Breaking the Clutter: Strategies for Handling Numerical Ties in Generalized Parallel Coordinate Plots

Denise Bradford

Abstract

High-dimensional data visualization is a cornerstone of modern analytics and is critical for revealing complex relationships across numerous variables and observations. Traditional visualization methods, such as scatterplots and Cartesian coordinate systems, struggle to manage datasets with more than three dimensions effectively, leading to challenges like overplotting, cognitive strain, and obscured patterns. Parallel Coordinate Plots (PCPs) have emerged as a powerful alternative for visualizing multidimensional data by mapping variables to parallel axes and representing observations as intersecting polylines. Despite their utility, PCPs face significant limitations, particularly in handling numerical ties, visual clutter, and mixed data types.

This dissertation focuses on advancing the design and functionality of PCPs to address these challenges. Key contributions include a novel approach to managing numerical ties by introducing visual spacing and jittering techniques, which preserve data integrity while enhancing interpretability. Adaptive methods for axis reordering and flipping are explored to improve the readability of highly correlated dimensions and reveal hidden patterns. Additionally, the integration of interactive features, such as dynamic filtering, brushing, and linking, is presented to facilitate real-time exploration and analysis of complex datasets.

This research systematically evaluates the performance of these enhancements across diverse applications, including genomics, finance, and social sciences. It demonstrates the potential of optimized PCPs to uncover clusters, trends, and outliers in large, multidimensional datasets. The findings underscore the importance of balancing data accuracy with visual clarity and provide a framework for future innovations in high-dimensional data visualization. These advancements establish PCPs as indispensable tools for exploratory analysis, supporting data-driven decision-making in an era of ever-growing complexity.

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 data sets exacerbate these issues by introducing visual clutter, making it harder to track individual observations (Heer and Bostock 2010). These challenges require innovative visualization techniques I don't think they have to be innovative… PCPs have been around for a while… tours, dimension reduction, etc. are also pretty straightforward. Aim less for a pretty sounding conclusion to the paragraph and more for simple and direct - we need ways to represent high p data which facilitate EDA and understanding of the structure of the data. that simplify complex data while preserving its structural integrity.

Parallel Coordinate Plots (PCPs) are a powerful way to examine data with more than one dimension. PCPs show each observation as a polyline on parallel axes, letting analysts follow data points across multiple single-variable axes (Inselberg 1985). While PCPs are less common than Cartesian graphs, with some training, it is possible to see correlations, identify clusters, and examine outliers in a multivariate dataset. These plots are especially good at showing relationships and trends between multiple variables. However, as with most visualization methods, overplotting reduces the effectiveness of PCPs, disrupting the ability to trace a single observation across multiple axes. This can complicate the viewer's ability to diagnose and cluster information visually. We begin by considering a series of plots which share characteristics with PCPs: multiple variables, each displayed on a vertical axis, with connecting structures which show the connection between variables. In Chapter … we demonstrate the use of a parallel coordinate plot in an exploratory setting, evaluating its effectiveness for exploratory analysis of publicly available data about small towns when used by residents of those towns. In Chapter … we examine the problem of numerical ties and overplotting, proposing a method for breaking ties visually which preserves numerical scaling while also allowing viewers to follow single observations across the plot. Chapter … discusses the computational implementation of the new methodology and its integration into the `ggpcp` R package. Finally, chapter … presents an empirical evaluation of the transformations proposed in Chapter …, along with the modifications to the display of categorical variables first presented in VanderPlas et al. (2023). The combined contribution of these chapters will improve the representation of data in generalized parallel coordinate plots, which should improve the visualization of large, complex data sets.

Coordinate Systems

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.

Include a 2D and 3D Cartesian plot here, along with a corresponding PCP. e.g. Row 1: 2D, 3D Cart, Row 2: 2D, 3D PCP. Highlight one or two points in each of the Cartesian plots, and the corresponding lines in the PCPs. It's fine to just rnorm() the data and it doesn't necessarily matter that the points have a strong linear relationship.

Evolution of data visualization methodologies

From its origins in the 19th century to the computational advances of today, visualization techniques have undergone significant transformations to address multidimensional, hierarchical, and mixed data challenges. This section delves into the historical development and interconnections among several key visualization techniques: Sankey diagrams, Alluvial plots, Parallel Sets plots, Hammock plots, Common Angle plots, and other adaptations which are similar to Parallel Coordinate Plots. Need to define basic inclusion criteria. What plots are in and out?

These innovations collectively chart a trajectory of creative solutions rooted in foundational principles to meet evolving challenges. This sentence says basically nothing – write tight.

Early Foundations: Statistical Graphics in the 19th Century

Statistical graphics of the 19th century - bar charts, pie charts, and flow maps - formed the groundwork for modern visualization techniques. Many of these early charts were based on industrial meters: pie charts share some connection to e.g. steam pressure gauges; early graphical innovators, like William Playfair, worked with such equipment during the industrial revolution (Berkowitz 2018). While these early tools primarily focused on small-scale, univariate or bivariate datasets, they offered a glimpse into the potential of data relationships. These methods demonstrated the power of visualizing data to uncover patterns, but with limited dimensionality and scope.

In addition, the period between the late 1700s and the late 1800s produced many different experiments in graphical representation. Some were persistent; bar and pie charts are still in use, while others faded from popular consciousness. In some cases, the same graphical structures were re-invented and popularlized later.

Figure 1 is from the Statistical Atlas of the United States published in 1898, visualizing the results of the eleventh US census. This proto-parallel coordinate plot shows the rankings of the most populous cities in the United States at each census from 1790 to 1890. Each axis in the plot represents a US census, and objects along the axes are cities, connected from census to census to show the relative rank of each urban area by population.

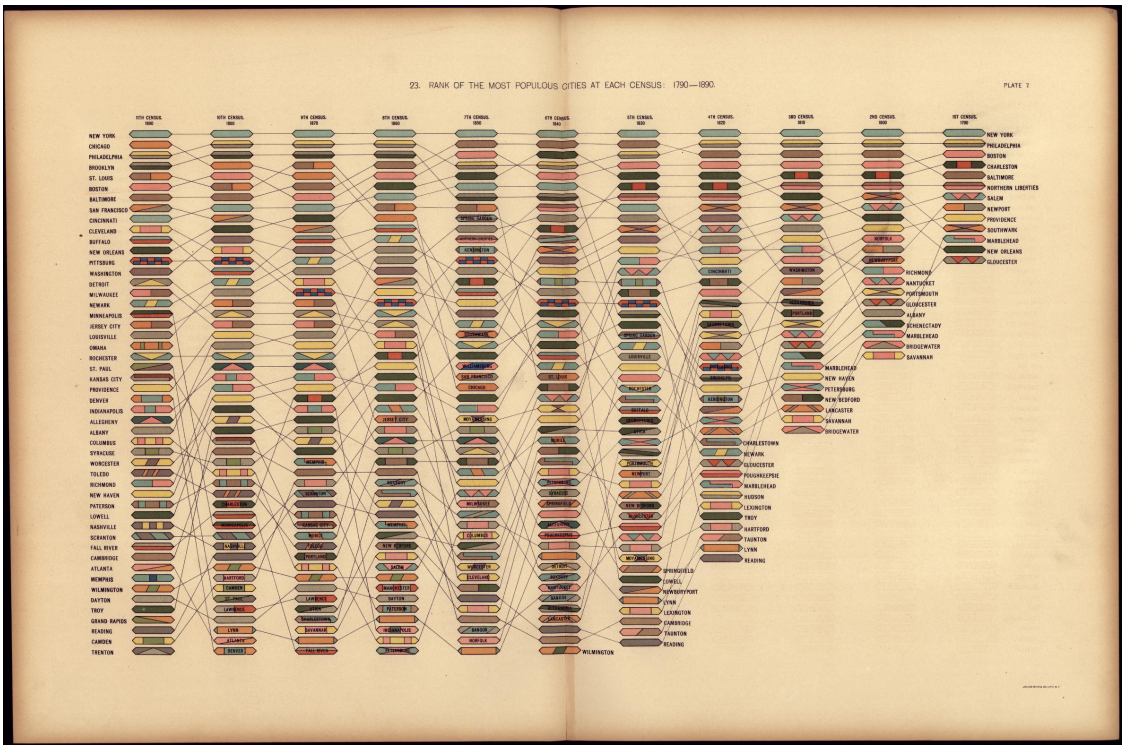


Figure 1: A 19th Century Parallel Coordinate Plot of U.S. cities. This plot shows the relative rank of each of the cities in the US over the course of each census which was conducted. This plot appeared in the Statistical Atlas created after the eleventh (1890) US Census (States Census Office" and Gannett 1898). The charts in the statistical atlases were created using lithography, which limits the number of unique colors used. In addition to color variations, markers are patterned differently to indicate distinct cities and facilitate tracing the city across the decades. Adjacent census observations are connected with thin black lines. Occasionally, cities are absorbed into other populous cities (e.g. Spring Garden, PA, which became part of Philadelphia between the 7th and 8th census) or disappear from the list of most populous cities entirely (e.g. Wilmington, which only appears in the 6th census.)

Around the same time period, Irish Captain Matthew Henry Phineas Riall Sankey published his diagram of energy efficiency in a steam engine. While this diagram, shown in **?@fig-sankey-orig**, does not conceptualize the energy flows shown as variables, this type of diagram became common in science and engineering; similar diagrams have been used to show the flow of heat and energy, but also the flow of items through a factory and, more recently, the life-cycle of products (Schmidt 2008).

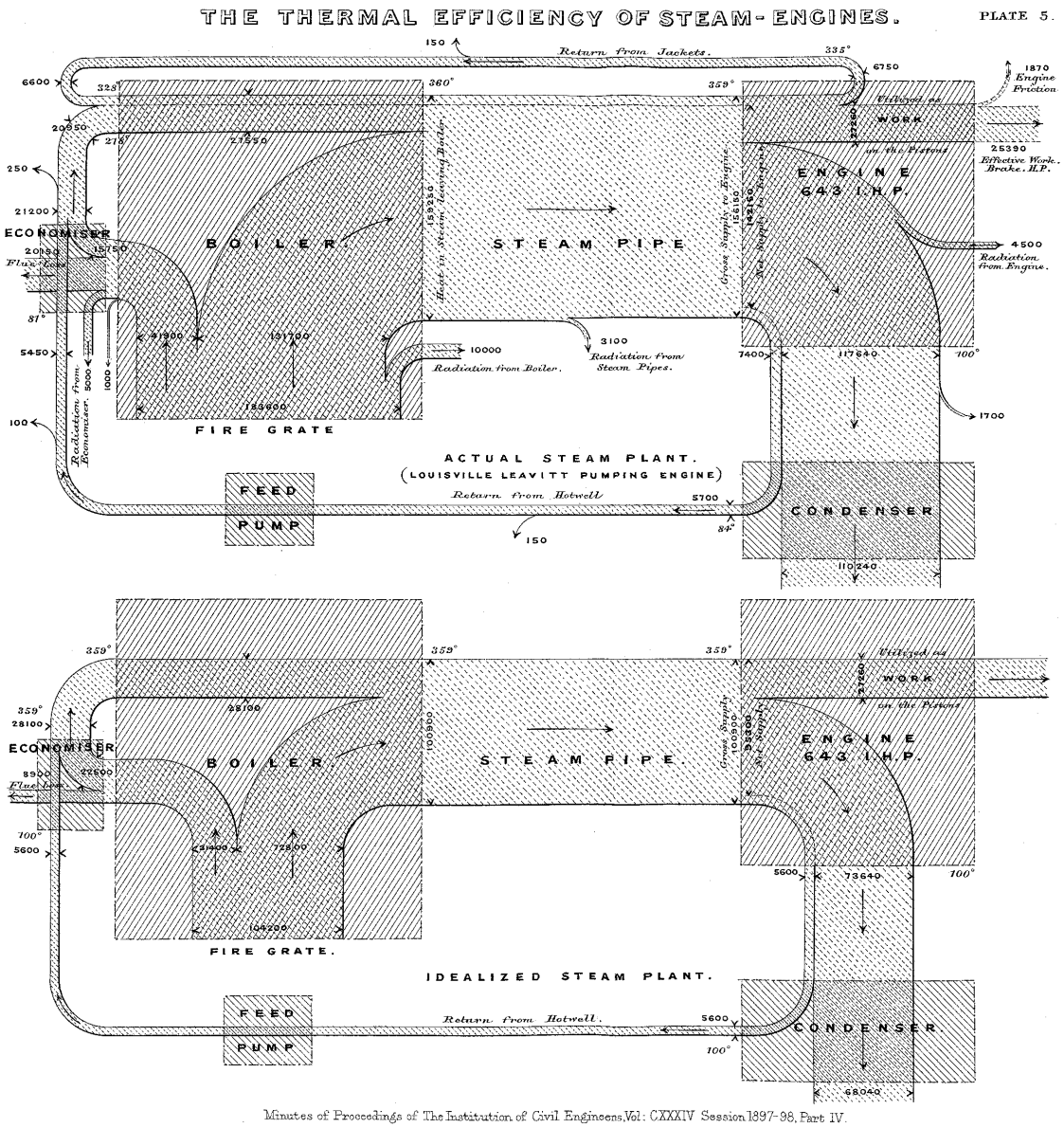


Figure 2: The original Sankey diagram of energy efficiency in a steam engine, reprinted from A. B. W. Kennedy and Sankey (1898) (Work in the public domain).

Sankey diagrams are visually similar to PCP variants, including parallel sets and hammock plots, and in many of the ways they are used today, PCPs may indeed be a better choice, but fundamentally, the intent of Sankey diagrams was to show the flow of a conservable quantity through a physical process of some sort. This is in contrast to PCPs, where the intent is to show a single entity's values across several variables. The overlap is slight, but present: one could examine a batch of 100 widgets, tracking their location or function at different time points; each repeated measurement would be a variable. A plot of this data using parallel axes for time points would be an example of an alluvial

plot; we will discuss these plots in the next section.

In addition, Sankey diagrams do not impose the traditional constraint of parallel axes (and while some modern variations use polar transformations, the basic principle of parallel axes is still maintained up until the final spatial transformation). Sankey diagrams also allow for multiple "exit points" - at each stage of the industrial process, there may be waste, leakage, and other processes which result in the observed quantity exiting what is presumed to be a continuous and perhaps cyclical system. While an important example, Sankey diagrams as they were originally conceived and used are different in both purpose and construction from parallel coordinate, parallel sets, and hammock plots.

Parallel Coordinate Plots

Inselberg (1985) introduced parallel coordinate plots as a solution to visualization of multidimensional numerical data. Parallel coordinate plots use the projection space, rather than Cartesian space, in order to show relationships between variables; a point in Cartesian space is represented by a line between vertical, parallel axes in projection space.]{.svp} Despite their utility for numerical data, early PCPs could not effectively represent categorical or mixed-type data. Figure 3 Pearson's correlation coefficient $r = \{-1, 0, 1\}$. When variables are negatively correlated, a 'knot', or inversion appears in projection space; when variables are positively correlated, the two axes are connected by a series of parallel lines. When no correlation is present, the lines are jumbled, with no clear patterns.

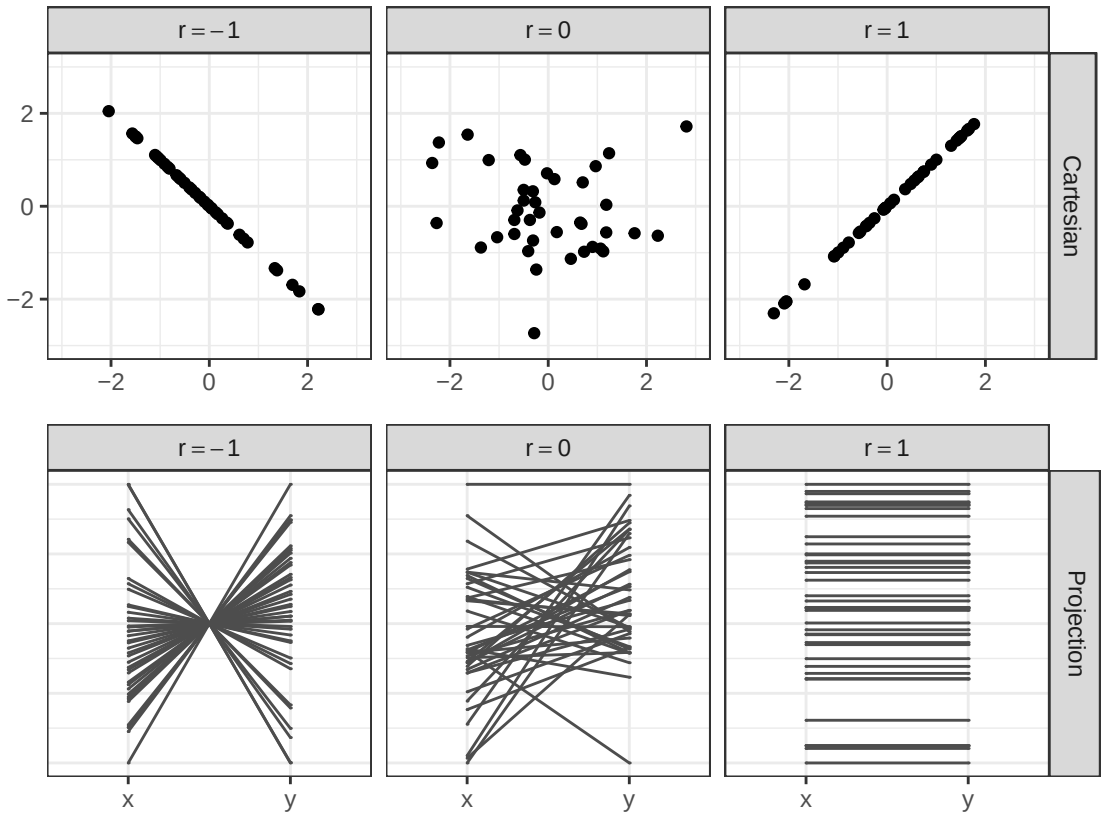


Figure 3: An illustration of extreme values of Pearson's correlation cofficient using scatterplots, which are created in Cartesian space (top row) and parallel coordinate plots, which utilize projection space (bottom row).

In effect, the parallel coordinate plot shows the visual equivalent of Kendall's τ, a robust version of Pearson's correlation coefficient which is calculated using ranks. The re-scaling of the vertical axes is equivalent to a ranking procedure.

Handling Categorical Variables

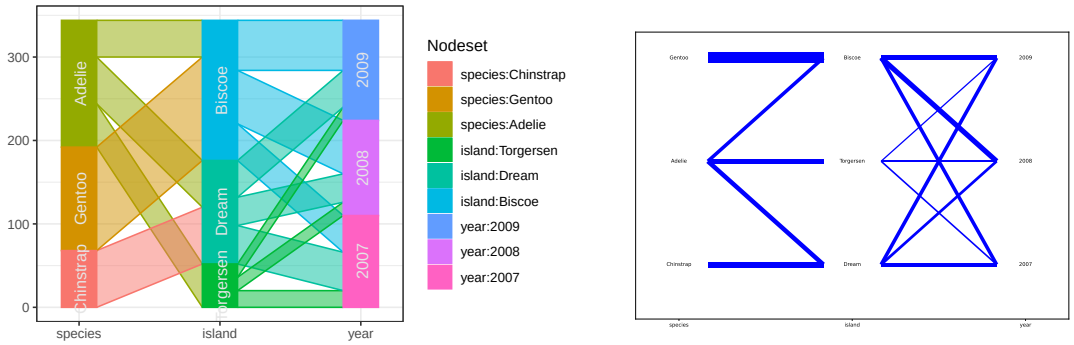
The transition to projection space from Cartesian space seems to have independently arisen several times, and in some cases, the origins of the practice are not entirely clear. Sankey diagrams, for instance, were not initially designed to leverage projection space; they were intended to visualize industrial processes, rather than variables, with axes that were often not parallel. That is, in mathematical terms, Sankey diagrams are a visual representation of flow through a directed, weighted graph. Alluvial diagrams, which are visually similar to both Sankey diagrams and parallel sets plots, are often considered to be a subset of Sankey diagrams; typically, alluvial diagrams visualize the change in weights in a graph over time - that is, the vertical axes are different time points, and the nodes are a set of categories which are consistent across time points. There is considerable confusion between alluvial diagrams and Sankey diagrams, as briefly mentioned in Holtz (2024). This confusion is compounded by different implementations of e.g. parallel sets plots, such as the `geom_parallel_sets` implementation in the `ggforce` R package (Pedersen 2024), which describes itself this way:

> If the variables has an intrinsic order the representation can be thought of as a Sankey Diagram. If each variable is a point in time it will resemble an alluvial diagram.

One obvious solution to the problem of how to visualize categorical variables using parallel axes is to convert the categorical variables into numeric variables and to use a standard parallel coordinate plot. This solution is not optimal even for ordered categorical variables, but for unordered variables, it is particularly unintuitive. In addition, of course, any information about the category frequency is lost in the numeric transformation.

Hammock plots and parallel sets plots arose independently in the early 2000s (Schonlau 2003; Bendix, Kosara, and Hauser 2005), though both plots represent categorical data on parallel axes. Parallel sets plots (Figure 4 a) were initially developed as part of an interactive environment for exploring crosstab data. Hammock plots (Figure 4 b) grew out of a desire to represent both numerical and categorical data effectively using parallel axes; in addition, these plots included support for visualizing missing data from the start. Parallel sets plots were one facet of a series of linked representations of categorical data, and included histograms as subplots within the parallel sets plot design.

There are some notable differences between the plots in construction: hammock plots do not divide up the full space of the vertical axis between categories; instead, the lines between adjacent axes are sized relative to the size of the intersection between categories on adjacent axes. In addition, hammock plots center all lines leaving a node, so the lines overlap. In parallel sets plots, the full axis length is divided up proportional to the marginal composition of the variable, and groups of observations which connect to different categories on the adjacent axis are not aligned. This facilitates part-to-whole comparisons of the joint probability values.



(a) Parallel sets plot

(b) Hammock plot

Figure 4: Hammock (Yang 2023) and parallel sets (Bendix, Kosara, and Hauser 2005), [implementation](Hofmann and Vendettuoli 2016) plots of categorical variables in the palmerpenguins (Horst, Hill, and Gorman 2020) data.

Hierarchical Data Visualization: Parallel Sets in the Early 2000s

Mid-2000s: Enhancements in Parallel Coordinate Plots

Advancements in Parallel Coordinate Plots during the mid-2000s expanded their functionality to accommodate mixed data types and interactive exploration. Features such as brushing, filtering, and dynamic axis reordering enabled users to uncover hidden patterns and insights within datasets. Generalized Parallel Coordinate Plots emerged during this era, offering robust solutions for visualizing mixed numerical and categorical datasets while maintaining scalability and clarity. Figure 5 is a parallel coordinates plot visualizing a multivariate, mixed-type dataset.

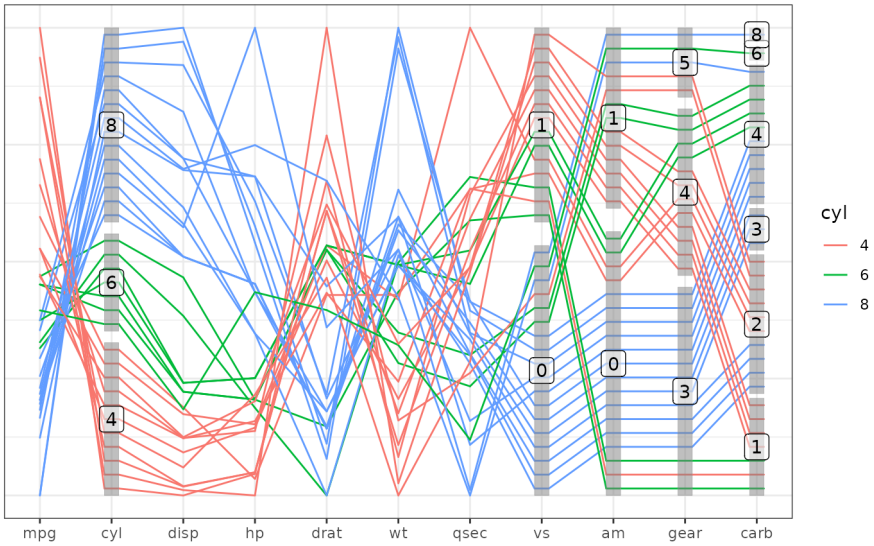


Figure 5: An example from the Generalized Parallel Coordinate Plots `ggpcp` package in R written by Hofmann et al

Late 2000s: Specialized Visualization Techniques

The late 2000s saw the development of specialized visualization techniques to address domain-specific challenges. Hofmann and Vendettuoli (2013) further examined the implications of the line-width illusion in categorical data visualizations. Their findings revealed that inconsistencies in line widths and unsuitable geometric representations exacerbated perceptual inaccuracies, complicating the interpretation of overlapping or closely related dimensions. ~~Common Angle Plots, for instance, employed radial layouts to visualize cyclic data, such as temporal trends, effectively.~~ THIS IS TOTALLY FALSE!!! I have no idea where this is coming from. Please fact-check everything you're saying and back it up with an accurate reference. By overcoming the limitations of linear arrangements for periodic patterns, these plots enhanced the interpretability and usability of such datasets. Figure 6

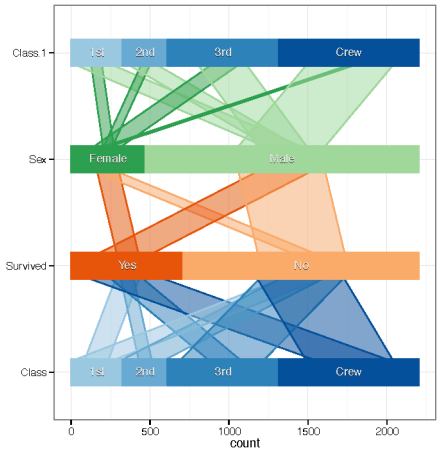


Figure 6

Interconnected Pathways of Evolution

The progression of these visualization techniques underscores a shared heritage of iterative improvement and cross-disciplinary inspiration. For instance, Parallel Sets informed hierarchical visualizations, which subsequently influenced flow-based approaches like Sankey and Alluvial Plots. Similarly, the iterative refinement of Parallel Coordinate Plots highlighted the necessity of integrating multiple data types while preserving clarity and scalability. These interconnections exemplify the synergistic nature of advancements in data visualization.

Parallel Coordinate Plots (PCPs): A Foundation

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 7 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.

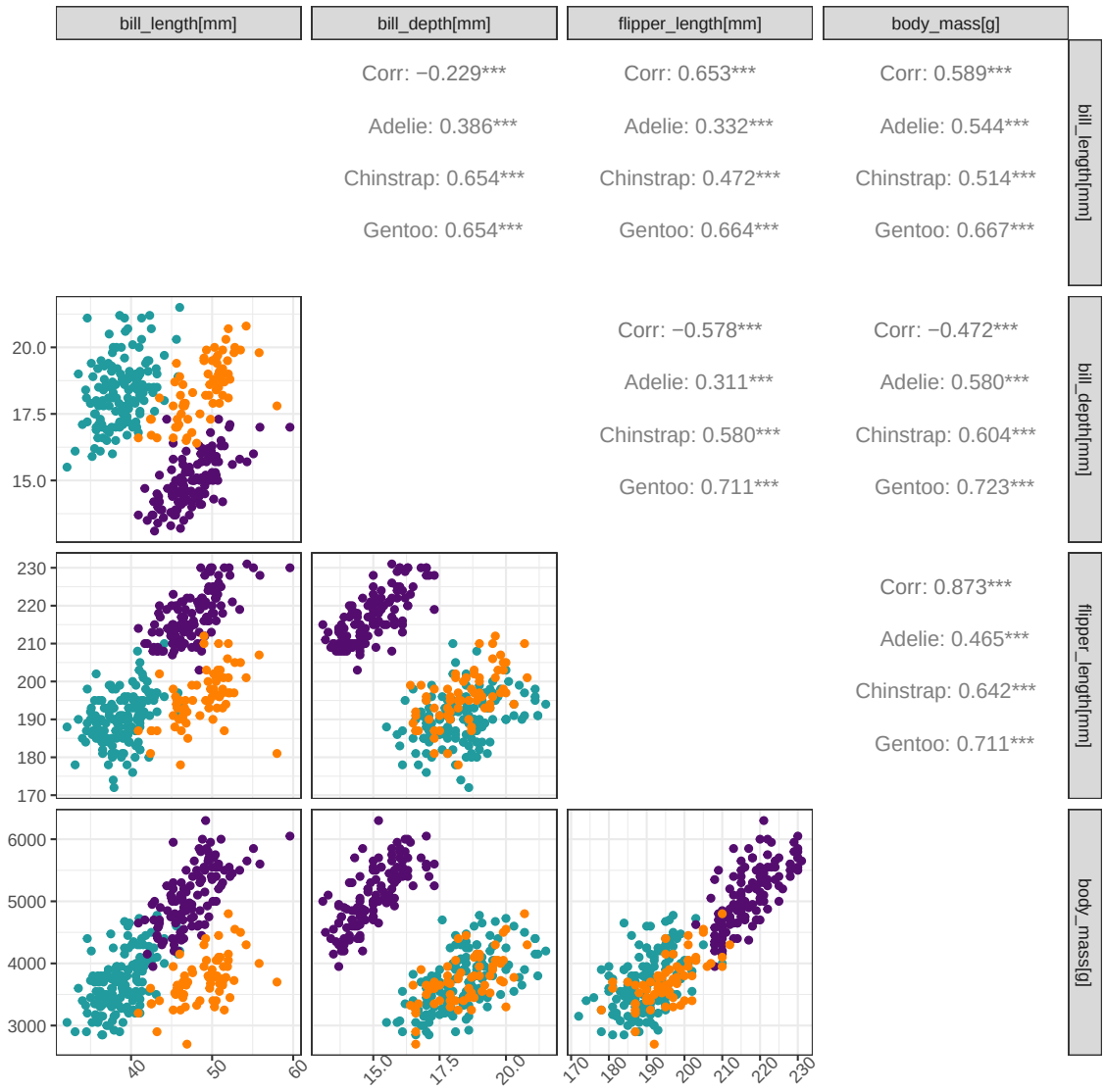


Figure 7: A generalized pairs plot

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.

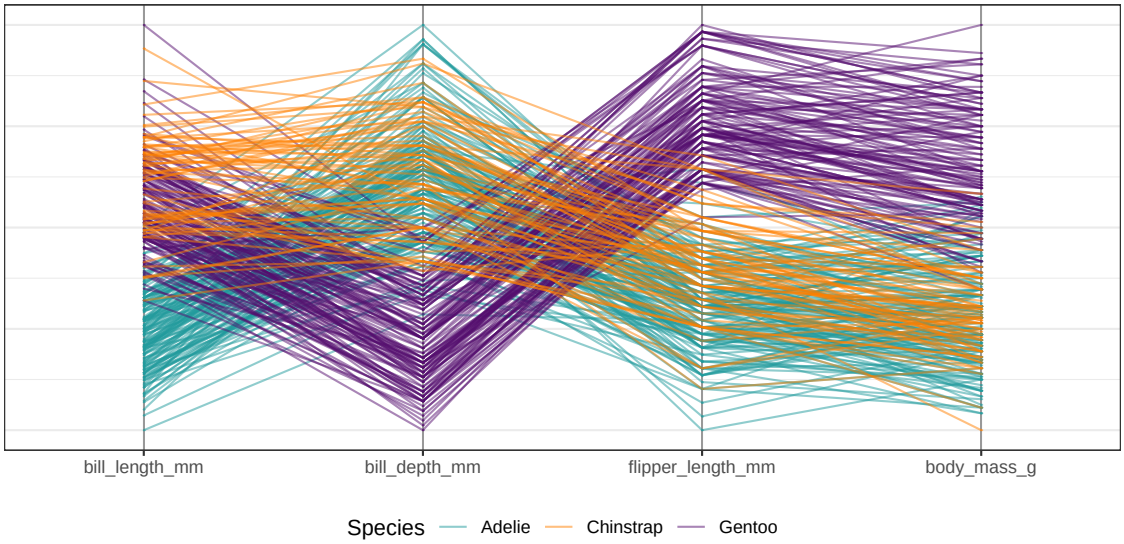


Figure 8: A parallel coordinate plot

In Figure 8, each line shows a different penguin, and the colors of the lines show the species. 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 7 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 9 replicates Figure 8 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.

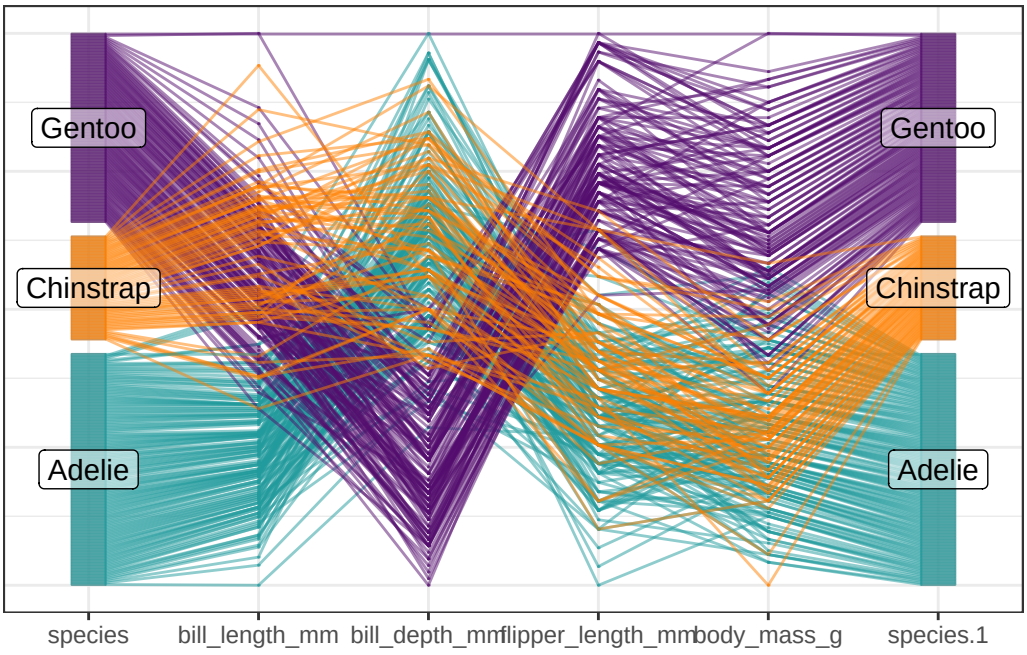


Figure 9: A generalized parallel coordinate plot, with species on the left and right of the plot; observations are ordered on the right side based on the value of body_mass_g, and on the left side based on the value of bill_length_mm. Translucent lines reduce the impact of overplotting and allow perception 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.

Parallel Sets Plots

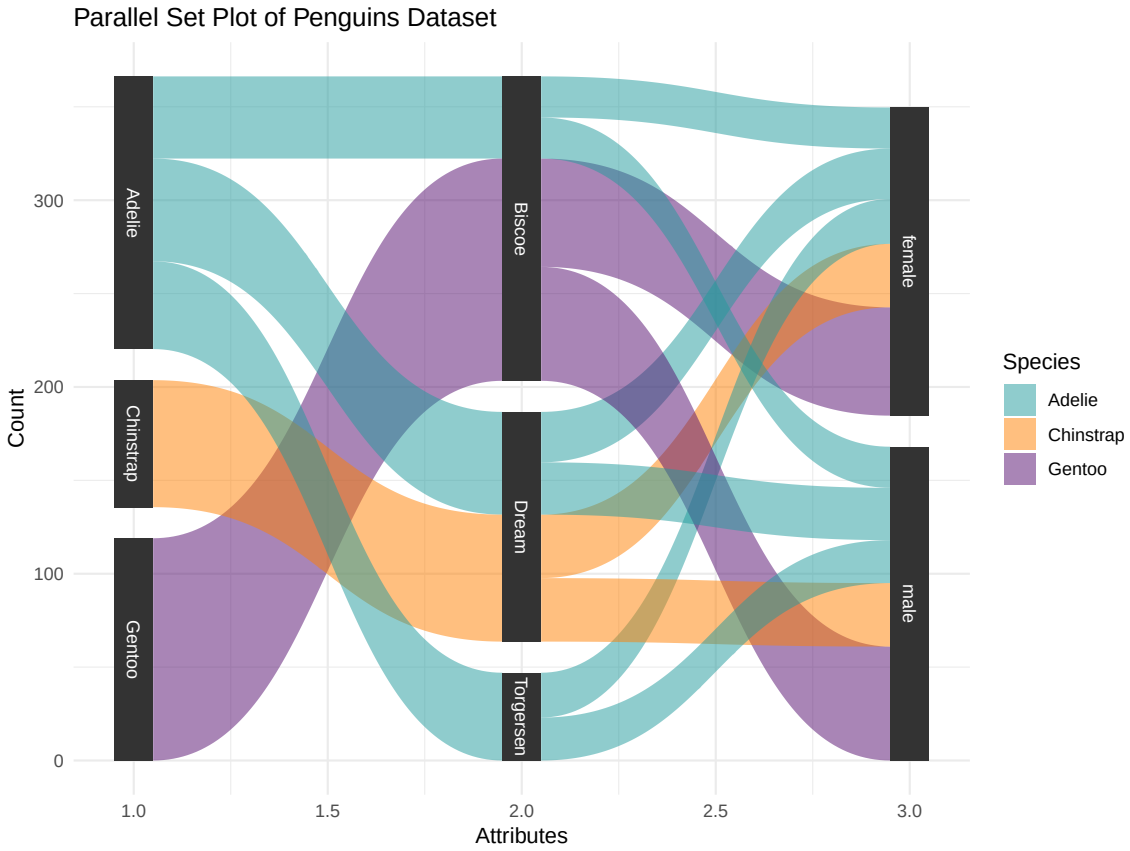


Figure 10: A Parallel Sets Plot

Parallel Sets, designed to visualize multidimensional relationships in numerical data, trace their origins to Parallel Coordinate Plots, first introduced by Alfred Inselberg in the 1950s. The adaptation for categorical data emerged in the late 1990s under Daniel Keim's and collaborators' leadership, addressing the growing need for tools capable of handling non-numerical datasets. Subsequent refinements, particularly by Kosara, Bendix, and Hauser (2006), emphasized intuitive designs that improved the interpretability of complex relationships. These advancements marked a significant step in broadening the scope of multidimensional visualization methods.

The 2010s witnessed a transformative era for Parallel Sets through dynamic and interactive computational tools such as D3.js and Matplotlib. Features like temporal zooming and filtering greatly enhanced the analytical depth of these visualizations. By allowing seamless navigation across dimensions, these tools solidified their role in exploratory analysis and historical research, as highlighted by Bostock, Ogievetsky, and Heer (2011).

Parallel Sets excel in revealing trends, anomalies, and intricate patterns within multidimensional datasets. However, they are not without limitations. Oversimplification of continuous data and challenges in readability—particularly in poorly designed implementations—can impede their effectiveness. Nonetheless, these tools have proven instrumental in bridging quantitative and qualitative approaches to historical studies. As computational capabilities expand and machine learning becomes increasingly integrated with data visualization, Parallel Sets hold the potential to redefine predictive modeling of historical trends and their implications for contemporary analysis.

The line-width illusion, part of the broader category of perceptual distortions known as the sine illusion, presents a significant challenge in visual data interpretation. This phenomenon causes variations in line width to distort perceived relationships or angles, leading to potential misinterpretations of visualized data. Early explorations of this issue in three-dimensional visualizations, such as those by Day and Stecher (1991), underscored its cognitive impact, highlighting the need for alternative visualization techniques to mitigate these biases.

Hofmann and Vendettuoli (2013) further examined the implications of the line-width illusion in categorical data visualizations. Their findings revealed that inconsistencies in line widths and unsuitable geometric representations exacerbated perceptual inaccuracies, complicating the interpretation of overlapping or closely related dimensions. By proposing Common Angle Plots as a solution, they demonstrated how perceptually robust designs could minimize distortions and enhance the extraction of meaningful insights from complex datasets.

These foundational studies underscore the necessity of addressing perceptual limitations in visualization design. They emphasize the critical balance between aesthetic simplicity and cognitive accuracy in developing tools that empower users to derive reliable conclusions from data.

Hammock Plots

Fig 11 is a Hammock plot that shows the relationship between two categorical variables, gear and cylinders (cyl). Each color-coded band represents a different category within these variables. The plot provides frequency but lacks precise counts or numeric values. Overplotting is evident, making it difficult to trace individual pathways. The plot shows the general distribution and relationships between gear and cylinder categories, but lacks numeric relationships or exact proportions.

Schonlau (2003) introduced Hammock Plots, a variation of Parallel Coordinate Plots, as a method for visualizing relationships between categorical variables. Designed to handle mixed data types, Hammock Plots provides a visual framework to simplify complex relationships and reduce visual clutter through "bandwidth" representations that emphasize connections between categories. This capability is particularly advantageous in health science datasets, where managing and interpreting intricate, multivariate data is essential.

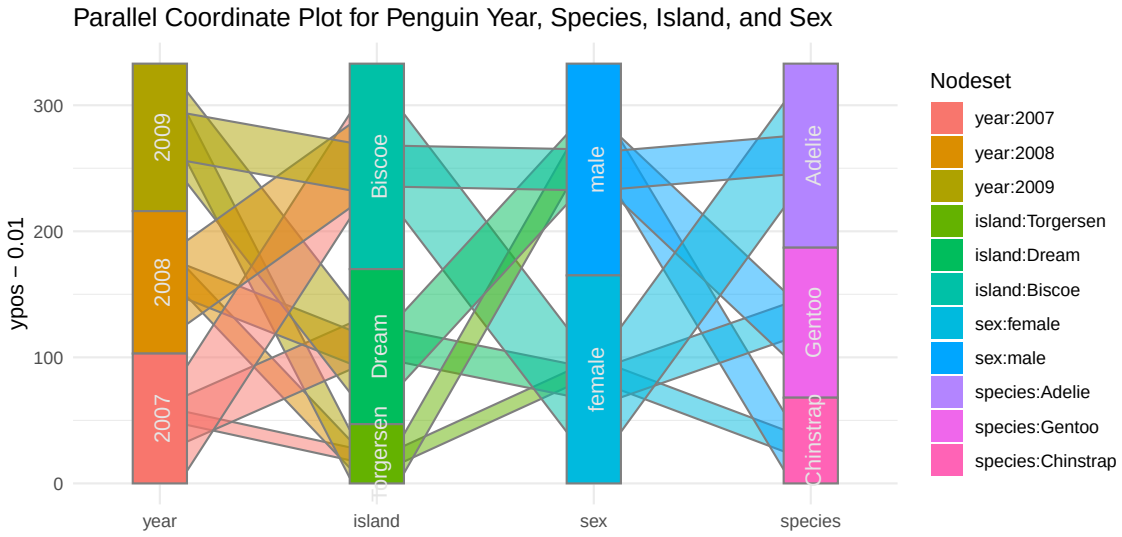


Figure 11: Hammock Plot (Schonlau 2024)

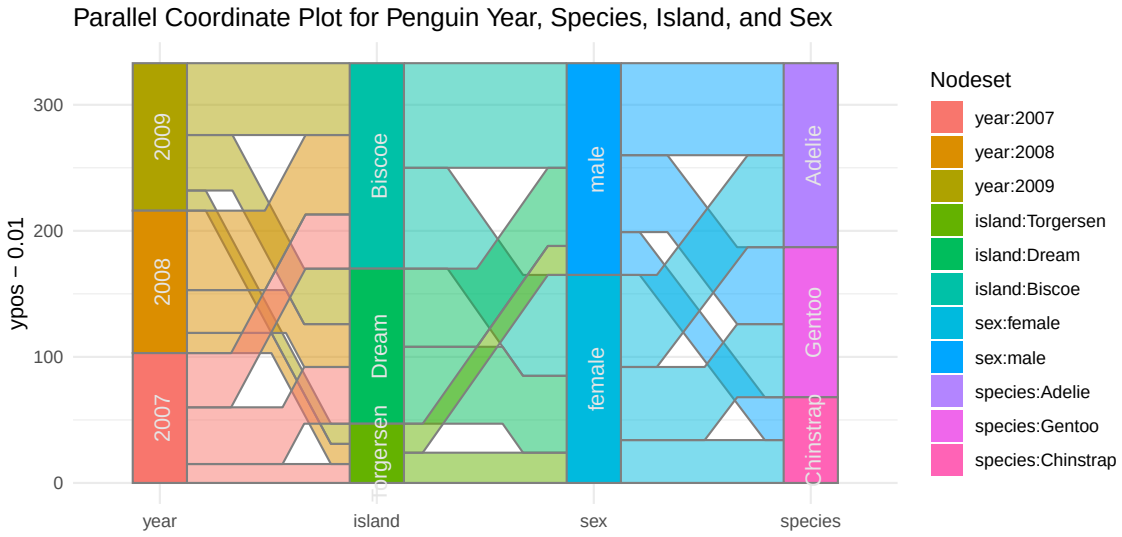


Figure 12: Common Angle Plot (Hofmann and Vendettuoli 2013)

Compared to mosaic plots, Hammock Plots offer a more compact and cognitively accessible representation of multidimensional categorical data. By alleviating the cognitive load on observers, these visualizations surpass traditional tools such as bar charts, making them indispensable for applications in health research and other fields dealing with mixed numerical and categorical variables.

Symanzik, Friendly, and Onder (2018) utilized Hammock Plots to analyze the Titanic dataset, highlighting their ability to align lines representing similar or identical values. This alignment facilitates the comprehension of relationships within the data, particularly in parallel coordinate spaces. However, they also acknowledged potential drawbacks, noting that Hammock Plots may introduce visual clutter when categorical distinctions are insufficiently defined, underscoring challenges in their application to continuous data.

Compared to mosaic plots, Hammock Plots offer a more compact and cognitively accessible representation of multidimensional categorical data. By alleviating the cognitive load on observers, these visualizations surpass traditional tools such as bar charts, making them indispensable for applications in health research and other fields dealing with mixed numerical and categorical variables.

The integration of Hammock Plots into grammar-based visualization frameworks further expanded their applicability. hofmannCommonAnglePlots2013 incorporated Hammock Plots into the `ggparallel` package, providing a grammar-driven interface for visualizing multivariate categorical data. This implementation preserved the foundational principles of parallel categorical displays while introducing capabilities for addressing edge cases, such as managing categorical ties and combining numerical and categorical data. schonlau2024 built on these advancements, adapting Hammock Plots to generalize Parallel Coordinate Plots by replacing lines with box elements. This modification leveraged box widths to represent the number of observations, enhancing their utility for both categorical and mixed data types. Schonlau's approach also tackled perceptual issues, such as the reverse line width illusion, by refining traditional Hammock Plot designs.

Hammock Plots and Common Angle Plots address similar visualization challenges but through distinct methodologies. Hofmann's Common Angle Plots (2013) introduced a constant angular structure to mitigate the line width illusion while maintaining visual continuity between categories. By ensuring consistent angular connections, this method preserved marginal proportions and provided perceptually robust visualizations.

Schonlau's adaptations, on the other hand, emphasized flexibility across diverse data types. By addressing both categorical and mixed data displays, Schonlau's Hammock Plots underscored the importance of adaptability in visualization tools. Despite their differing approaches, both methods shared a common goal: enhancing the accuracy and usability of visualizations for categorical and mixed data.

Hofmann and Schonlau's contributions represent complementary advances in the visualization of multivariate categorical data. Hofmann's focus on perceptual clarity through angular consistency and Schonlau's emphasis on adaptability illustrate two facets of the ongoing effort to refine data visualization techniques. Together, these approaches reflect a shared commitment to overcoming traditional visualization challenges, paving the way for more accurate, intuitive, and versatile tools for understanding complex datasets.

Alluvial plots & Sankey diagrams

Alluvial plots provide a compelling framework for visualizing flows and transitions within categorical data, capturing changes across dimensions or over time. Inspired by stratified

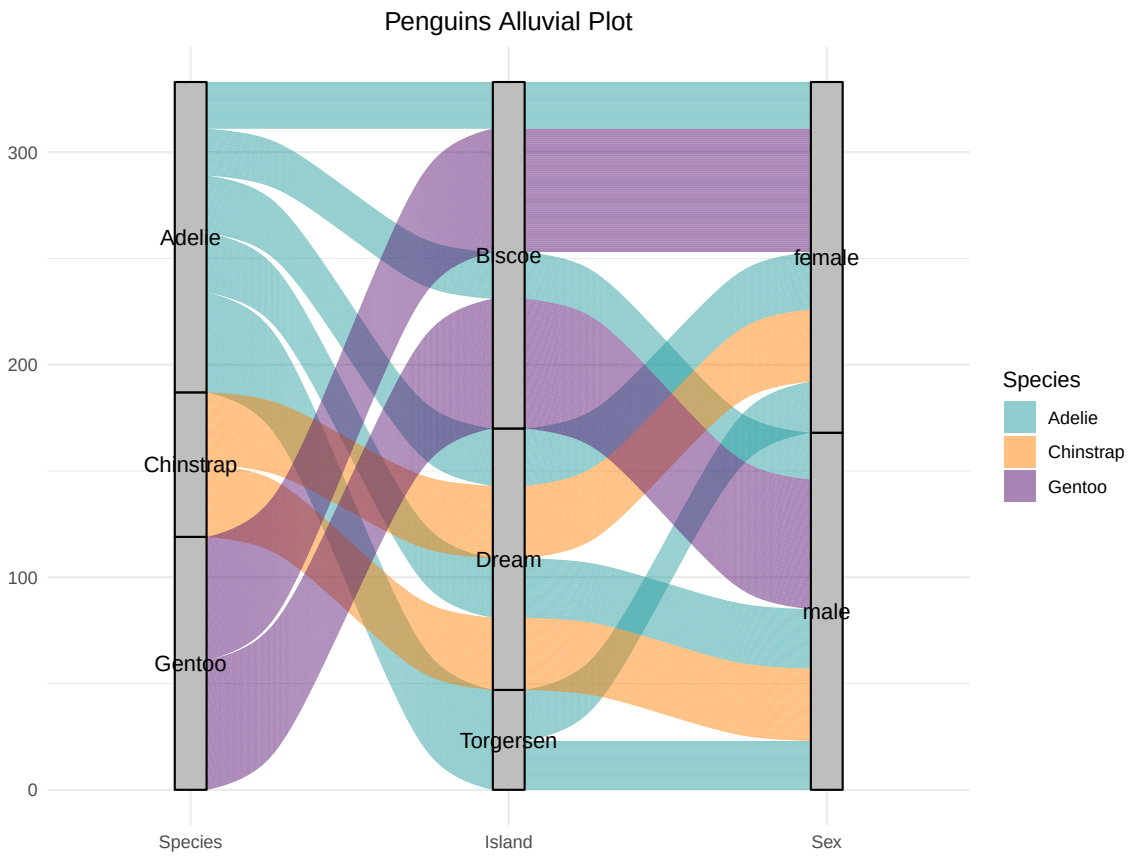


Figure 13: A Alluvial Plots

data representations, these plots evoke the imagery of sedimentation patterns in geology, reflecting layered and interconnected structures. Since their popularization in the early 2000s, alluvial plots have gained widespread application across fields such as sociology and genomics, where understanding transitions and relationships within datasets is essential (Rosvall and Bergstrom 2008).

The design of alluvial plots is centered around parallel axes representing categorical dimensions, with connecting bands illustrating relationships between elements across these axes. The width of these bands often reflects magnitude or relative importance, though the primary emphasis remains on the qualitative nature of transitions. For example, in bibliometrics, alluvial plots effectively trace the evolution of research topics over time. These plots intuitively convey intricate categorical relationships, offering a visually accessible means of interpreting complex datasets.

Sankey diagrams, first introduced by A. B. Kennedy and Sankey (1898) during his analysis of steam engine efficiency, were initially designed to illustrate energy flows. Over the years, their application has expanded significantly, encompassing areas such as resource allocations, material transfers, and financial systems. These diagrams have found utility across diverse fields, including environmental science, logistics, and economics.

The defining characteristic of a Sankey diagram is its proportional representation of flow magnitude. Nodes, representing entities or processes, are connected by links, whose widths are scaled to reflect the quantity being transferred. This proportional emphasis makes Sankey diagrams particularly well-suited for visualizing resource distributions, diagnosing inefficiencies, and analyzing complex systems with substantial quantitative components. For instance, they are commonly employed to trace energy inputs and

losses in power plants or to map the flow of financial resources (Edward R. Tufte and Graves-Morris 1983).

Sankey diagrams and alluvial plots share a focus on flow visualization but differ in their primary purpose and design. Sankey diagrams are optimized for representing quantitative data flows, such as energy or resource transfers, with a focus on the magnitude of connections. In contrast, alluvial plots are designed to visualize transitions and correlations within categorical datasets, emphasizing relationships rather than quantities.

The structural distinctions between these techniques further highlight their unique strengths. Sankey diagrams use nodes and proportional links to depict numerical flows, making them invaluable in energy analysis, financial tracking, and logistics. Conversely, alluvial plots rely on parallel axes and connecting bands to represent categorical correlations, enabling an intuitive understanding of trends across dimensions.

Both visualizations are cross-disciplinary, scalable to diverse datasets, and offer extensive customization options. Their complementary strengths make them powerful tools for exploring and interpreting complex data in both numerical and categorical domains.

The transition from categorical parallel axis plots to mixed-type visualizations represents a significant evolution in data visualization, driven by the growing need to analyze and represent heterogeneous datasets simultaneously. Traditional parallel axis plots were initially designed to handle numerical or categorical data independently, often limiting their utility when datasets contained both types. The development of mixed-type plots addressed this limitation by incorporating flexible mechanisms to integrate numerical values with categorical groupings, ensuring a seamless representation across diverse dimensions. Key innovations, such as dynamic axis scaling, combined use of lines and bands, and interactive filtering, have enhanced the capacity of these plots to capture complex relationships within mixed datasets. These advancements allow users to analyze continuous trends alongside discrete transitions, making them particularly valuable in domains such as genomics, social sciences, and finance, where datasets frequently encompass numerical measurements and categorical classifications. By bridging the gap between purely categorical and numerical representations, mixed-type visualizations extend the interpretive power of parallel axis plots, enabling more nuanced and comprehensive insights into multidimensional data.

Mixed Variable Data Parallel Representations

The evolution of Parallel Coordinate Plots (PCPs) from tools for numerical data visualization to instruments for mixed-type datasets reflects a rich history of advancements in data visualization. This body of literature explores key contributions and methodologies that have extended PCPs to represent heterogeneous datasets comprising categorical and numerical attributes. These developments build upon foundational optimization principles, such as those proposed by Hurley, and emphasize clustering, interactivity, and dimensionality reduction innovations to enhance readability, interpretability, and analytical depth.

Historical Foundations and Early Innovations

Friendly and Denis (2001) provide a comprehensive account of the evolution of statistical graphics, including parallel coordinate systems, tracing their origins to foundational thematic cartography. By emphasizing the historical significance of statistical graphics in multivariate data visualization, their work sets the stage for subsequent innovations,

including Hurley's algorithms. Hurley (2004) made a pivotal contribution by introducing the end-link algorithm, which optimizes variable ordering in parallel plots to enhance the clarity and interpretability of multidimensional datasets. This foundational work has since informed modern parallel axis plots, enabling better cluster recognition and visual analytics.

Application-Driven Advancements

Sartorio et al. (2004) utilized parallel plots to examine the relationships between BMI, age, and muscular power output, demonstrating the relevance of PCPs in biomedical studies. Hemming (2004) employed parallel plotting methods to analyze geological and climate data, offering a practical application of Hurley's clustering principles in environmental datasets. These studies underscore the versatility of PCPs across domains, from healthcare to environmental science.

Mixed-Type PCPs: Expanding Dimensions

Kwak and Huh (2008) advanced PCPs by applying them to mixed-type datasets, linking categorical and numerical variables through correlation metrics. This approach, grounded in Hurley's optimization principles, improved the interpretive clarity of PCPs and extended their applicability to heterogeneous data. Wang et al. (2022) further innovated with Set-Stat-Map, a mixed-type PCP method for spatial data visualization. By clustering variables and grouping categorical attributes, this approach facilitated the integration of geospatial and numerical data, enhancing usability in geospatial contexts.

Optimizing High-Dimensional Data

Arif and Basalamah (2012) combined PCPs with similarity-dissimilarity metrics to highlight patterns in biomedical datasets, optimizing axes for classification accuracy and cluster separation. Shaham (2015) introduced tree-based methods for mapping mixed-type data into PCPs, emphasizing hierarchical relationships and employing clustering and dimensionality reduction techniques. These works extend the capabilities of PCPs to manage high-dimensional data, particularly in domains where hierarchical or clustered relationships are critical.

Interactive and Real-Time Exploration

Garrison et al. (2021) emphasized interactivity in their Dimlift framework, incorporating dimensional bundling for real-time exploration of mixed-type and incomplete datasets. By extending Hurley's clustering techniques, Dimlift enabled dynamic adjustments to visualizations, fostering deeper analytical engagement. Tortora and Palumbo (2024) introduced probabilistic clustering methods for PCPs, blending probabilistic models with deterministic visualization approaches to handle hybrid datasets effectively.

Practical Applications and Broader Contexts

Jones et al. (2004) applied statistical visualizations akin to Hurley's principles to diagnose renal obstructions, illustrating the role of PCPs in healthcare analytics. Markwick and Valdes (2004) utilized parallel axis methods to model elevation and climate data for simulation experiments, demonstrating their utility in palaeoenvironmental research. Gower and Dijksterhuis (2004) explored mathematical optimization techniques in data visualization, providing theoretical underpinnings that support PCP advancements.

Ma et al. (2020) advanced generative modeling techniques for heterogeneous datasets, visualizing results through pairwise PCPs. This approach merged data generation with visualization, aligning with optimization techniques inspired by Hurley. Integrating generative models into PCPs highlights the potential for uncovering more profound insights into complex mixed datasets.

These contributions collectively showcase Parallel Coordinate Plots' versatility and adaptability in visualizing mixed-type datasets. By incorporating principles of clustering, optimization, and interactivity, researchers have enhanced the utility of PCPs across diverse domains, from biomedical diagnostics to geospatial analysis. Integrating probabilistic and generative models further underscores the potential for future advancements, paving the way for more intuitive and powerful visualization tools. These innovations build upon Hurley's foundational work, reinforcing PCPs as indispensable instruments for exploring multidimensional data and uncovering hidden insights.

Perceptual Challenges with PCPs

Parallel Coordinate Plots (PCPs) have become increasingly useful yet sophisticated ways to show high-dimensional correlations as datasets become more complicated and multidimensional. Still, the inherent difficulties of PCPs—from visual clutter and perceptual biases to cognitive and computational demands—continue to restrict their accessibility and efficacy.

The interaction of technical details of visualization design with cognitive processes, including perception and attention, is a vast field of research. Insights from cognitive psychology, human-computer interaction, and statistical graphics emphasize the need for careful methods that address user heterogeneity and design constraints. These developments, which address problems such as line-width illusions and clutter and incorporate immersive technology, highlight the necessity of merging technical developments with user-centered strategies.

While these advancements solve many issues related to PCPs, they also highlight an even more general opportunity: improving PCPs to make them more scalable, efficient, and understandable. We can approach thoroughly realizing PCPs in high-dimensional data visualization by investigating methods that include dimensionality reduction, subspace projections, and rendering improvements. This shift toward optimization represents the next phase in PCP advancement, broadening their applicability for even more challenging analytical assignments.

In the following section, we explore optimization techniques that improve PCP performance and scalability and open the path for a new generation of high-dimensional visualization tools.

Balancing Accuracy and Comprehensibility in Data Visualization

The accuracy of data interpretation is highest when working directly with numerical values or one-dimensional visual representations, as noted by Spence (1990). However, as datasets grow increasingly large and complex, this approach becomes less practical. The escalating size and dimensionality of modern datasets necessitate methods for simplifying and consolidating information without losing critical insights. In this context, data visualization emerges as a pivotal intermediary step, bridging the gap between raw data and an observer's decision-making process.

Effective data visualization distills the essence of a dataset, presenting it in a format that is both accessible and comprehensible. While this often involves a trade-off—sacrificing some degree of numerical precision relative to tabular representations—the result is a more intuitive understanding of relationships among observations or variables. By prioritizing clarity and interpretability, well-designed visualizations enable observers to quickly grasp complex patterns and draw meaningful conclusions, making them indispensable tools for data-driven decision-making. This balance between accuracy and comprehensibility ensures that visualization remains a powerful ally in the ever-evolving landscape of high-dimensional data analysis.

Stimulus-Responsive Attention in Data Visualization and Its Implications for Design

Stimulus-responsive attention plays a critical role in data visualization, guiding observers to graphical elements that signal valuable information and refining their subsequent search processes to improve judgment accuracy. Spence (1990) highlighted the significance of visual psychophysics in understanding how fundamental graphical components can be leveraged to effectively convey information. His work underscored that the comprehensibility of graphs depends heavily on the cognitive and visual capabilities of the observer, emphasizing the need to design visualizations that align with these human factors.

The foundational work of Cleveland and McGill (1984) on graphical perception revealed notable variability in individuals' abilities to decode multidimensional data. Their findings underscore the importance of considering cognitive and perceptual differences in the design of visualizations such as generalized parallel coordinate plots. These plots, which integrate multiple dimensions within a single visual glyph, often present subtle variations that increase the complexity of interpreting intricate data relationships. The cognitive demands of such designs make it imperative to account for the diversity in user abilities and perceptual processing when presenting complex datasets.

Subsequent studies have expanded on these principles. Heer and Bostock (2010) and Simkin and Hastie (1987) demonstrated that errors occur at various stages of information processing, showing that the effectiveness of different graphical representations varies depending on the task and user accuracy requirements. This research reinforces the idea that visualizations must be tailored to the specific needs of their intended users to maximize their utility.

Carswell (1992) and Shah and Freedman (2011) further examined the interplay between task difficulty and perceptual processes. They emphasized that the interaction of these factors significantly influences how individuals interpret multidimensional data. For generalized parallel coordinate plots, these insights suggest that designs must address the dual challenges of cognitive complexity and perceptual variability to effectively convey

information. Incorporating these considerations into the design process ensures that multidimensional visualizations are not only aesthetically compelling but also cognitively accessible and functionally precise.

Tackling Clutter

For PCPs, visual clutter remains a key difficulty, especially when displaying dense or large-scale data. Focusing on dimensionality reduction, opacity management, axis reordering, and visual bundling, researchers have presented a spectrum of methods to reduce clutter while maintaining data integrity.

One fundamental approach to clearing clutter in PCPs is to reduce the number of dimensions seen there. Subspace projections and principal component analysis (PCA) help simplify data representation to visualize challenging datasets better. By demonstrating how effectively subspace projections reduce overplotting and produce cleaner multivariate data representations, Cheng (2018) illustrates high-dimensional biological datasets. Raidou (2019) also combines dimensionality reduction with clustering, emphasizing the flexibility of these methods.

Dynamic line transparency shows hidden patterns in areas of dense data. Using GPU acceleration to maximize opacity control, Stumpfegger et al. (2022) significantly improved clustering performance and visual clarity. Cao, Lin, and Gotz (2015) demonstrated that highly packed datasets remain interpretable by proving the value of opacity-based methods for separating clusters in probabilistic multi-label data.

Reducing overlapping lines and improving cluster separation mostly depends on strategic axis rearrangement. While Hurter (2017) developed filtering and reordering approaches to control clutter in dense datasets efficiently, Akbar and Gabrys (2018) underlined the role of data-driven reordering in increasing interpretability.

One creative way to reduce clutter without major information loss is to aggregate overlapping pathways into coherent bundles. Using bundling methods for Alzheimer's disease-related data, Lhuillier, Hurter, and Telea (2017) enhanced interpretability while preserving important data trends.

While decluttering improves PCP structural clarity, ensuring interpretive accessibility is still quite an important goal. The non-intuitive character of PCPs and the cognitive challenges of high-dimensional data analysis call for creative solutions to aid user understanding. Recent studies have concentrated on interactive features, design improvements, and cognitive psychology insights to make PCPs more approachable for users of all experience levels.

Improving Interpretive Accessibility

The interpretive difficulties with PCPs have motivated various techniques to improve user understanding and lower cognitive burden. These include interaction models, aspect ratio optimization, and cognitive aids, such as glyph-based upgrades and focus+context views.

PCP interpretability is greatly improved by interactivity. Linked brushing and glyph-based enhancements allow users to investigate particular data sets, lowering cognitive load and increasing engagement. While Chung et al. (2015) presented glyph sorting as an interactive tool for enhancing multi-dimensional analysis, Rzeszotarski (2017) emphasized the role of focus and context views in making PCPs less cognitively demanding.

PCP users interpret them largely through visual perception. Meka (2024) examined how aspect ratios affect line perception accuracy and developed practical design rules to improve usability. These results highlight how crucial visual design is for lowering misinterpretation and enhancing user satisfaction.

Extensive research on the cognitive mechanisms underlying PCP interpretation has helped to align visual designs with human natural processes. By investigating these processes and suggesting interaction methods to improve user comprehension, McColeman (2017) emphasizes the importance of customized training and cognitive tools for new users. Resch (2019) examined how conventional and current visualization tools are understood differently across user groups.

Accepting visual design uncertainty can help lower cognitive biases and promote subtle interpretation. Nowak (2023) showed that ambiguity can support more in-depth analytical insights by suggesting ways to construct PCPs that encourage sense-making in risk analysis and prediction.

Challenges in Isolating Patterns

Parallel Coordinate Plots (PCPs) provide a framework for representing multidimensional data by mapping variables to parallel axes and depicting observations as polylines intersecting these axes. In contrast, their design allows for the simultaneous visualization of multiple dimensions; however, the overlapping and intersection of lines in dense datasets often result in a "visual tangle," obscuring individual trends and patterns. As Liu et al. (2014) emphasize, this limitation significantly impacts the utility of PCPs in real-world applications involving large and complex data datasets.

One of the primary obstacles in using PCPs is the difficulty in achieving clarity for high-dimensional datasets. In their comprehensive survey, Julian Heinrich and Weiskopf (2013b) highlight critical limitations, noting that the accumulation of lines and intersections exacerbates clutter, making it challenging to discern meaningful insights. As proposed in their work, techniques such as dynamic filtering and opacity control provide incremental improvements but fail to address the core issues comprehensively. Firat et al. (2022) underscore the importance of user literacy in interpreting PCPs, advocating for interactive pedagogical tools to help novice users better understand these visualizations. Despite these initiatives, the complexity of tracking individual observations across dimensions remains a significant challenge.

Interactive tools have emerged as promising solutions to the challenges of clutter and readability in PCPs. Kaur and Karki (2018) introduced bifocal parallel coordinate plots, which focus user attention on selected dimensions while de-emphasizing others, creating a more streamlined visual experience. Gruendl et al. (2016) expanded on this concept by integrating time-series plots within PCPs, offering a hybrid approach that provides temporal context to multidimensional data. While these methods enhance interpretability, their reliance on sophisticated computational techniques may still overwhelm users in scenarios involving highly complex datasets.

Alternative visualization techniques aim to address the inherent limitations of PCPs. Mitku et al. (2020) proposed linked visualizations that combine scatter plots with parallel axes, offering complementary views of dimensional relationships. Similarly, Itoh et al. (2017) advocated for the interactive construction of low-dimensional PCPs derived from high-dimensional datasets. These methods reduce complexity while preserving essential data relationships, providing a simplified yet effective means of visual exploration.

While advanced interaction technologies such as virtual reality (VR) fall outside the primary scope of this discussion, they present an intriguing perspective on enhancing the usability of Parallel Coordinate Plots (PCPs). Tadeja, Kipouros, and Kristensson (2020) explored how immersive PCPs could be utilized in engineering design processes, allowing users to interactively navigate and manipulate high-dimensional data. This approach offers a novel way to comprehend relationships between dimensions, leveraging VR's spatial affordances to provide deeper insights. However, the significant infrastructure and training requirements for such solutions pose challenges to their broader adoption, making them a compelling, though currently impractical, avenue for exploration in the context of general visualization challenges.

Despite advancements in interaction and hybrid visualizations, isolating patterns in PCPs remains an ongoing challenge. Researchers like Moustafa (2011) have proposed mathematical transformations, such as density plots and edge bundling, to reduce clutter and enhance clarity. These methods aggregate overlapping paths and emphasize critical trends, making the visualizations more interpretable. However, such approaches can be unintuitive for non-expert users, creating a barrier to broader adoption.

Line Width Illusion Susceptibility

The line width illusion and its impact on data visualization, particularly in Parallel Coordinate Plots (PCPs), have garnered increasing attention in recent research. Variations in line width and color often introduce perceptual biases, leading viewers to overemphasize certain data points or trends that may lack statistical significance. Bok, Kim, and Seo (2020) noted the prevalence of such biases in PCPs, where non-standardized visual encodings can mislead viewers and compromise the accurate representation of multivariate data. Tyagi et al. (2022) further demonstrated that these perceptual distortions are particularly acute in dense datasets, where the interplay of line width and color exacerbates interpretive challenges. They propose interactive axis reordering as a dynamic solution, enabling users to refine visualizations and reduce the effects of such biases.

The influence of line width on perception extends beyond PCPs. Hofmann and Vendettuoli (2013) found that categorical visualizations, including parallel sets and hammock plots, are also vulnerable to these distortions. Their work underscores the importance of perception-true visualization designs that prioritize accuracy. Bulatov et al. (2024) expanded this inquiry by studying geometric visual illusions within PCPs. They concluded that consistent and deliberate plotting techniques can reduce misjudgments, advocating for standardized design practices to minimize perceptual errors.

Mitigating these biases remains a nuanced challenge. Liu et al. (2014) argued that while standardizing line width can decrease perceptual distortions, this approach may come at the cost of visual richness and aesthetic appeal. They emphasized the need for designers to balance clarity with the engaging qualities of visualizations. Viviani and Stucchi (1989) added complexity to this issue by demonstrating that movement velocity in animated visualizations can amplify size illusions, further complicating the interpretation of PCPs. Meanwhile, Henriques and Soechting (2003) highlighted individual variability in sensitivity to perceptual biases, suggesting that user-centered design and targeted training programs can help users navigate these challenges effectively.

Fermüller and Malm (2004) added another dimension to the discussion by exploring the role of top-down cognitive influences in amplifying perceptual distortions. Their findings suggest that user training to recognize and counteract biases is essential for improving interpretive accuracy. This perspective aligns with a broader consensus that visualization

tools must go beyond visual appeal to ensure interpretive reliability through thoughtful design and user education.

In summary, the susceptibility of PCPs to line-width illusions and related perceptual biases poses significant challenges for data visualization. Research points to standardization, user training, and interactive techniques as effective strategies to address these issues. However, achieving a balance between interpretive accuracy and visual engagement remains critical. As data visualization evolves, integrating these insights into design frameworks will be vital for creating robust, bias-resistant tools that support reliable data interpretation.

Despite these advancements, PCPs remain complex tools that demand careful optimization to balance interpretability, performance, and usability. Addressing dimensional overload, computational inefficiencies, and scalability will enhance their utility. Optimization techniques, such as dimensionality reduction, subspace projections, and efficient rendering strategies, offer promising solutions to these persistent issues. By refining the underlying mechanics of PCPs, these methods aim to streamline data visualization workflows and make high-dimensional analysis more accessible.

In the next section, we delve into optimizing PCPs and explore approaches that enhance their efficiency and effectiveness. From computational techniques to algorithmic advancements, these strategies promise to redefine the boundaries of what PCPs can achieve in high-dimensional data visualization.

Optimization of PCPs

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. 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

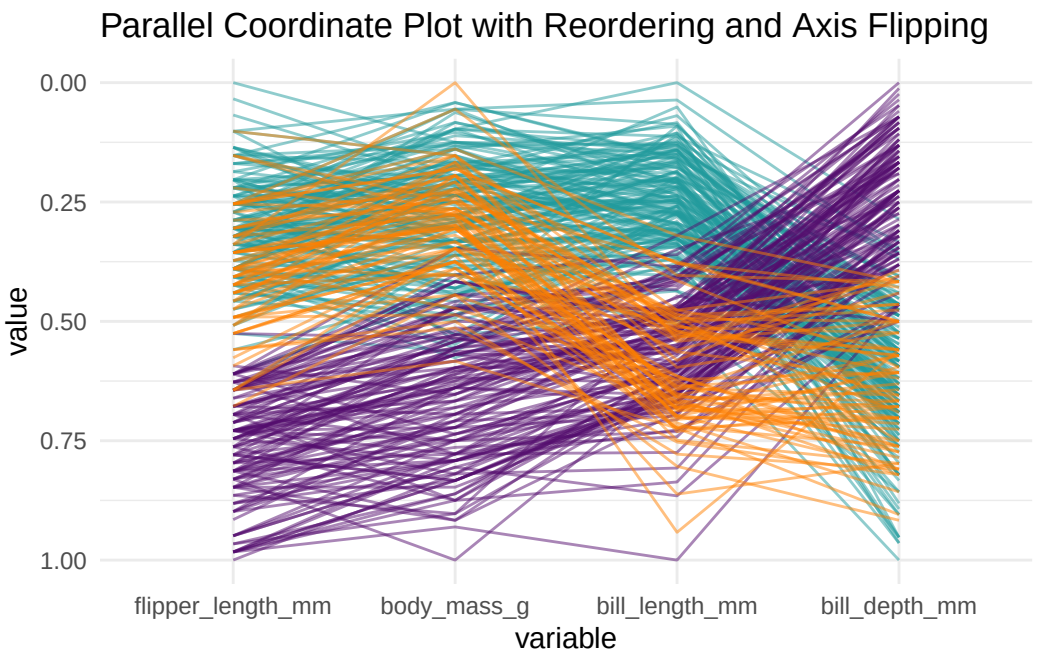


Figure 14: Reordering and Axis Flipping in PCP

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.

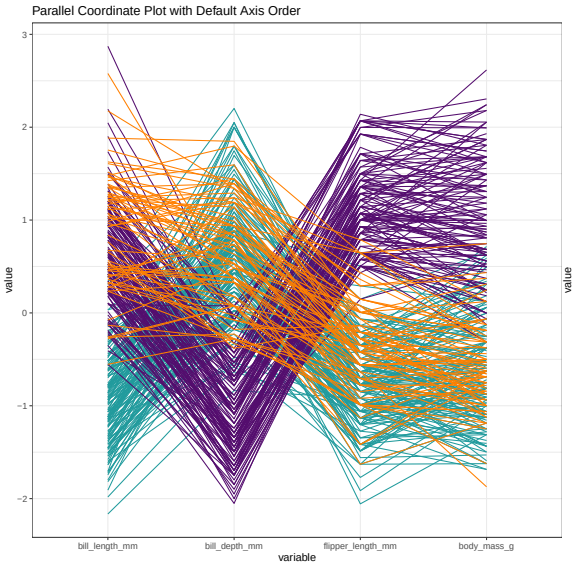


Figure 15

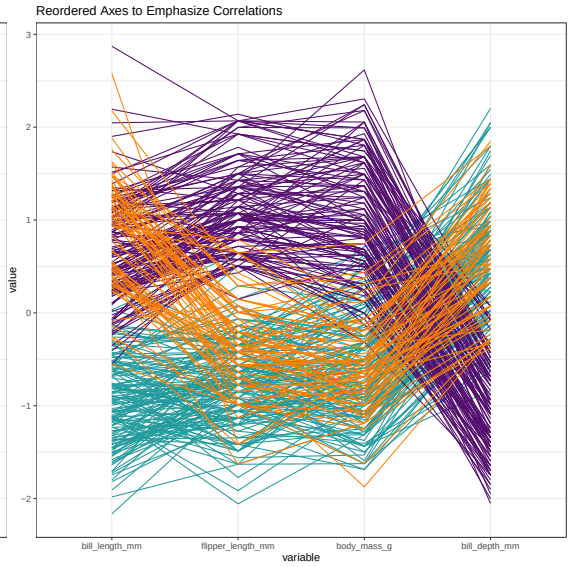


Figure 16

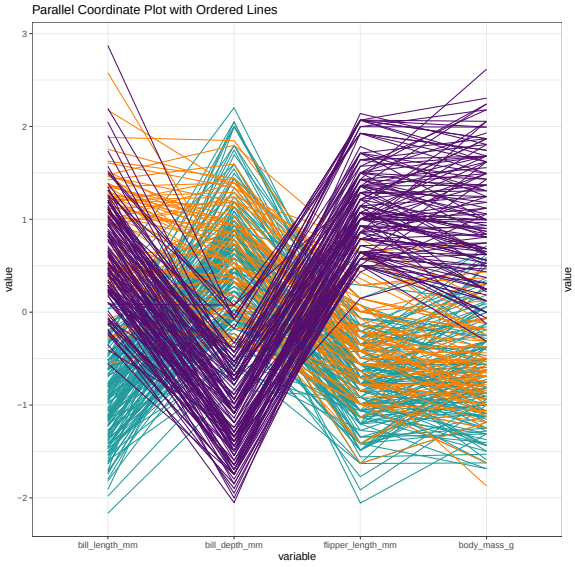


Figure 17

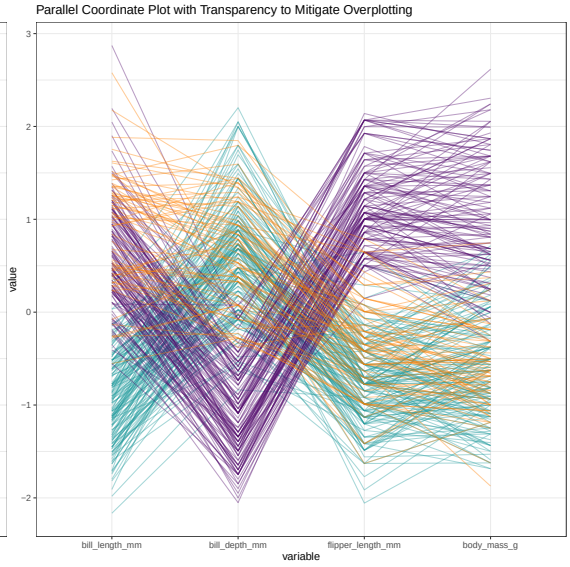


Figure 18

Pargnostics

Figure 15 shows the default axis arrangement without optimization. While the data is represented faithfully, the lack of deliberate axis order makes it challenging to discern meaningful relationships or clusters due to overlapping lines and clutter.

The order of axes in a PCP is a primary determinant of clarity. Dasgupta and Kosara (2010) introduced Pargnostics, a framework leveraging screen-space metrics to optimize axis arrangements. Pargnostics helps identify axis sequences that emphasize meaningful relationships between variables by focusing on metrics like clutter reduction and outlier detection. For instance, swapping adjacent axes can unveil previously obscured correlations, enabling users to discern patterns and trends better. This technique has proven effective across diverse applications, including climate data analysis and medical diagnostics, where understanding variable interdependencies is most important. By reordering the axes based on meaningful relationships between variables, Figure 16 highlights correlations that were previously obscured. For example, adjacent placement of highly correlated dimensions simplifies the identification of patterns, making trends more accessible to the observer.

In addition to axis ordering, the arrangement of lines within PCPs plays a crucial role in mitigating visual clutter and improving pattern visibility. As developed by Dasgupta and Kosara (2011), adaptive sorting techniques use clustering algorithms to reorder lines dynamically. By prioritizing critical data subsets and de-emphasizing less relevant observations, these methods enable users to focus on regions of interest while maintaining the integrity of the broader dataset. This approach is particularly valuable in high-density datasets, where overlapping lines obscure important relationships. Figure 17 showcases line ordering based on clustering techniques. Critical data subsets are prioritized, with similar trajectories grouped together to reduce visual clutter and emphasize key patterns. This approach enhances the viewer's ability to focus on important data points while preserving overall dataset integrity.

The drawing order of lines in PCPs determines which data points are most visible in overlapping regions. Julian Heinrich and Weiskopf (2013b) emphasized the importance of visualization layering, where higher-priority lines are drawn on top to prevent occlusion by less critical data. Techniques such as transparency adjustments and overplotting mitigation further enhance legibility, ensuring that essential patterns and outliers remain prominent. These methods allow users to navigate complex datasets more precisely, even when dealing with dense or highly variable data.

Integrating line and axis ordering strategies creates a synergistic effect, optimizing cluster detection and correlation identification. Firat, Swallow, and Laramee (2023) demonstrate the effectiveness of these combined approaches in ensemble data visualization and anomaly detection. These methods achieve a higher degree of clarity and functionality by simultaneously addressing axes and line arrangements, making PCPs more versatile and impactful across various fields. Adjusting line transparency effectively mitigates the effects of overplotting in dense datasets, Figure 18. Subtle variations in opacity help to reveal underlying structures and clusters, ensuring that both prominent and less frequent patterns remain visible without overwhelming the visualization.

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.

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 (2012a). 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 (2012b) 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.

Interactivity

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 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.

Handling Numerical Ties

The challenge of managing tied values in statistical data analysis has long baffled 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 longstanding challenges in statistical analysis.

Edward R. 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.

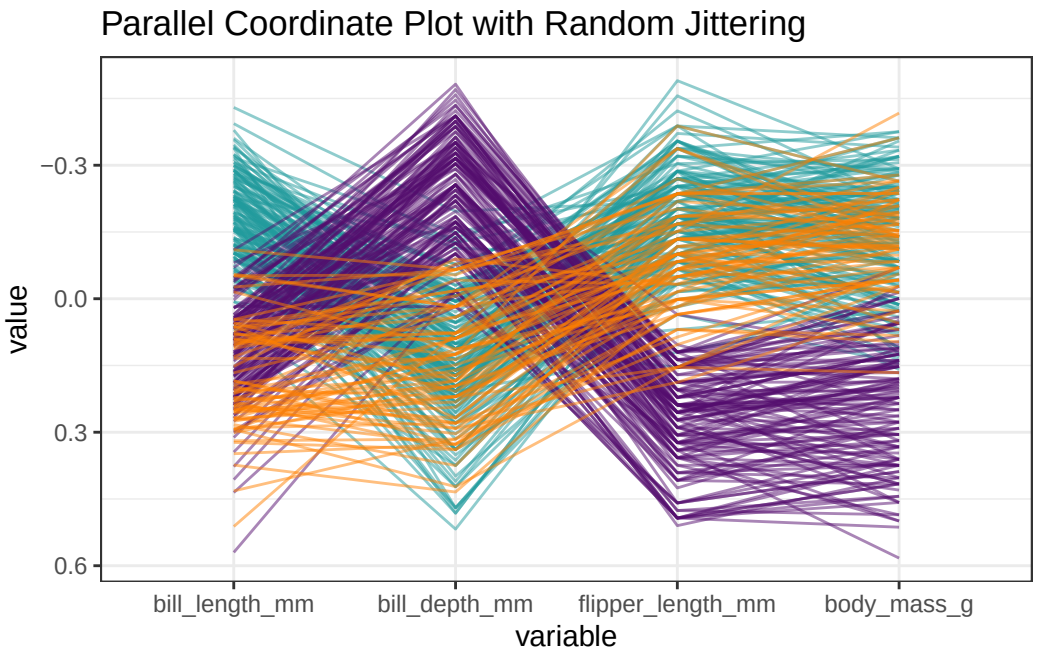


Figure 19: Random Jittering of Data Points in PCP

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

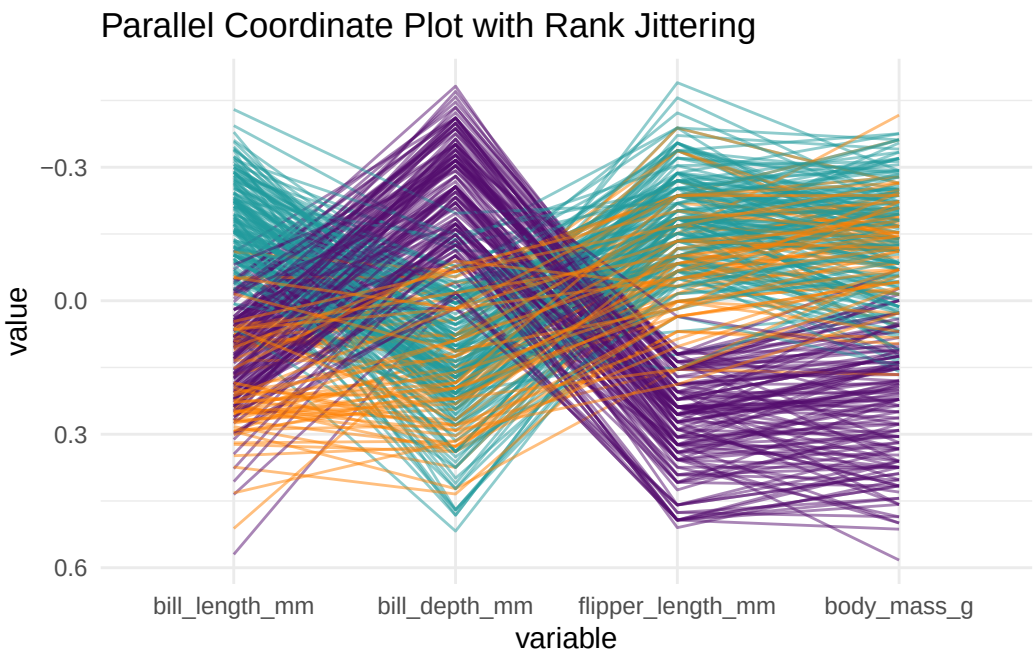


Figure 20: Rank Jittering of Data Points in PCP

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 es-

pecially 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.

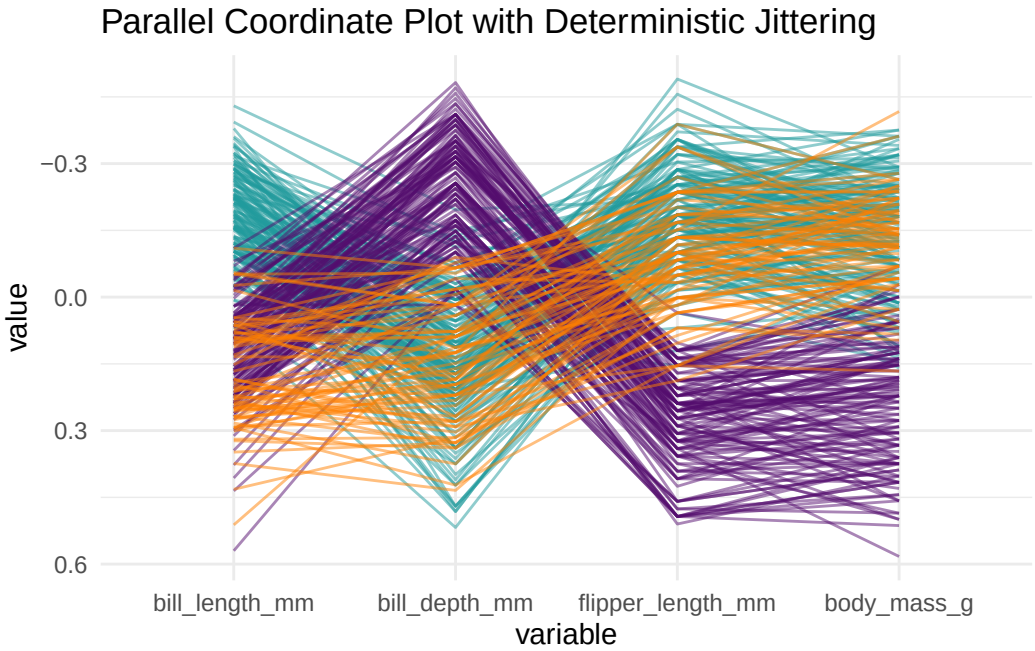


Figure 21: Deterministic Jittering of Data Points in PCP

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:

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

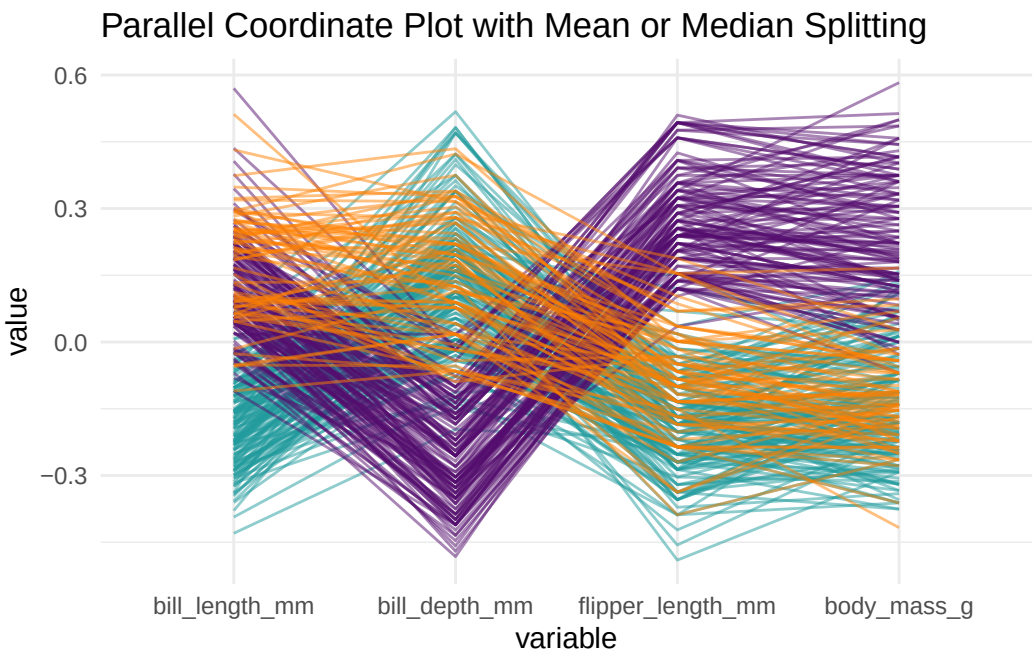


Figure 22: Mean or Median Splitting of Data Points in PCP

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.

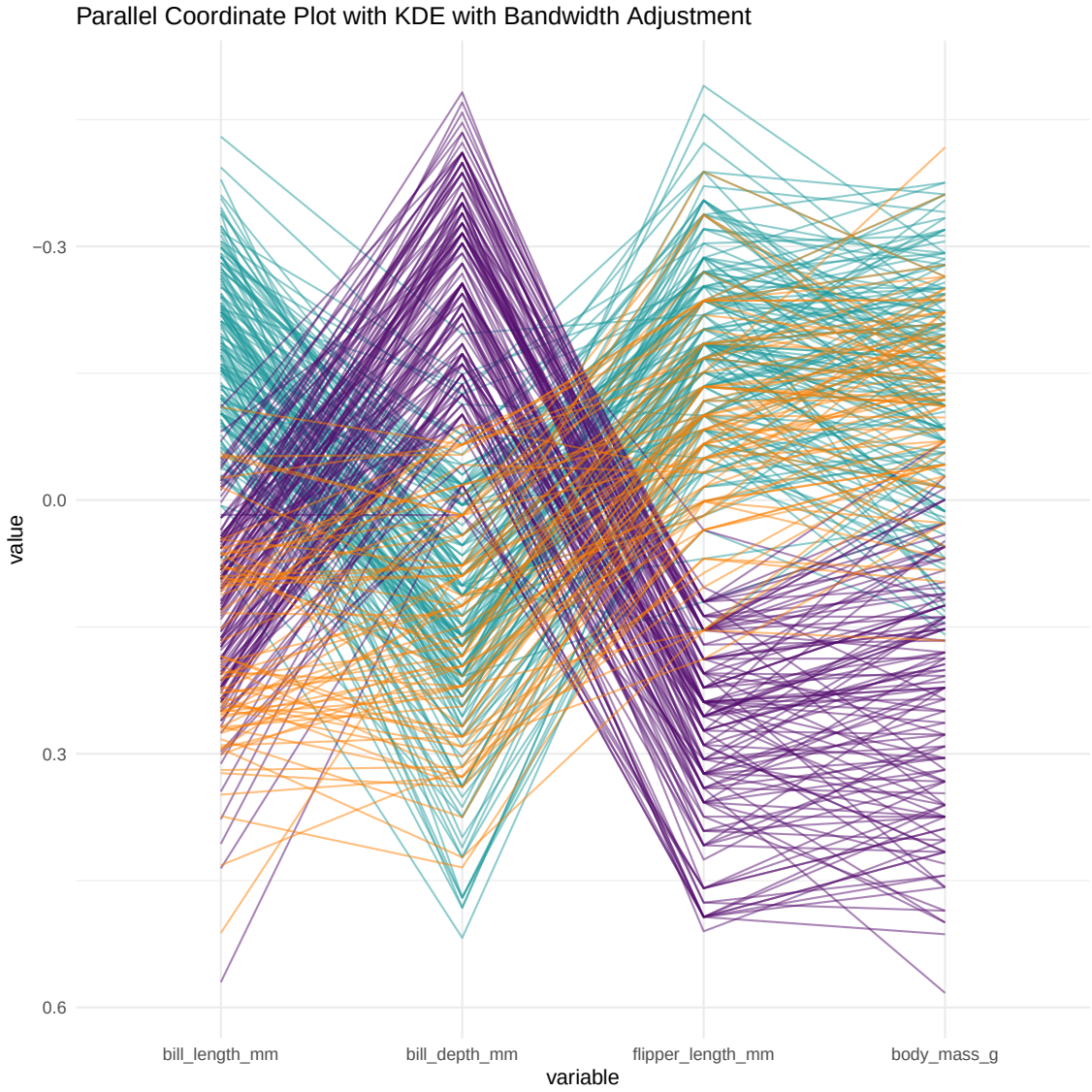


Figure 23: KDE with Bandwidth Adjustment of Data Points in PCP

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.

Handling Ties in PCPs

The growing complexity and volume of datasets have fueled the advancement of data visualization approaches, which require creative ways to blend clarity and detail. Early visualization techniques underlined closeness and alignment to expose patterns and linkages, especially in network research and multidimensional scaling. Effective for smaller datasets, these methods failed as the data grew in complexity and scale, resulting in problems, including visual clutter.

Analyzing high-dimensional data has found great use for parallel coordinate plots (PCPs). But overlapping lines and numerous intersections—which hide important data structures—often compromise their efficacy. Especially with big datasets, this visual complexity makes it challenging for users to spot trends and linkages.

Early studies focused on raising "visual distance"—the geographical separation of visual elements—" to address overplotting in PCPs." On page 201, Chang, Dwyer, and Marriott (2018), for example, compared PCPs with scatterplot matrices (SPLOMs) and discovered that complicated cross-line patterns caused users to find difficulty identifying correlations in PCPs. They found, "In PCP, participants had to reorder the axes to examine all dimensions, whereas, in SPLOM, it is straightforward to identify correlation." This shows how improving visual distance between overlapping items simplifies cognitive grouping without upsetting the continuity of the material.

Researchers have devised several techniques to increase visual distance in PCPs: Line bundling: This method clusters related pathways, minimizing the visual anarchy by overlapping lines. Bundling may hide unusual pathways or outliers even when it helps to visualize things more easily. Bundling is sometimes used with transparency methods or line spacing changes to help minimize this and guarantee that important data points are still clear.

Introduced by Guo et al. (2005), the nested-means approach method methodically avoids overlapping lines by ensuring the mean value of every variable is centrally situated on its axis. This method solves overplotting and maintains data-central tendencies, improving analytical value and clarity.

Changing the spacing between lines or reordering axes to align relevant dimensions lowers cognitive burden and facilitates tracing links between variables. Early studies showed that these changes enhanced users' capacity to recognize trends across dimensions and interpret intersections.

Visual distance offers cognitive advantages beyond only aesthetics. Reducing visual clutter helps users spot trends, patterns, and outliers more successfully. These techniques act as cognitive tools, focusing on important components of challenging datasets. For PCPs, for instance, adding small separations between connected values highlights intersections and relationships, allowing more accurate analysis without changing the fundamental data structure. It is especially helpful in maintaining the integrity of the graphic and stressing important trends such as line bundling and spacing changes. These methods guarantee that anomalies or unusual trends remain obvious and enable users to concentrate on the general structure of the data.

Visual distance has evolved outside PCPs to various visualization methods as datasets become more complex. Techniques improving visual distance help scatterplots, network graphs, and heat maps. Dynamic interactivity—brushing, filtering, and axis flipping—allows users to explore relationships across dimensions more naturally, improving the usability of current visualizations.

For example, Schonlau (2003) modify visual distance ideas using proportional box widths to show numerical and categorical data. Likewise, Wegman and Luo (1997) work on parallel coordinate density plots uses transparency and density changes to handle overplotting in high-dimensional data.

Visual distance is today understood as a fundamental element of good data visualization. Visual distance approaches simplify cognitive processing and help analysts find insights in even the most complicated datasets by lowering perceptual overload. These developments help current visualizations achieve a careful mix of clarity and detail, boosting confidence in data interpretation.

Data visualization keeps changing by extending early research and improving these methods, thereby arming users with strong instruments to make sense of their data. Visual distance stays first as both a spatial and a cognitive tool, improving our capacity to negotiate and understand ever-complex datasets.

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.

Akbar, Muhammad Sajjad, and Bogdan Gabrys. 2018. "Data Analytics Enhanced Data Visualization and Interrogation with Parallel Coordinates Plots." In *2018 26th International Conference on Systems Engineering (ICSEng)*, 1–7. IEEE.

Arif, Muhammad, and Saleh Basalamah. 2012. "Similarity-Dissimilarity Plot for High Dimensional Data of Different Attribute Types in Biomedical Datasets." *International Journal of Innovative Computing, Information and Control* 8 (2): 1275–97.

Bendix, Fabian, Robert Kosara, and Helwig Hauser. 2005. "Parallel Sets: Visual Analysis of Categorical Data." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 133–40. IEEE.

Berkowitz, Bruce. 2018. *Playfair: The True Story of the British Secret Agent Who Changed How We See the World*. Fairfax, Virginia: George Mason University.

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.

Bulatov, Aleksandr, Vilius Marma, Natalija Bulatova, and Artūras Grigaliūnas. 2024. "Combined Manifestation of Two Geometric Visual Illusions." *Attention, Perception, & Psychophysics* 86 (7): 2456–74.

Cao, Nan, Yu-Ru Lin, and David Gotz. 2015. "Untangle Map: Visual Analysis of Probabilistic Multi-Label Data." *IEEE Transactions on Visualization and Computer Graphics* 22 (2): 1149–63.

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

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

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

Cheng, Shenghui. 2018. "Visual Analytics for Relation Discovery in Multivariate Data." PhD thesis, State University of New York at Stony Brook.

Chung, David HS, Philip A Legg, Matthew L Parry, Rhodri Bown, Iwan W Griffiths, Robert S Laramee, and Min Chen. 2015. "Glyph Sorting: Interactive Visualization for Multi-Dimensional Data." *Information Visualization* 14 (1): 76–90.

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.

———. 2011. "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2241–48.

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

Day, Ross H, and Erica J Stecher. 1991. "Sine of an Illusion." *Perception* 20: 49–55.

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.

Fermüller, Cornelia, and Henrik Malm. 2004. "Uncertainty in Visual Processes Predicts Geometrical Optical Illusions." *Vision Research* 44 (7): 727–49.

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

Firat, Elif E, Alena Denisova, Max L Wilson, and Robert S Laramee. 2022. "P-Lite: A Study of Parallel Coordinate Plot Literacy." *Visual Informatics* 6 (3): 81–99.

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

Friendly, Michael, and Daniel J Denis. 2001. "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization." *URL Http://Www.datavis.ca/Milestones* 32: 13.

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

Garrison, Laura, Juliane Müller, Stefanie Schreiber, Steffen Oeltze-Jafra, Helwig Hauser, and Stefan Bruckner. 2021. "Dimlift: Interactive Hierarchical Data Exploration Through Dimensional Bundling." *IEEE Transactions on Visualization and Computer Graphics* 27 (6): 2908–22.

Gower, John C, and Garmt B Dijksterhuis. 2004. *Procrustes Problems*. Vol. 30. OUP Oxford.

Gruendl, Henning, Patrick Riehmann, Yves Pausch, and Bernd Froehlich. 2016. "Time-Series Plots Integrated in Parallel-Coordinates Displays." In *Computer Graphics Forum*, 35:321–30. 3. Wiley Online Library.

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): 113–32.

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, 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.

Hemming, Sidney R. 2004. "Heinrich Events: Massive Late Pleistocene Detritus Layers of the North Atlantic and Their Global Climate Imprint." *Reviews of Geophysics* 42 (1).

Henriques, Denise YP, and John F Soechting. 2003. "Bias and Sensitivity in the Haptic Perception of Geometry." *Experimental Brain Research* 150 (1): 95–108.

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.

———. 2016. *Ggparallel: Variations of Parallel Coordinate Plots for Categorical Data*. https://cran.r-project.org/package=ggparallel.

Holtz, Yan. 2024. "Sankey Diagram. Data to Viz." June 24, 2024. https://www.data-to-viz.com/graph/www.data-to-viz.com/caveat/sankey.html.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

Hurley, Catherine B. 2004. "Clustering Visualizations of Multidimensional Data." *Journal of Computational and Graphical Statistics* 13 (4): 788–806.

Hurter, Christophe. 2017. "Antoine Lhuillier." PhD thesis, Université Toulouse 3 Paul Sabatier.

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.

Itoh, Takayuki, Ashnil Kumar, Karsten Klein, and Jinman Kim. 2017. "High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots." *Journal of Visual Languages & Computing* 43: 1–13.

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.

Jones, Richard A, Marcos R Perez-Brayfield, Andrew J Kirsch, and J Damien Grattan-Smith. 2004. "Renal Transit Time with MR Urography in Children." *Radiology* 233 (1): 41–50.

Kaur, Gurminder, and Bijaya B Karki. 2018. "Bifocal Parallel Coordinates Plot for Multivariate Data Visualization." In *VISIGRAPP (3: IVAPP)*, 176–83.

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

Kennedy, A B W, and H R Sankey. 1898. "The Thermal Efficiency of Steam Engines. Report of the Committee Appointed to the Council Upon the Subject of the Definition of a Standard or Standards of Thermal Efficiency for Steam Engines: With an Introductory Note. (Including Appendixes and Plate at Back of Volume)." *Minutes of the Proceedings of the Institution of Civil Engineers* 134 (1898): 278–312. https://doi.org/10.1680/imotp.1898.19100.

Kennedy, Alex BW, and H Riall Sankey. 1898. "THE THERMAL EFFICIENCY OF STEAM ENGINES. REPORT OF THE COMMITTEE APPOINTED TO THE COUNCIL UPON THE SUBJECT OF THE DEFINITION OF a STANDARD OR

STANDARDS OF THERMAL EFFICIENCY FOR STEAM ENGINES: WITH AN INTRODUCTORY NOTE.(INCLUDING APPENDIXES AND PLATE AT BACK OF VOLUME).” In *Minutes of the Proceedings of the Institution of Civil Engineers*, 134:278–312. 1898. Thomas Telford-ICE Virtual Library.

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.

Kwak, Il-Youp, and Myung-Hoe Huh. 2008. “Parallel Coordinate Plots of Mixed-Type Data.” *Communications for Statistical Applications and Methods* 15 (4): 587–95.

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.

Lhuillier, Antoine, Christophe Hurter, and Alexandru Telea. 2017. “State of the Art in Edge and Trail Bundling Techniques.” In *Computer Graphics Forum*, 36:619–45. 3. Wiley Online Library.

Liu, Shixia, Weiwei Cui, Yingcai Wu, and Mengchen Liu. 2014. “A Survey on Information Visualization: Recent Advances and Challenges.” *The Visual Computer* 30: 1373–93.

Ma, Chao, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. 2020. “VAEM: A Deep Generative Model for Heterogeneous Mixed Type Data.” *Advances in Neural Information Processing Systems* 33: 11237–47.

Markwick, Paul J, and Paul J Valdes. 2004. “Palaeo-Digital Elevation Models for Use as Boundary Conditions in Coupled Ocean–Atmosphere GCM Experiments: A Maastrichtian (Late Cretaceous) Example.” *Palaeogeography, Palaeoclimatology, Palaeoecology* 213 (1-2): 37–63.

McColeman, Caitlyn. 2017. “Exploring Human Cognition Through Multivariate Data Visualization.”

Meka, Leon. 2024. “Line Perception in Parallel Coordinates Under Different Aspect Ratios.” PhD thesis, Technische Universität Wien.

Mitku, Aweke A, Temesgen Zewotir, Delia North, and Rajen N Naidoo. 2020. “Exploratory Data Analysis of Adverse Birth Outcomes and Exposure to Oxides of Nitrogen Using Interactive Parallel Coordinates Plot Technique.” *Scientific Reports* 10 (1): 7363.

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

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

Nowak, Stan. 2023. “Designing for Ambiguity in Sensemaking: Visual Analytics in Risk Analysis and Prediction.”

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

Pedersen, Thomas Lin. 2024. “Create Parallel Sets Diagrams — Geom_parallel_sets. Ggforce Documentation.” January 22, 2024. https://ggforce.data-imaginist.com/reference/geom_parallel_sets.html.

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

———. 2012a. “Comparing Clusterings Using Bertin’s Idea.” *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2510.

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. 2019. “Visual Analytics for the Representation, Exploration,

and Analysis of High-Dimensional, Multi-Faceted Medical Data." *Biomedical Visualisation: Volume 2*, 137–62.

Resch, Gabriel. 2019. *Denaturalizing Information Visualization*. University of Toronto (Canada).

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.

Rzeszotarski, Jeffrey M. 2017. "Uncovering Nuances in Complex Data Through Focus and Context Visualizations." PhD thesis, Carnegie Mellon University.

Sartorio, A, M Proietti, PG Marinone, F Agosti, F Adorni, and CL Lafortuna. 2004. "Influence of Gender, Age and BMI on Lower Limb Muscular Power Output in a Large Population of Obese Men and Women." *International Journal of Obesity* 28 (1): 91–98.

Schmidt, Mario. 2008. "The Sankey Diagram in Energy and Material Flow Management Part 1: History." *Journal of Industrial Ecology* 12 (1): 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x.

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.

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

Shaham, Yoav. 2015. "Visualizing Mixed Variable-Type Multidimensional Data Using Tree Distances." PhD thesis, Monterey, California: Naval Postgraduate School.

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

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

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

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

States Census Office", "United, and Henry Gannett. 1898. *Statistical Atlas of the United States Based on the 11th Decennial Census*. Washington DC. https://www.loc.gov/resource/g3701gm.gct00010/?sp=36.

Stumpfegger, Josef, Kevin Höhlein, George Craig, and Rüdiger Westermann. 2022. "GPU Accelerated Scalable Parallel Coordinates Plots." *Computers & Graphics* 109: 111–20.

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."

Tadeja, Slawomir K, Timoleon Kipouros, and Per Ola Kristensson. 2020. "IPCP: Immersive Parallel Coordinates Plots for Engineering Design Processes." In *AIAA Scitech 2020 Forum*, 0324.

Tortora, Cristina, and Francesco Palumbo. 2024. "FPDclustering: A Comprehensive r Package for Probabilistic Distance Clustering Based Methods." *Computational Statistics*, 1–24.

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

Tufte, Edward R, and Peter R Graves-Morris. 1983. *The Visual Display of Quantitative Information*. Vol. 2. 9. Graphics press Cheshire, CT.

Tyagi, Anjul, Tyler Estro, Geoff Kuenning, Erez Zadok, and Klaus Mueller. 2022. "Pc-Expo: A Metrics-Based Interactive Axes Reordering Method for Parallel Coordinate Displays." *IEEE Transactions on Visualization and Computer Graphics* 29 (1): 712–22.

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.

Viviani, Paolo, and Natale Stucchi. 1989. "The Effect of Movement Velocity on Form Perception: Geometric Illusions in Dynamic Displays." *Perception & Psychophysics* 46 (3): 266–74.

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, Tiancheng. 2023. "Hammock-Plot: Hammock - Visualization of Categorical or Mixed Categorical/Continuous Data." https://github.com/TianchengY/hammock_plot.

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.