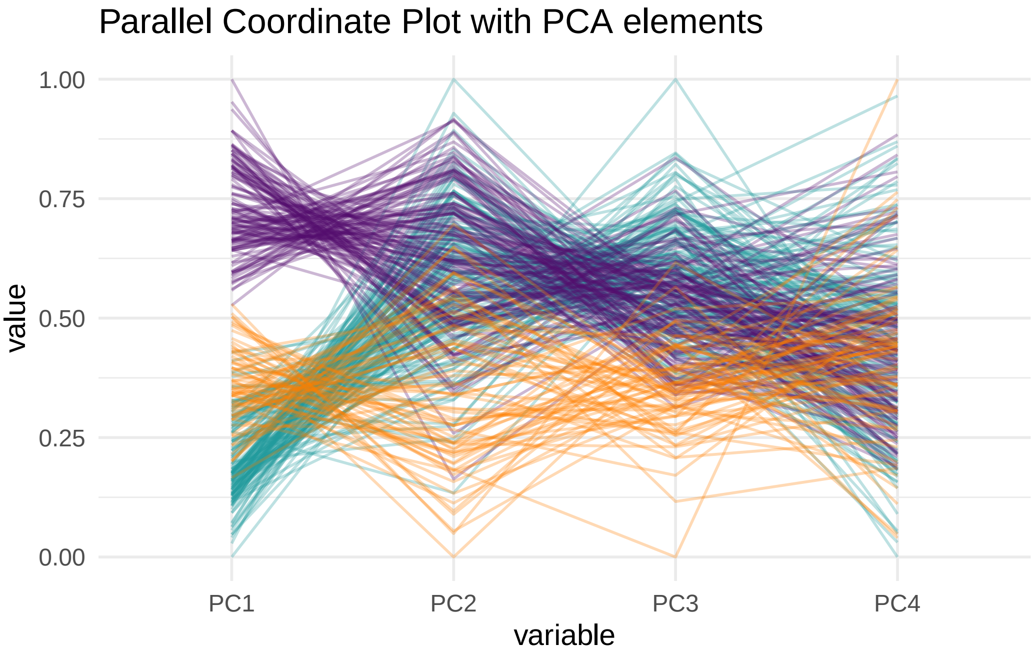

Figure 11: A PCA in Parallel Coordinate Plot

emerge along specific principal components, showcasing PCA's utility in revealing broad group-level trends.

Clusters around the PC axes indicate each species' general concentration of values, hinting at their central tendencies in the reduced space. While explicit distribution shapes are not depicted, these clusters offer insight into species-specific patterns. Color-coded lines and groupings emphasize the relationships between PCs and species, suggesting that specific principal components contribute to species separation. However, the bundling effect of overlapping lines makes detecting outliers and understanding individual variability more challenging.

This visualization demonstrates the potential of PCA in reducing dimensional complexity and enabling the identification of trends in large datasets. While PCA's abstraction simplifies relationships, trade-offs remain in preserving fine-grained details and outlier visibility, underscoring the importance of complementary techniques for in-depth analysis.

Cluster-Based and Hierarchical PCPs

Clustering techniques, such as spectral clustering, have proven instrumental in improving the interpretability of parallel coordinate plots (PCPs) by grouping related data points and revealing underlying structures in high-dimensional datasets. These methods allow analysts to uncover patterns and relationships not immediately evident in raw data, providing insights into the natural grouping of data points (Zhao and Kaufman 2012). Visualizing clusters in PCPs makes connections between data points and cluster features more apparent, facilitating a deeper understanding of complex datasets.

Before generating PCPs, data reduction techniques such as hierarchical clustering and principal component analysis (PCA) are often employed. These strategies focus on the most significant components that explain the majority of the variance in the data, limiting the number of dimensions visualized and reducing visual clutter. This preprocessing step enhances the clarity and utility of PCPs, mainly when dealing with large, multidimensional datasets.

The study Structure Revealing Techniques Based on Parallel Coordinates Plot by Zhao and Kaufman (2012) highlights the limitations of traditional PCPs in uncovering important patterns within high-dimensional data, particularly when faced with challenges like overplotting. To address these issues, the authors propose clustering and sorting techniques designed explicitly for PCPs. By leveraging spectrum theory, algorithms are developed to group similar polylines, reducing visual clutter and exposing hidden trends and correlations.

Additionally, the study introduces a correlation-based sorting method to arrange axes in a way that highlights relationships between variables. This systematic approach enables the visualization of trends across dimensions, making it easier to interpret complex data. Including view-range metrics further improves data clarity by aggregating and limiting noisy datasets. Experimental results demonstrate that these enhancements significantly improve PCPs' ability to reveal valuable patterns, trends, and correlations, enabling more efficient data analysis.

For multidimensional datasets, clustering, and hierarchical PCPs provide structured and organized visualizations, making comparing clusters and identifying trends easier. Hierarchical clustering enhances the narrative of PCPs by presenting high-level overviews and detailed views, as argued by Fua, Ward, and Rundensteiner (1999). This dual perspective allows analysts to observe the big picture while examining finer details within subsets of the data.

Fua, Ward, and Rundensteiner (1999) pioneered hierarchical PCPs with a zoomable interface, enabling users to explore data clusters and delve into specific subsets for detailed analysis. Geng, Deng, and Ali (2005) expanded on this concept by applying clustering to group similar data points, highlighting patterns, and making general trends more accessible. Further advancements by Poco et al. (2011) incorporated hierarchical clustering with user-defined granularity levels, providing flexible and interactive visualizations tailored to analytical needs.

While clustering simplifies data visualization and enhances the accessibility of patterns in PCPs, it comes with trade-offs. The grouping of data points can obscure individual data, particularly outliers. Geng, Deng, and Ali (2005) observed that clustering effectively identifies general patterns but may mask unique or significant findings hidden within homogenous groups. To strike a balance, analysts must carefully weigh the benefits of clustering against the potential loss of detail, ensuring that the visualization supports broad overviews and detailed investigations.

By integrating clustering techniques, hierarchical approaches, and correlation-based sorting methods, PCPs have evolved into more powerful tools for exploring high-dimensional datasets. These advancements enhance the visualization's ability to expose hidden structures while maintaining flexibility for detailed analysis, making PCPs indispensable in modern data analysis workflows.

Optimized Layout Algorithms

Layout algorithms for parallel coordinate plots (PCPs) have been refined to improve clarity by minimizing line crossings, which is critical when visualizing high-dimensional data. An optimized layout enhances the readability of PCPs by reducing visual clutter and facilitating smoother navigation. Effective design often determines whether a

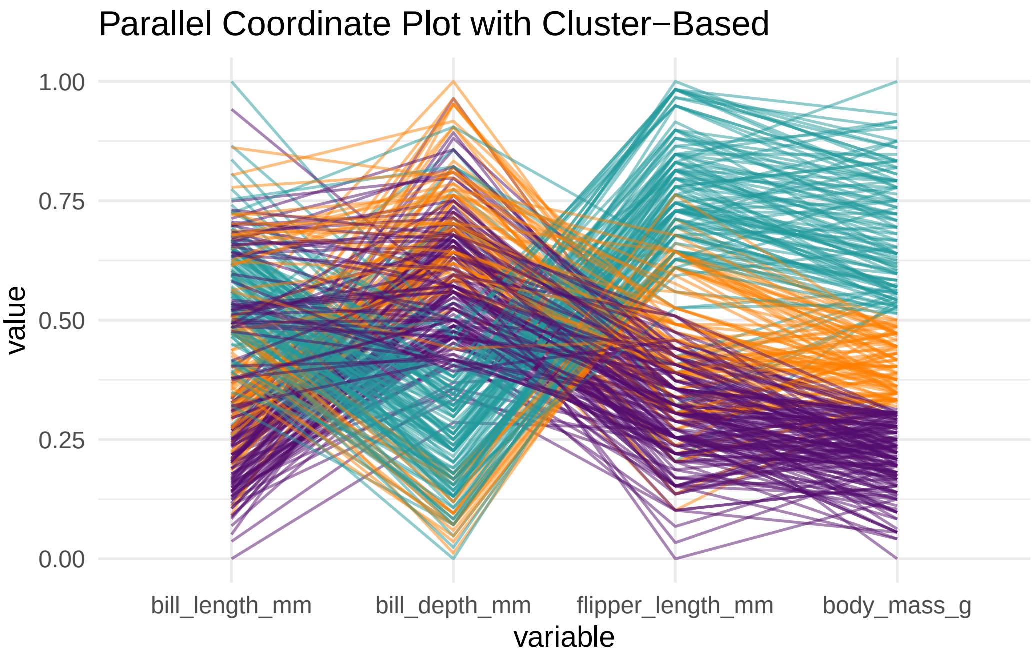

Figure 12: Cluster-Based PCP

PCP is comprehensible or chaotic (Ankerst, Berchtold, and Keim 1998). In an empirical analysis using three visualization techniques, including PCPs, researchers compared conventional sequential layouts with similarity-based arrangements. Their findings revealed that similarity-based configurations significantly enhanced readability by grouping comparable dimensions, making patterns, correlations, and functional connections more apparent. This alignment of similar dimensions allows for more precise identification of trends, underscoring the impact of thoughtful axis arrangement.

Heuristic layout algorithms in PCPs often generate results misaligned with specific user-defined analytical needs (Johansson et al. 2005). To address this limitation, Johansson et al. (2005) introduced an optimized axis positioning algorithm to maximize spacing between overlapping lines, thereby improving data distinction. In this context, data distinction refers to the ability to expose significant structures within dense PCPs, such as cluster formations, local outliers, and intensity fluctuations.

Johansson's work incorporates clustering, high-precision texturing, and transfer functions (TFs) to highlight multiple data properties. These innovations allow users to discern cluster structures, detect local outliers, and examine intensity variations in large datasets. Techniques like logarithmic TFs reveal dense structures and subtle patterns, while variance-based feature animation highlights areas requiring closer analysis. These methods, evaluated on datasets with up to 100,000 data points, demonstrate the capacity of PCPs to visualize complex trends and sub-clusters efficiently.

Two notable approaches for optimizing PCPs include the Similarity Clustering Approach (Ankerst, Berchtold, and Keim 1998) and the Clustered Parallel Coordinates with High-Precision Textures (Johansson et al. 2005). Both methods enhance interpretability by focusing on dimensional relationships, albeit with differing priorities.

- Similarity Clustering Approach (Ankerst, Berchtold, and Keim (1998)): This method emphasizes arranging dimensions based on functional similarities and correlations rather than explicitly reducing line crossings. Grouping related variables facilitates the discovery of patterns and connections, enhancing interpretability in datasets with strong inter-variable relationships. However, it is less effective for categorical data where reducing line crossings is critical.

- Clustered PCPs with High-Precision Textures (Johansson et al. (2005)): This approach addresses visual clutter by aggregating data points into clusters and representing them with high-precision textures. Transfer functions and outlier detection reduce overplotting and highlight significant data features without altering dimension arrangements. While effective for trend and structure visualization, it does not tackle dimension reordering to minimize crossings.

The `ggpcp` method minimizes line crossings in PCPs, particularly for categorical variables. It provides a direct solution for visual clutter by optimizing variable order, making it ideal for datasets where reducing overlap is a priority. However, its focus on categorical data limits its applicability to broader multidimensional contexts.

In contrast, the methods proposed by Ankerst, Berchtold, and Keim (1998) and Johansson et al. (2005) address broader interpretive challenges in PCPs. The former prioritizes grouping similar dimensions to highlight functional correlations, while the latter enhances visual representation through clustering and advanced texturing techniques. These approaches suit numerical and high-dimensional datasets where trends and relationships require detailed exploration.

Each method offers unique strengths and is best suited for specific use cases. The Similarity Clustering Approach and Clustered PCPs excel in revealing trends and relationships in numerical data, while `ggpcp` is better equipped to handle categorical data by reducing line crossings. These methods illustrate the diverse strategies for optimizing PCPs, ensuring their adaptability to various analytical needs and datasets. While they share a common goal of improving interpretability, their distinct methodologies cater to different aspects of the visualization challenge, from structural clarity to categorical simplification.

Perceptual Challenges with PCPs

- Clutter
- Interpretation isn't natural
- following an observation across the plot
- line width illusion susceptibility

Optimization of PCPs

Interactivity

Interactive and Dynamic

The integration of interactive features has significantly improved the utility and flexibility of parallel coordinate plots (PCPs). These enhancements allow users to dynamically reorder axes, filter data based on specified criteria, and invert axes to view data from different perspectives. Brushing and connecting highlight specific data points across axes while filtering reduces clutter by hiding irrelevant lines. These capabilities facilitate

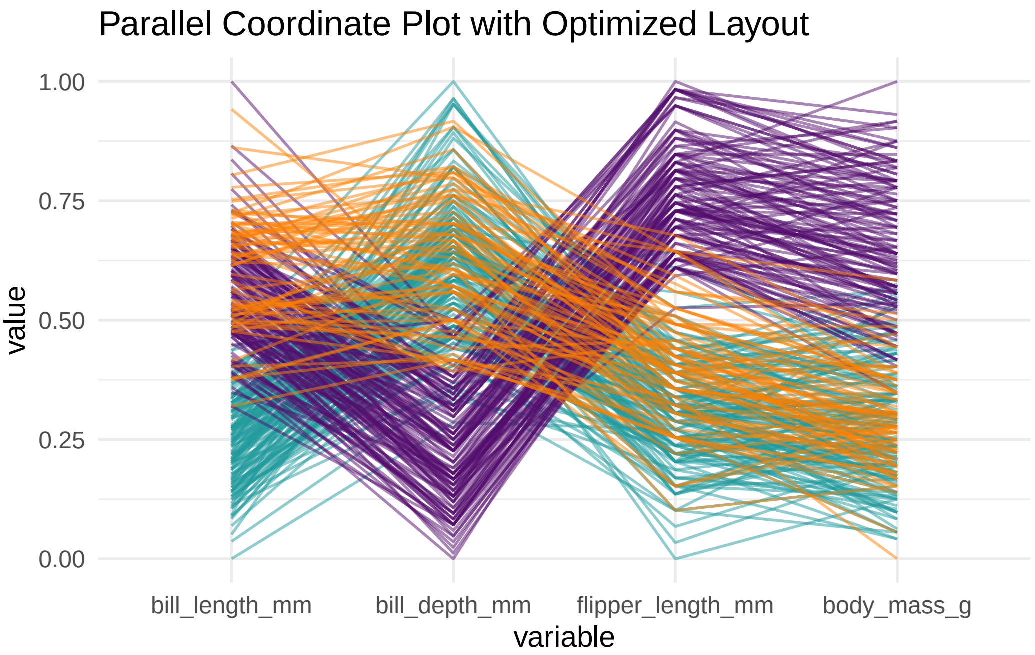

Figure 13: While the PCP employs an optimized layout, its clarity is compromised by overplotting, lack of clustering or feature enhancement, and minimal reduction in line crossings. Clustering, high-precision textures, and transfer functions could greatly improve readability by simplifying data representation and directing attention to key features.

real-time data analysis, enabling users to uncover patterns and correlations intuitively. Interactive PCPs are particularly valuable for exploratory data analysis, especially when dealing with large and complex datasets.

Julian Heinrich and Weiskopf (2013a)'s State of the Art of Parallel Coordinates provides a comprehensive review of PCP visualization techniques, categorizing them into a taxonomy of approaches. The study highlights modeling methods, visualizing, and interacting with PCPs, emphasizing their applications in knowledge discovery tasks like sorting, clustering, and regression. Key advancements include geometric modeling, interpolation techniques, point-line duality, and foundational principles of PCPs. Further innovations discussed include density-based visualizations, improved axis ordering, and interactive techniques such as brushing and bundling, which address challenges like overplotting and poor axis arrangements.

The study also explores practical applications of PCPs in engineering and life sciences, underscoring their adaptability for high-dimensional data analysis. Techniques like clustering, density estimation, and interactive filtering enhance data exploration, enabling users to map patterns through curves, shapes, and density plots.

Interactive features in PCPs have revolutionized how users interact with high-dimensional datasets. Tools like brushing, linking, and filtering allow precise and focused analysis. These capabilities help users engage intuitively with data, selectively highlighting areas of interest to identify patterns and generate hypotheses. Real-time interaction, such as reordering axes or filtering data points, makes large datasets more manageable and supports discovery.

The early groundwork for interactive PCPs was laid by Wegman (1990), who introduced fundamental techniques for multidimensional data interaction. Building on this foundation, Siirtola and Räihä (2006) implemented brushing and linking to improve dimensional analysis, allowing for easier comparisons across attributes in large datasets. Inselberg (2009) further expanded these tools, introducing dynamic filtering and axis rearrangement, enabling users to tailor PCP visualizations to their needs.

While interactivity enhances insight, it can also introduce challenges. Increased complexity can overwhelm novice users and demand higher computational resources, potentially slowing analysis in high-dimensional datasets. As Inselberg (2009) noted, the cognitive load of multi-step filtering and interaction can lead to user fatigue. Addressing this balance requires careful interface design to ensure that interactive features remain accessible and intuitive.

Modern implementations, such as the live parallel coordinate plots on Plotly's ggplot2 platform, demonstrate how interactivity can streamline data analysis. Users can filter and separate data ranges by moving their mouse across parallel axes, with tooltips providing detailed information about individual data points. Drag-and-drop axis reordering allows users to adjust dimensions to uncover hidden patterns and correlations. Brushing options dynamically emphasize relevant data paths, making trends and outliers visible while downplaying less critical information ("Parallel Coordinates Plot in Ggplot2," n.d.). Interactive Parallel Coordinate Plot

Interactive PCPs continue to evolve, offering increasingly intuitive and adaptable methods for high-dimensional data exploration. By combining dynamic features like axis reordering, filtering, and brushing, PCPs enable users to extract meaningful insights from complex datasets. However, balancing usability with computational demands and minimizing user fatigue remain ongoing challenges. Future developments will likely focus on enhancing user interfaces and optimizing algorithms to make interactive PCPs more efficient and accessible for a broader range of users and applications.

Reordering and Axis Flipping

Adaptive reordering and axis flipping are essential techniques for simplifying the analysis of multidimensional relationships in parallel coordinate plots (PCPs). Reordering aligns highly correlated or thematically related dimensions, minimizing intersections and reducing visual clutter. When implemented effectively, it transforms PCPs from a tangled mass of lines into a clear roadmap of relationships.

Axis flipping complements reordering by addressing negative correlations between adjacent dimensions. Inverting the scale of one or more axes eliminates unnecessary line crossings, revealing patterns that might otherwise remain hidden. Together, these methods improve the interpretability of complex datasets, making relationships and trends across dimensions more discernible.

Early works, such as Inselberg and Dimsdale (1990), proposed axis reordering to enhance the interpretability of PCPs, especially for datasets with highly interrelated variables. Building on this foundation, LeBlanc, Mellor-Crummey, and Fowler (1990) introduced correlation-based reordering, which positions axes to align more related dimensions and reduce visual noise. Later advancements by Peng et al. (2004) incorporated automatic reordering algorithms that adjust axis positions and flip orientations based on user-defined weights, providing greater flexibility for tailoring PCPs to specific analytical needs.

Dynamic reordering techniques optimize the alignment of correlated variables, enabling more precise identification of associations and trends (Yuan et al. 2009). By grouping linked variables and separating unrelated ones, reordering reduces the visual noise caused by crossings between weakly connected variables. Researchers Johansson et al. (2005) and Wegman and Luo (1997) found that well-designed reordering algorithms reveal patterns, clusters, and outliers more effectively while mitigating overplotting. These methods are particularly valuable for analyzing datasets with numerous dimensions, where clear visual grouping can facilitate discovery.

Axis flipping addresses challenges associated with negative correlations by inverting the scale of one or more axes, reducing unnecessary line crossings, and improving clarity. This adjustment is beneficial when adjacent axes contain negatively correlated factors, revealing trends and relationships that standard PCPs might obscure. Automated axis flipping algorithms, such as those explored by Dasgupta and Kosara (2010), dynamically adapt plots based on data properties, reducing the cognitive burden on users while improving interpretability.

Flipping axes enhance the visual representation of negative correlations and help uncover subtle patterns in datasets with unrelated or oppositely related variables. By turning “problematic” relationships into familiar shapes, axis flipping clarifies connections and aids in exploring complex datasets. However, frequent flipping can disrupt the analytical flow, as users may need help maintaining a stable mental model of the data structure (Julian Heinrich and Weiskopf 2013b).

While automatic reordering and axis flipping greatly enhance PCP readability, they can sometimes misalign dimensions critical to specific analytical questions. LeBlanc et al. cautioned that automated ordering might sacrifice user intent by prioritizing generic readability over context-specific relevance. Continuous reconfiguration of axes can also introduce inconsistency, challenging users to track relationships as layouts change dynamically. These methods offer limited benefits for datasets with minimal correlations, as reordering may not significantly clarify insights.

Reordering and axis flipping should balance automation with user-defined parameters to address these concerns. Allowing users to influence axis arrangement ensures the visualization aligns with their analytical goals. Tools like user-adjustable weights for reordering algorithms and toggle options for axis flipping provide this flexibility, enabling users to tailor PCPs to their needs without compromising clarity or interpretability.

Reordering and axis flipping are complementary techniques that address overplotting in PCPs. Reordering aligns correlated variables to reduce crossings while flipping axes minimizes interference in negatively correlated dimensions. When used together, these methods create a cleaner, more interpretable visualization, making it easier to detect patterns and explore relationships in large datasets (Healey and Enns 1999).

Studies by Julian Heinrich and Weiskopf (2013a) show that combining reordering and flipping significantly enhances the accuracy and usability of PCPs, improving their ability to reveal groups, trends, and outliers. These advancements are critical for data analysts, enabling more informed decision-making based on visual insights. However, the potential for disruption caused by continuous reordering highlights the importance of thoughtful implementation. Ensuring stability and clarity in PCPs requires balancing dynamic adjustments with consistency, helping users build a reliable mental model of the data.

By integrating these techniques, PCPs become powerful tools for exploring and understanding multidimensional data, providing analysts the flexibility and precision to uncover meaningful insights.

In Figure 14, the parallel coordinate plot enhances interpretability by rearranging variables and flipping axes to highlight patterns and connections between penguin species.

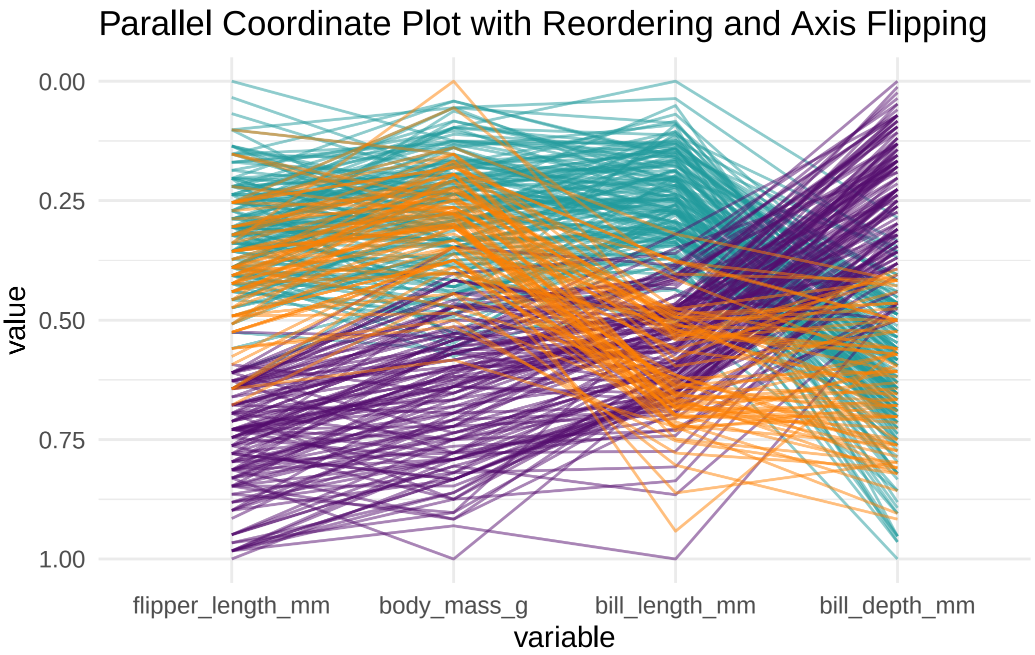

Figure 14: Reordering and Axis Flipping in PCP

The visualization brings species-based trends into sharper focus by placing related factors, such as body mass and flipper length, closer together and flipping specific axes. For example, the Gentoo (purple) species exhibits a pattern of shallower bills paired with larger bodies and longer flippers. In contrast, the Adelie (pale blue) species displays opposite traits.

This rearrangement reduces visual clutter caused by unrelated variables being placed adjacent, making inter-variable trends more discernible. However, overplotting remains a challenge, particularly in the middle ranges of each axis where values converge. The dense overlap makes distinguishing individual lines or peaks difficult, complicating the identification of subtle trends or outliers.

While the exact distribution shapes of the variables are not explicitly visible, the density and spread of lines along each axis provide a clear impression of the range of values for each species. The rearranged order improves the visibility of inter-variable relationships, revealing species-based correlations in measurement trends that were previously obscured. For instance, the changes between adjacent axes emphasize how species cluster around specific measurement ranges, enabling a clearer understanding of their distinguishing features.

Despite these improvements, individual relationships between data points still need to be discerned due to persistent line crossings. This limitation underscores the trade-offs of PCPs: while reordering and axis flipping enhance clarity at the macro level by exposing group-level patterns and trends, they are less effective at resolving micro-level details or distinguishing outliers within dense clusters.

The figure demonstrates the power of these adaptive techniques to make species-based correlations more apparent, even as challenges in visualizing fine-grained details persist. Combining these strategies with additional methods, such as filtering or clustering, further enhances the utility of PCPs for analyzing complex datasets.

Handling Ties in PCPs

Adding distance to visual ties has evolved significantly alongside advancements in data visualization, mainly as the need to balance detail and clarity in large datasets has grown. Early visualization methods, especially in network analysis and multidimensional scaling, relied heavily on proximity and alignment to convey patterns and relationships. These techniques were effective for smaller datasets but struggled to scale as the complexity and volume of data increased.

Visual clutter emerged as a significant challenge, particularly in parallel coordinate plots (PCPs), where overlapping lines and intersections often obscured underlying data structures. As PCPs became more common for high-dimensional data analysis, the density of lines made it increasingly difficult for users to discern meaningful trends and relationships. To address this, researchers began developing techniques to increase the "visual distance" between elements in these visualizations, reducing overplotting while preserving the continuity of the visualization data.

Early studies of PCPs highlighted the importance of visual distance in aiding cognitive interpretation. Chang, Dwyer, and Marriott (2018a) compared PCPs and scatterplot matrices (SPLOMs), observing that users struggled to identify correlations in PCPs due to the complexity of cross-line patterns. They noted, "In PCP, participants had to reorder the axes to examine all dimensions, whereas in SPLOM, it is straightforward to identify correlation." This underscores how increasing visual distance between overlapping elements can simplify cognitive grouping without fragmenting the data.

Researchers have explored methods to enhance visual distance, including line bundling, spacing adjustments, and axis reordering. These approaches create separation between lines and intersections, allowing individual paths to emerge more distinctly. For instance, in PCPs, early experiments showed that increasing visual distance improved the clarity of intersections, enabling users to trace relationships across dimensions better and understand the overall structure of the data.

One foundational technique that has influenced the development of visual distance in PCPs is the "nested-means" approach introduced by D. Guo et al. (2005). This method creates a more organized and interpretable visualization by reducing overlapping lines and ensuring that the mean value of each variable is consistently placed at the midpoint of its axis. D. Guo et al. (2005) noted that this approach not only addresses overplotting issues but also preserves the central tendencies of the data, enhancing both visual clarity and analytical utility.

Line bundling, another key advancement, groups similar paths together, reducing the visual chaos caused by overlapping lines. While this method simplifies the overall visualization, it introduces a trade-off by obscuring some outliers or unique paths. Combining bundling with line spacing adjustments or transparency techniques helps mitigate this limitation, ensuring that critical data points remain visible while reducing clutter.

These methods collectively illustrate how visual distance is a spatial and cognitive tool that enables users to navigate increasingly complex datasets more efficiently. By strategically separating visual elements, researchers can guide users' attention to key patterns and trends without disrupting the continuity of the data.

As datasets grow more intricate, the role of visual distance has expanded beyond PCPs to encompass a wide range of visualization techniques. From scatterplots to network graphs, methods that enhance visual distance are critical in reducing clutter and improving interpretability. For example, in modern PCPs, combining visual distance with dynamic interactivity—such as brushing, filtering, and axis flipping—further enhances users' ability to explore relationships across dimensions.

Today, visual distance is recognized as an essential component of effective data visualization. It reduces perceptual overload and facilitates cognitive processing, helping analysts uncover insights in even the most complex datasets. By applying lessons learned from early studies and continuing to refine these techniques, modern visualizations can strike the delicate balance between detail and clarity, empowering users to make sense of their data confidently.

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. (2020) 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.

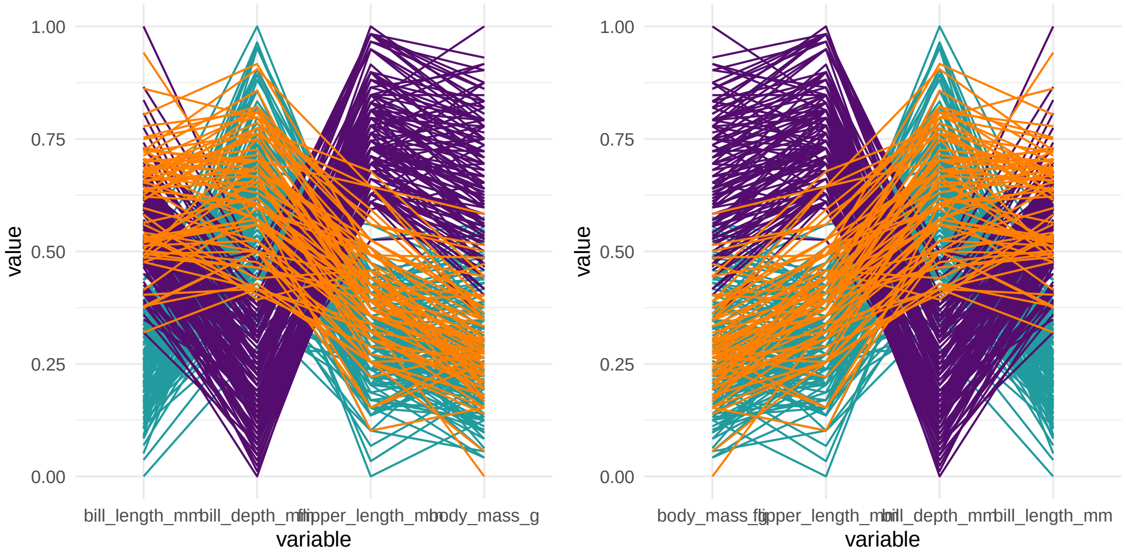

Figure 15: PCP Without Axis Reordering

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, referred to as Quantized Generalized Parallel Coordinate Plots (QGPCP), is a technique that groups close-together values into bins, reducing visual clutter and improving the readability of tightly packed data. This method is particularly effective for managing large datasets, where overplotting and excessive line crossings obscure meaningful trends and patterns.

By aggregating similar values into quantized bins, QGPCP simplifies visualizations without significantly altering the overall data structure. As Moustafa explains, "By quantizing close data points, we can minimize overplotting and improve discernment of overall trends." This highlights the trade-off inherent in quantization: sacrificing granular detail to enhance interpretability. Tests on real-world data have demonstrated that this approach preserves important patterns while reducing visual noise, making it an invaluable tool for exploring high-dimensional datasets.

Quantization excels in scenarios where high-dimensional datasets result in extensive line overlap. By reducing the number of distinct lines in a parallel coordinate plot (PCP), this method creates a cleaner, more navigable visualization. Key benefits of quantization include:

- Improved Clarity: Aggregating close values minimizes line crossings, enabling users to discern trends more quickly. Patterns previously obscured by overplotting become more apparent, enhancing the user's ability to interpret the data.
- Noise Reduction: By removing small-scale variations, quantization reduces the visual complexity of the plot. This makes it easier for analysts to focus on larger trends and relationships across dimensions.
- Scalability: Quantization is particularly effective for large datasets, where traditional visualization methods struggle to handle the volume of overlapping data points. Aggregating values optimizes visualizations for datasets with high-dimensional complexity.

While quantization simplifies visualizations, it has limitations that make it less suitable for specific analytical contexts. Chief among these is the potential loss of detail. Aggregating values into bins can obscure small but meaningful differences, which may be critical in scenarios requiring precision. For example:

- Loss of Granularity: By grouping data points, quantization may mask outliers or subtle variations that provide valuable insights.
- Bias in Representation: The choice of bin size and boundaries can influence the visualization, potentially introducing artifacts or skewing interpretations.

- Limited Utility for Precision Tasks: Quantization may oversimplify the visualization for analyses that demand exact data representation, such as identifying specific outliers or subtle correlations.

As Moustafa notes (Rida EA Moustafa 2009), the method's strength lies in its ability to simplify densely packed datasets, which comes at the cost of fine-grained detail. Thus, the technique is best suited for exploratory analyses where the goal is to identify broad patterns rather than focus on precise data points.

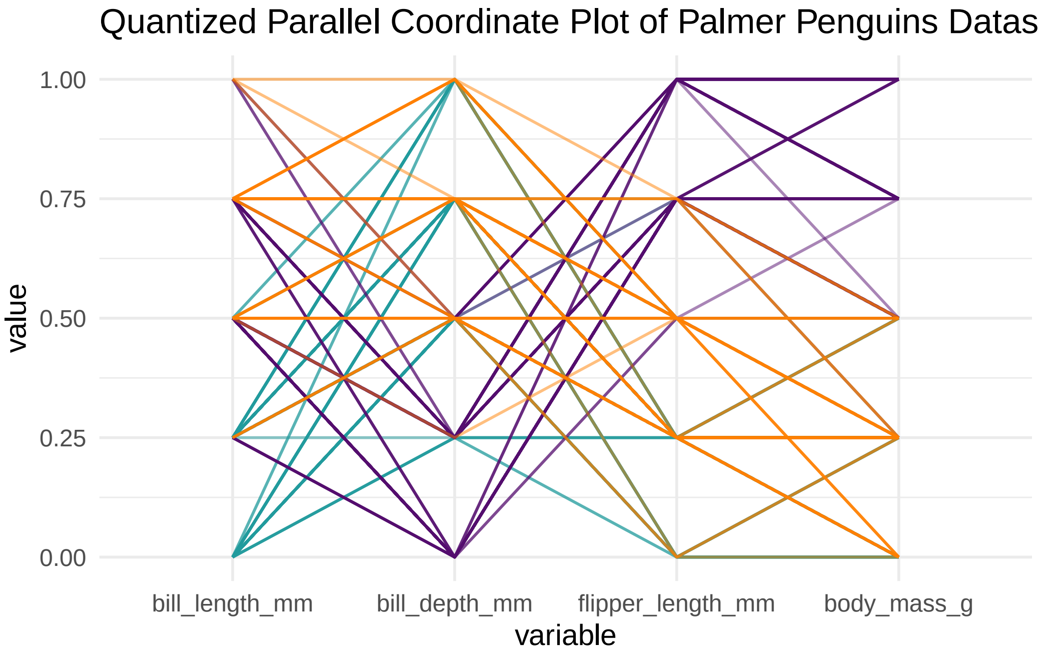

Figure 16: Quantized Generalized Parallel Coordinate Plots

In Figure 16, 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.

Quantization, or QGPCP, is a powerful method for simplifying dense datasets in parallel coordinate plots and similar visualizations. Grouping close-together values into bins minimizes overplotting, reduces noise, and highlights overarching trends. While it sacrifices some granularity, the technique is highly effective for exploratory data analysis and high-dimensional datasets where clarity is important. When applied thoughtfully, quantization balances detail and readability, making it a valuable tool for modern data visualization.

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, Kim, and Seo (2020), 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. 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, Dwyer, and Marriott (2018b), 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. In areas with many people, it could be useful because overlapped lines can make it harder to see.

Handling Numerical Ties

The challenge of managing tied values in statistical data analysis has long perplexed experts striving for clarity in visualizations. Traditionally, statistical approaches relied on rank-based methods or binary data handling, which often struggled with tied observations. These ties could distort distributions or complicate their interpretation. Early solutions involved modifying ranking systems or introducing tie-breaking rules, but these adjustments frequently sacrificed accuracy or interpretability.

As graphical data analysis grew, pioneers such as Tukey, Chambers, and Cleveland developed innovative approaches to handle tied values visually. Techniques such as gradually increasing the spacing between tied observations have emerged, enabling more accurate and comprehensible visualizations without altering the underlying data structure. These advancements underscore the importance of visualization for effectively addressing long-standing challenges in statistical analysis.

Tufte (2001)'s seminal work, The Visual Display of Quantitative Information, highlights the detrimental impact of visual noise—any element that crowds or complicates the visual field—on data interpretation. Tufte's "graphical excellence" principle emphasizes designing visualizations that present data as clearly and efficiently as possible. He argues that separating tied values in complex datasets can significantly reduce overlapping lines and visual clutter, enhancing the viewer's ability to accurately identify patterns and compare variables.

In line with Tufte's philosophy, distancing tied values in dense visualizations, such as parallel coordinate plots (PCPs), reduces interpretative difficulty. By decreasing visual noise, users can more rapidly and effectively perceive quantitative relationships, leading to more precise insights from the data.

In complex visualizations like PCPs, tied values often cause overlapping lines that obscure other data points, masking trends, and outliers. Tufte describes excessive clutter as "chartjunk," unnecessary visual elements that hinder comprehension. He emphasizes that even small reductions in visual noise can substantially improve a viewer's ability to process information quickly and accurately.

For example, introducing slight separations between tied values can make patterns and relationships more discernible in large datasets where similar values cluster tightly. This adjustment reduces cognitive load, allowing viewers to extract meaningful insights without effort. By addressing ties visually rather than through statistical manipulation, modern visualization techniques preserve data integrity while improving clarity, aligning with Tufte's vision of effective data presentation.

Jittering Points

Random Jittering of Data Points

Random jittering is a widely used technique in data visualization, where small random values are added to tied observations to prevent excessive density and improve clarity. This method is particularly effective in scatter plots, dot plots, and other visualizations where overlapping data points can obscure patterns and relationships. By subtly adjusting the position of tied values, random jittering ensures that individual points remain visible, making the overall distribution easier to interpret without altering the underlying data.

First introduced in Graphical Methods for Data Analysis (Chambers 1983), random jittering is a simple yet powerful tool for managing overlaps in data visualizations. Typically, the added values are drawn from a uniform or normal distribution, with the range or standard deviation carefully chosen to maintain the integrity of the dataset. This approach preserves the overall shape of the data distribution while addressing visual clutter.

Experiments have shown that jittering significantly enhances the readability of visualizations, especially when overlaps are minimal or occur infrequently. It is beneficial for datasets with many tied values, where dense clustering can make it difficult to discern individual data points. In these scenarios, jittering highlights trends and patterns without compromising the viewer's ability to understand the broader dataset.

However, random jittering has its limitations. One notable drawback is the potential loss of precision, as the exact values of tied observations become obscured. If the jittering distance is too large, it may create an impression of variability where none exists, misleading viewers. To mitigate this issue, the amount of jitter should be kept minimal and proportional to the scale of the data.

In practice, the effectiveness of jittering depends on the nature of the dataset and the visualization context. For example, small-scale jittering can separate overlapping points in scatter plots where points represent discrete categories or measurements without distorting the categorical structure. In contrast, jittering must be applied cautiously for continuous data to avoid introducing apparent trends or deviations.

Additionally, the choice of distribution for generating jitter values plays a crucial role. Uniform distributions are often preferred for their simplicity and predictability, ensuring consistent point spacing. On the other hand, normal distributions can mimic natural variations, which may be more suitable for datasets representing real-world phenomena. In either case, transparency and precise documentation of the jittering process are essential to maintain trust and interpretability in the visualization.

Random jittering is a valuable tool for improving the clarity of data visualizations, particularly in scenarios where tied observations lead to visual clutter. By carefully balancing the amount and distribution of jitter, analysts can preserve the integrity of the data while enhancing its interpretability. While not without limitations, when used judiciously, jittering can reveal hidden patterns, reduce cognitive load, and provide a clearer picture of the underlying data, making it an indispensable technique in the modern data visualization toolkit.

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.

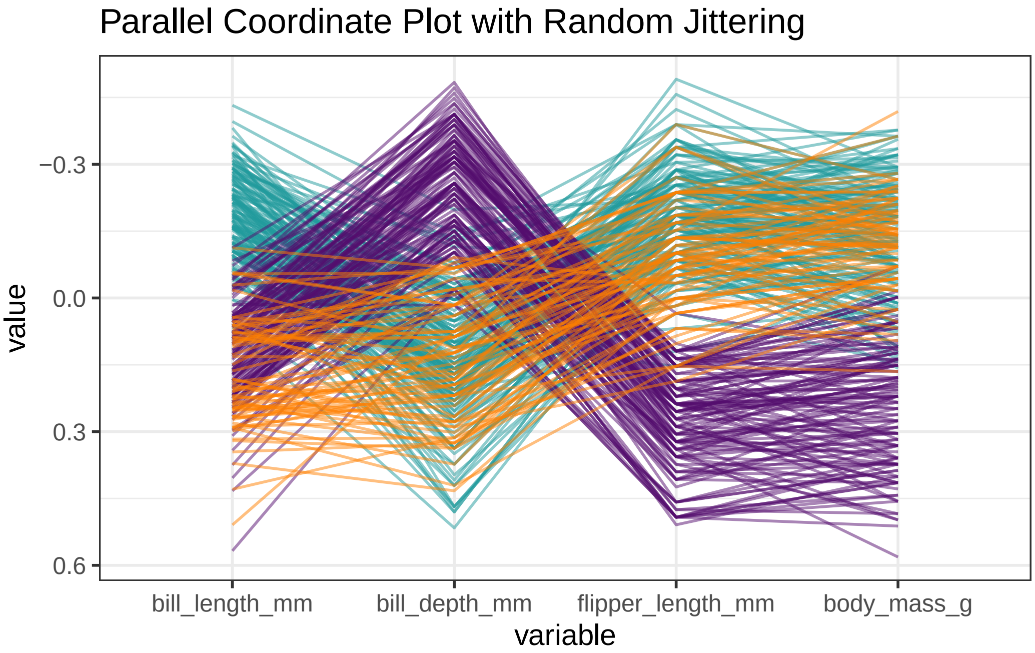

Figure 17: Random Jittering of Data Points in PCP

Rank Jittering (Rank Adjustment)

Rank jittering is an effective technique used primarily in non-parametric analyses to address the challenges tied values pose. Instead of modifying the data values, this method adjusts the ranks of tied observations, making it easier to discern ordinal relationships in visualizations and rank-based tests such as the Wilcoxon signed-rank test. By preserving the ordinal nature of the data, rank jittering maintains the integrity of non-parametric analyses while reducing the impact of ties on interpretability.

As detailed in Conover (1999)'s Practical Nonparametric Statistics, rank jittering retains the ordinal structure of the dataset without significantly altering the underlying distribution. This ensures that results from rank-based analyses remain valid and reliable. The primary advantage of rank jittering lies in its ability to maintain rank-based interpretations while minimizing tied observations' visual or analytical impact. This is especially beneficial in statistical contexts where the rank order carries critical meaning.

Rank jittering involves adding a slight, random, or systematic adjustment to the ranks of tied observations, effectively breaking ties without altering the data values. The adjustments can be random (introducing variability) or systematic (maintaining proportionality among tied ranks). This technique is commonly applied to datasets undergoing rank-based testing, ensuring more straightforward visualizations and more precise analyses.

One systematic approach to rank jittering leverages binary relationships within the data to create proportional separations among tied ranks. As discussed by Ipkovich, Héberger, and Abonyi (2021), rank-based separation distributes tied observations evenly while maintaining interpretability and reducing visual clutter. This method is beneficial in parallel coordinate plots (PCPs), where overlapping lines from tied values often obscure trends and patterns. By assigning spacing factors based on rank differences, rank jittering effectively spreads overlapped lines, enhancing clarity and readability.

The key strength of rank jittering lies in its ability to reduce visual and analytical noise while preserving the order and interpretability of the data. It is especially well-suited for categorical or ordinal datasets where maintaining order is critical. For example, in PCPs or other rank-based visualizations, rank jittering highlights relationships and reduces clutter, making it easier to identify trends or categories.

However, rank jittering also presents challenges. Computational complexity increases with the size of the dataset, as handling large numbers of tied values requires significant processing power. Additionally, excessive adjustments to ranks may inadvertently alter test significance levels, potentially affecting the outcomes of non-parametric analyses. This highlights the importance of carefully calibrating the extent of jitter applied, particularly in large datasets with many ties.

Rank jittering is particularly effective in scenarios where maintaining the ordinal nature of data is essential, such as non-parametric statistical tests or visualizations of categorical and ordinal data. Its application is most valuable when:

- Non-parametric analyses require clear distinctions among ranks to ensure accurate results (e.g., Wilcoxon signed-rank test).
- Visual clarity is needed in datasets with overlapping ranks, such as PCPs or rank-based heatmaps.
- Order preservation is critical to the interpretability of the data, particularly in ordinal or categorical contexts.

Despite its advantages, rank jittering should be used judiciously. Careful calibration of adjustments is necessary to balance the benefits of reduced clutter and improved interpretability against the risks of computational inefficiency and potential distortions in statistical results.

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,
- $\delta_i \sim \mathcal{N}(0, \sigma_{rank}^2)$
- σ_{rank} are small values ensuring that the adjustment is minor.

Rank jittering is a powerful tool for addressing ties in non-parametric analyses and visualizations. This method enhances clarity and interpretability by preserving the ordinal structure of data while reducing the impact of ties. Though computationally intensive for large datasets, its ability to maintain rank-based relationships makes it indispensable for categorical or ordinal data applications. When applied carefully, rank jittering provides a robust solution for resolving ties without compromising the integrity of the data or analysis.

Deterministic Jittering (Fixed Perturbation)

Deterministic jittering offers a structured and repeatable solution for addressing tied observations in data visualizations. Unlike random jittering, which relies on chance, deterministic jittering adds a small, fixed value (epsilon) to each tied observation. This predictable adjustment ensures that each data point is visually separated while maintaining the integrity of the dataset. As detailed in David and Tukey (1977)'s Exploratory Data Analysis, deterministic jittering is particularly valuable in scenarios where random

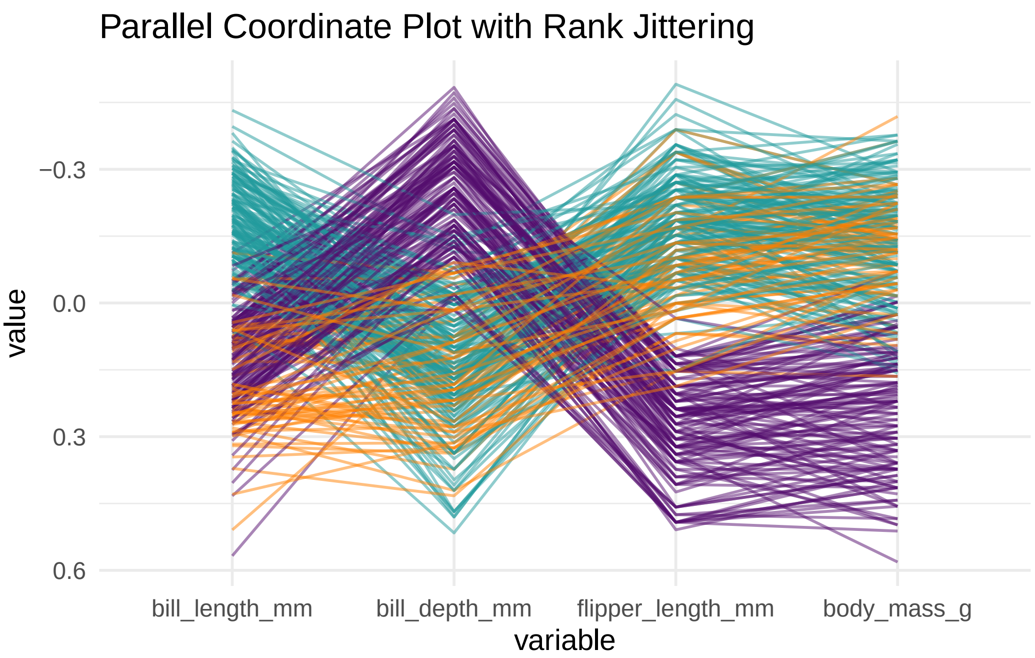

Figure 18: Rank Jittering of Data Points in PCP

noise is undesirable, as it guarantees that the process is transparent, replicable, and free from stochastic variability.

Deterministic jittering is more straightforward to understand and control than random jittering, making it an appealing choice for analysts who prioritize clarity and consistency. However, it must be applied cautiously, especially in scaled datasets, to avoid introducing artificial patterns or misleading trends. Properly calibrated, deterministic jittering can enhance visual clarity without compromising the statistical structure of the data.

Jittering, whether random or deterministic, is a common technique for managing ties in visualizations like parallel coordinate plots (PCPs). In PCPs, tied values often cause overlapping lines, obscuring patterns and making the data harder to interpret. By adding small amounts of noise, jittering creates a visual separation between tied observations, improving clarity and facilitating analysis.

Swayne et al. (2003) highlights jittering as an intuitive and effective way to handle ties in 2-D correlation tours, where dense overlapping values often distort visualizations. They note that jittering separates tied values without distorting their underlying statistical relationships, preserving the dataset's integrity. Similarly, Few and Edge (2008) describes jittering as a practical response to the perceptual challenges caused by overplotting. By introducing subtle adjustments to tied values, jittering highlights individual lines or points in dense datasets while maintaining the overall structure of the visualization.

While jittering enhances clarity, its effectiveness depends on selecting the appropriate amount of noise. Excessive jittering can introduce artificial variability, potentially misleading viewers, while insufficient jittering may fail to separate tied observations adequately. Deterministic jittering mitigates this issue by providing a fixed, controlled adjustment that can be calibrated to balance clarity with fidelity.

In their review, Few and Edge (2008) emphasize the importance of regulating jittering to avoid compromising the visualization's accuracy. The careful application ensures that jittering improves visual clarity without introducing distortions. This approach is especially critical in dense datasets, where the balance between noise and structure can significantly affect interpretability.

Deterministic jittering's predictability makes it a preferred choice in applications where transparency and repeatability are essential. For example, in datasets requiring precise replication of visualizations, deterministic jittering ensures that tied observations are consistently adjusted. In contrast, random jittering is better suited for exploratory analyses, where variability is acceptable and often necessary to reveal hidden patterns.

Both methods have their strengths and limitations. Deterministic jittering excels in creating structured, predictable visualizations but risks introducing artificial patterns if overapplied. While more flexible, random jittering may obscure the exact values of tied observations if the added noise is excessive or inconsistent. The choice between the two depends on the specific analytical goals and the dataset's characteristics characteristics.

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.

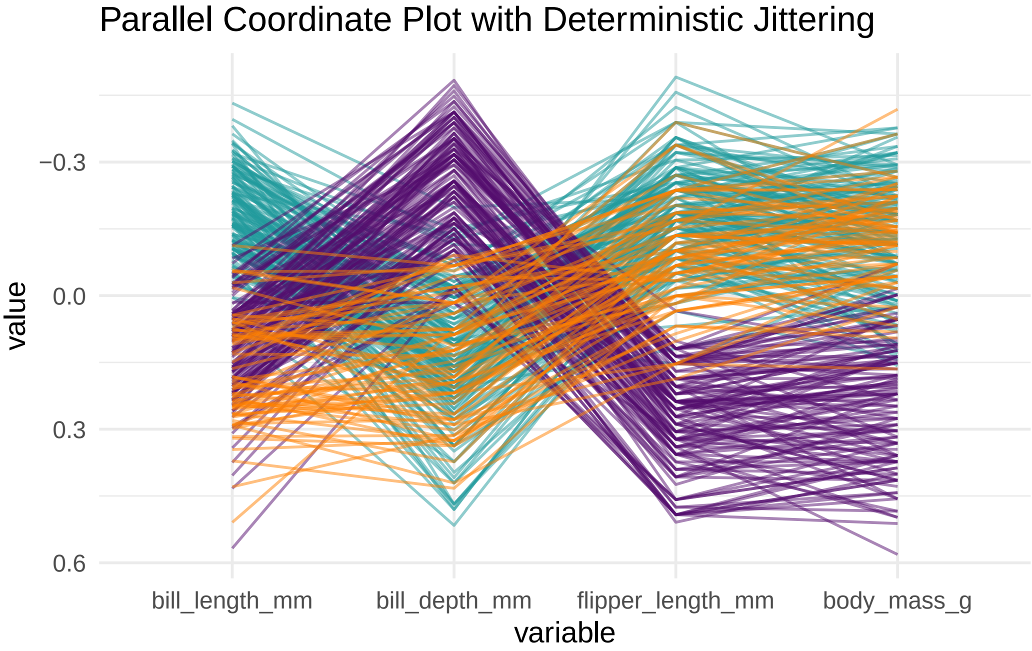

Figure 19: Deterministic Jittering of Data Points in PCP

Whether random or deterministic, jittering is an indispensable tool for managing ties in data visualizations. Deterministic jittering stands out for its clarity, repeatability, and ability to preserve the dataset's statistical structure. When applied carefully, it ensures that tied observations are visually distinct without introducing artificial patterns or distortions. For dense visualizations like PCPs, jittering enhances interpretability and reduces the challenges posed by overplotting. By balancing noise and structure, deterministic jittering provides a reliable and effective solution for creating clear and insightful data visualizations.

Mean or Median Splitting

Mean or median splitting is another widely used method for resolving ties in data analysis and visualization. This approach replaces tied values with their mean or median, effectively centering the tied observations on their central tendency. This technique works particularly well for symmetric distributions, where the mean and median closely align, and for visualizations that emphasize central values, such as box plots.

As detailed in Wilcox (2017)'s Modern Statistics for the Social and Behavioral Sciences, mean or median splitting provides a balanced and fair representation of tied data, mainly when used in boxplot visualizations. Assigning tied observations to their central values ensures that the visual representation reflects the dataset's core tendencies without distorting its overall structure.

This method is especially well-suited for visualizations prioritizing central tendencies, such as box plots or histograms. By centering tied values, mean or median splitting emphasizes the median line or the mean's location, enhancing the clarity of the visualization. It is beneficial for datasets with minimal ties, as it preserves the integrity of the distribution while reducing visual clutter caused by overlapping points or lines.

Mean or median splitting is advantageous in cases where the tied group size is small, as it minimizes the impact on the range and variability of the dataset. For example, replacing tied values with their mean or median preserves the dataset's balance in symmetric distributions, ensuring the visualization remains accurate and interpretable.

Despite its strengths, mean or median splitting has limitations, particularly for datasets with numerous ties. Excessive use of this method can obscure the variability within tied groups, creating a homogenized view that may not accurately reflect the diversity of the data. This can be problematic when understanding the range of tied observations is critical, such as in scatter plots or other point-based visualizations.

Moreover, mean or median splitting assumes that the central value represents the tied group, which may not hold for skewed distributions. In these cases, using the mean or median may introduce bias, potentially distorting the interpretation of the data. Therefore, it is essential to consider the dataset's characteristics when applying this technique carefully.

In practice, mean or median splitting involves identifying groups of tied values and replacing each observation with the mean or median of the group. For example, this technique can emphasize the data's central line (median) or central tendency (mean) in a box plot, reducing ambiguity caused by overlapping points. The simplicity of implementation and its alignment with central tendency measures make this method a popular choice for summarizing tied data in visualizations.

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:

$$x'_i = x_{mean} \text{ or } x'_i = x_{median}$$

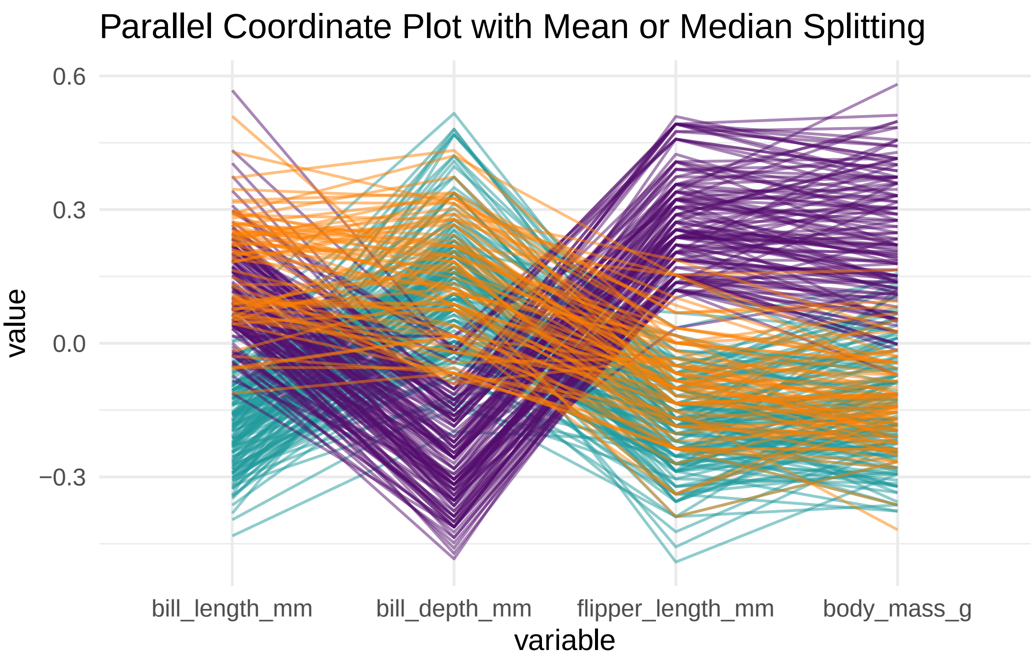

Figure 20: Mean or Median Splitting of Data Points in PCP

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

Mean or median splitting offers a centered and balanced approach to handling ties, making it ideal for symmetric distributions and visualizations focused on central tendencies. While it enhances clarity in datasets with few ties, its limitations in preserving variability and representing skewed distributions highlight the importance of careful application. By centering tied observations on their mean or median, this technique provides a straightforward solution for improving the interpretability of tied data, particularly in visualizations like box plots that prioritize centrality.

Kernel Density Estimation (KDE) with Bandwidth Adjustment

Kernel Density Estimation (KDE) is a powerful method for creating smooth density visualizations by spreading overlapping observations. By assigning slightly different weights to tied values, KDE maintains the original data while providing a more continuous and visually interpretable representation of the distribution. This technique is especially effective for multimodal distributions, where tied observations can obscure peaks or create misleading impressions of uniformity.

As discussed in Silverman (2018)'s Density Estimation for Statistics and Data Analysis, adjusting the bandwidth in KDE enables an accurate depiction of multimodal distributions, reducing errors caused by ties. Bandwidth selection is critical. Determining the degree of smoothing applied to the data and choosing an appropriate bandwidth balance the need to spread out tied values while preserving the distribution's key features. Incorrect bandwidth choices, particularly in datasets with tightly clustered values, can lead to over-smoothing or under-smoothing, distorting the visualization. KDE excels in density plots, offering a visually intuitive way to represent the data's underlying distribution. However, its reliance on careful bandwidth tuning requires expertise, especially for datasets with high variability or dense clusters.

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.

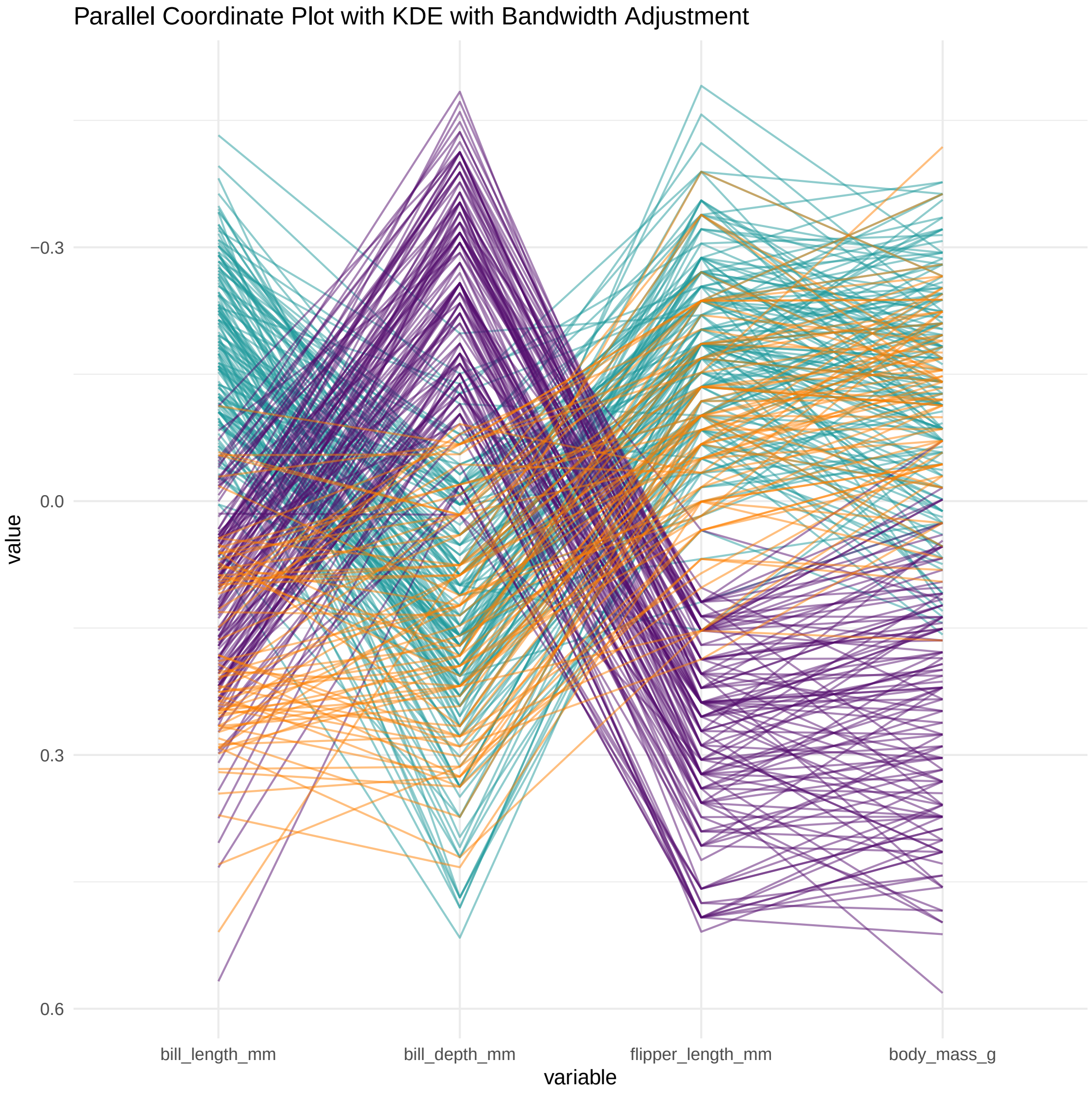

Figure 21: KDE with Bandwidth Adjustment of Data Points in PCP

Depth Cues

Traditional methods for handling ties often need more effective high-dimensional visualizations, such as parallel coordinate plots (PCPs), as overlapping lines obscure trends and connections. Advanced visual techniques, such as Depth Cue Parallel Coordinates (DCPC), address this challenge by introducing depth perception cues to create a three-dimensional effect.

As demonstrated in Johansson, Ljung, and Cooper (2007)'s work, DCPC combines line transparency and brightness adjustments to layer information visually. Temporal binning and perception-based coloring add further clarity, enabling users to distinguish overlapping trends effectively. These techniques are designed to handle temporal data with high dimensionality, where conventional visualization methods struggle to reveal patterns.

DCPC preserves the original dataset by ensuring each line remains identifiable, even in densely populated plots. While this method significantly enhances clarity, it requires advanced rendering tools, which may only be available in some visualization settings. The computational complexity of generating depth cues and transparency effects can pose challenges, especially for large datasets.

Different techniques for handling ties excel in specific contexts, depending on the type of visualization and the nature of the data:

- Jittering (random or deterministic): Ideal for scatterplots and similar two-dimensional visualizations, jittering separates tied values by adding small random or fixed displacements. This approach effectively reveals hidden patterns and reduces visual clutter in dense datasets.
- KDE and rank-based methods: are best suited for density plots and ordinal visualizations. KDE spreads out tied values smoothly, preserving the overall distribution, while rank-based methods emphasize ordinal relationships without distorting the data.
- DCPC and advanced PCP techniques: For multidimensional data, especially in PCPs, traditional methods often fail to address overlapping lines adequately. DCPC introduces visual depth and transparency to resolve these issues, providing clarity in complex datasets.

While existing methods improve clarity in specific visualization contexts, PCPs present unique challenges due to their high-dimensional nature and the tendency for lines to overlap extensively. Traditional techniques like jittering and KDE, designed for two-dimensional visualizations, only sometimes scale effectively to PCPs.

A promising solution involves systematically introducing spacing between tied values across multiple axes in PCPs. This method would distribute tied observations more evenly, reducing overlaps and revealing otherwise obscured patterns. Researchers could uncover trends and connections by designing this approach to account for relationships across dimensions while preserving the dataset's structure and integrity.

Such a method could combine rank-based adjustments, deterministic jittering, and transparency techniques to create a more precise and insightful visualization of high-dimensional data. Incorporating this approach into PCPs would significantly enhance their utility for exploring complex, multidimensional datasets.

Each method for handling ties—whether jittering, KDE, rank-based adjustments, or depth cueing—has unique strengths tailored to specific visualization needs. While jittering and KDE are effective for scatterplots and density plots, advanced techniques like DCPC are better suited for high-dimensional visualizations like PCPs. Developing a dedicated method for handling ties in PCPs, such as planned spacing across axes, could further advance the field, enabling researchers to explore complex data more effectively while maintaining visual clarity and data integrity. This innovation would provide a transformative step in high-dimensional data visualization, making uncovering trends and connections across diverse datasets easier.

References

Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." *WIREs Computational Statistics* 2 (4): 433–59. https://doi.org/10.1002/wics.101.

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.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. "D^3 Data-Driven Documents." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2301–9.

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

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

———. 2018a. "An Evaluation of Perceptually Complementary Views for Multivariate Data." In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, 201. IEEE.

Claessen, Jarry HT, and Jarke J Van Wijk. 2011. "Flexible Linked Axes for Multivariate Data Visualization." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2310–16.

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.

Dennig, Frederik L, Maximilian T Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A Keim, and Evanthia Dimara. 2021. "Parsetgnostics: Quality Metrics for Parallel Sets." In *Computer Graphics Forum*, 40:375–86. 3. Wiley Online Library.

Few, Stephen, and Perceptual Edge. 2008. "Solutions to the Problem of over-Plotting in Graphs." *Visual Business Intelligence Newsletter*.

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.

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, Diansheng, Mark Gahegan, Alan M MacEachren, and Biliang Zhou. 2005. "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach." *Cartography and Geographic Information Science* 32 (2): 122.

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.

Healey, Christopher G, and James T Enns. 1999. "Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization." *IEEE Transactions on Visualization and Computer Graphics* 5 (2): 145–67.

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." *Eurographics (State of the Art Reports)*, 95–116.

———. 2013b. "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.

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.

Hofmann, Heike, and Marie Vendettuoli. 2013. "Common Angle Plots as Perception-True Visualizations of Categorical Associations." *IEEE Transactions on Visualization and Computer Graphics* 19 (12): 2297–2305. https://doi.org/10.1109/TVCG.2013.140.

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.

Ipkovich, Ádám, Károly Héberger, and János Abonyi. 2021. "Comprehensible Visualization of Multidimensional Data: Sum of Ranking Differences-Based Parallel Coordinates." *Mathematics* 9 (24): 3203.

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, and Matthew Cooper. 2007. "Depth Cues and Density in Temporal Parallel Coordinates." In *EuroVis*, 7:35–42.

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.

Kosara, Robert, Fabian Bendix, and Helwig Hauser. 2006. "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data." *IEEE Transactions on Visualization and Computer Graphics* 12 (4): 558–68.

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.

Munzner, Tamara. 2014. *Visualization Analysis and Design*. CRC press.

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. 2012a. "Comparing Clusterings Using Bertin's Idea." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2506–15.

———. 2012b. "Comparing Clusterings Using Bertin's Idea." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2510.

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.

Rosvall, Martin, and Carl T Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105 (4): 1118–23.

Schonlau, Matthias. 2003. "Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots." In *Proceedings of the Joint Statistical Meetings, Section on Statistical Graphics*. American Statistical Association. https://schonlau.net/

publication/03jsm_hammockplot.pdf.

———. 2024. "Hammock Plots: Visualizing Categorical and Numerical Variables." *Journal of Computational and Graphical Statistics*, 1–16.

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.

Swayne, Deborah F, Duncan Temple Lang, Andreas Buja, and Dianne Cook. 2003. "GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization." *Computational Statistics & Data Analysis* 43 (4): 423–44.

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.

Velagala, Vijay, Weitao Chen, Mark Alber, and Jeremiah J Zartman. 2020. "Multiscale Models Coupling Chemical Signaling and Mechanical Properties for Studying Tissue Growth." In *Mechanobiology*, 173–95. Elsevier.

Wang, Shisong, Debajyoti Mondal, Sara Sadri, Chanchal K Roy, James S Famiglietti, and Kevin A Schneider. 2022. "Set-Stat-Map: Extending Parallel Sets for Visualizing Mixed Data." In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, 151–60. IEEE.

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.