A Grammar of Graphics Approach to Managing Numerical Ties in Parallel Coordinate Plots

Denise Bradford

2026-02-01

Table of contents

1 Introduction and Motivation

Parallel coordinate plots (PCPs), first popularized by Alfred Inselberg, are a powerful technique for investigating patterns across multiple attributes (variables) simultaneously (Inselberg 2009). PCPs assign each dimension of an n-dimensional dataset to a vertical axis. Each vertical axis is then positioned along one dimension, arranged in parallel (Wegman 1990). Observations are drawn as polylines connecting a single point along the sequence of dimensional axes, providing a different perspective on hard-to-visualize multidimensional data that can be used to identify clusters, outliers, and other facets of a dataset.

When multiple observations share the same value in a given dimension, their polylines perfectly overlap, creating "visual collisions." This masks information about the joint distribution between variables as well as the density along a single axis. The treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg (1985) and Wegman (1990).

PCPs, which were canoncially used to show continuous random variables, have always had this issue, and the use of PCPs for categorical data only emphasized the problem. Categorical parallel axis plot solutions, such as Sankey[1] and Alluvial diagrams, Categorical PCPs (Pilhöfer and Unwin 2013), and Parallel Sets (Kosara, Bendix, and Hauser 2006) plots all developed (largely independently!) as

[1]Sankey diagrams do not always have parallel axes, but they are visually similar and frequently use parallel axis constructions.

methods for using parallel axes to represent categorical, bivariate frequency information. Variations on these plots, such as common angle plots (Hofmann and Vendettuoli 2013)

Hammock plots (Schonlau 2003), another type of PCP that accommodates categorical and continuous variables, shows multiple bivariate relationships using parallelograms drawn between axes, obviating the need for tie-breaking and preserving the density information. Because only bivariate relationships are shown, it is not possible to mentally reconstruct the full joint distribution using a hammock plot.

The development of Generalized PCPs (and the software implementation `ggpcp` (VanderPlas et al. 2023)) represent a major step forward in representing the full joint distribution of the data while showing individual observations. One innovation in the `ggpcp` package is the handling of ties in continuous variables: ties are broken and the observations are spread out within the box representing the marginal frequency of each value of the categorical variable. Sorting methods are implemented to distribute the tied values in a way that reduces the overall complexity of the plot, as random line-crossings can make PCPs difficult to read and interpret.

This project aims to extend the treatment of categorical variables in generalized PCPs, implementing a solution for the treatment of continuous variables which has the following properties:

- marginal density information (approximately) faithfully represented on the vertical axis
- individual lines can be visually distinguished, at least for data where the number of observations does not induce overplotting.

An immediate consequence of the second objective is that the viewer can, at least in theory, reconstruct the full joint distribution between variables from the plot. We will extend the method used in `ggpcp` to represent categorical data to continuous variables, modifying it to maintain the approximate scaling of the continuous variable, with slight local distortions when tied values occur. In addition, we will develop a consistent visual representation to provide visual cues which indicate the presence of a local distortion due to tied values, while preserving the visual representation of singular values on the parallel axis.

This combination of the advantage of Hammock plots – maintaining the marginal density information – and generalized PCPs, which preserve individual observations and provide the ability to reconstruct the full joint distribution, will enhance GPCPs. Real-world data sets frequently have both categorical and numerical data, and duplicated values in either type of variable. Our extension to GPCPs will facilitate the visual representation of these data sets.

1.1 Categorical Variables and Ties in `ggpcp`

The `ggpcp` package currently addresses categorical ties through a combination of sorting and tie-breaking algorithms. The package implements hierarchical sorting through the `pcp_arrange(data, method, space)` function, which orders observations based on a hierarchical application of variable values. The `method` parameter determines the sequence in which variables are considered when resolving ties in the arrangement. The `space` parameter specifies the proportion of the y-axis dedicated to empty space between levels of categorical variables (Figure 1).

```
# 1. Modular Data Pipeline (The correct Step 1, 2, 3)
# Note: 'pcp_data' is used here as a variable name for the transformed object.
pcp_data <- mtcars %>%
  mutate(across(c(cyl, am, gear, carb), as.factor)) %>%
  pcp_select(cyl, am, gear, carb) %>%     # Step 1: Reshape data
  pcp_scale(method = "uniminmax") %>%     # Step 2: Scale axes
  pcp_arrange(method="from-left")         # Step 3: Break Ties,
                                          #         sorting from left
```

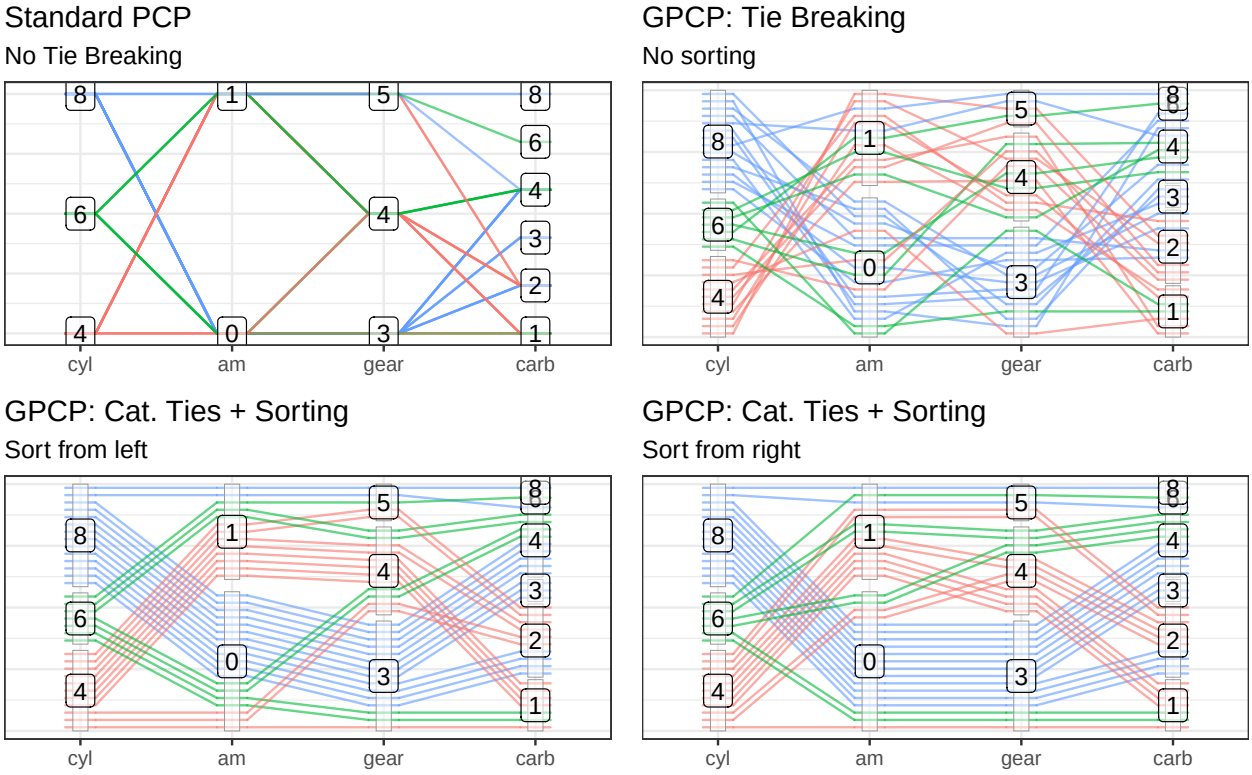


Figure 1: Comparison of tie-breaking methods for categorical variables in parallel coordinate plots using mtcars data. Standard PCP (top-left) shows overlapping lines without tie-breaking. GPCP with tie-breaking but no sorting (top-right) spreads observations evenly within categories. GPCP with sorting from left (bottom-left) and from right (bottom-right) apply hierarchical sorting to minimize line crossings, with light gray boxes indicating category groupings.

The ability to follow individual observations is key to being able to reconstruct the full joint distribution across parallel axes. `ggpcp` creates equally spaced points within each category that span the portion of the vertical axis dedicated to the category (preserving the marginal frequency information, which is optionally emphasized with `geom_pcp_boxes`). The top left plot in Figure 1 shows the PCP without this categorical tie-breaking approach. However, `ggpcp` goes one step further by using hierarchical sorting (from the left, right, or both) to minimize unnecessary line crossings induced by the treatment of tied categorical variables. The absence of the hierarchical sorting approach is emphasized in the top right plot in Figure 1, while the bottom left and right plots show sorting from the left and right respectively. This hierarchical sorting approach serves as a form of "external cognition," reducing the cognitive load required to "untangle" and group crossed

lines. The sorting also ensures that only necessary crossings[2] occur, minimizing visual clutter.

It is worthwhile to examine how a viewer would follow a line across the plot in detail, using a generalized PCP with both numeric and categorical variables to illustrate the challenges introduced when numerical ties are present.

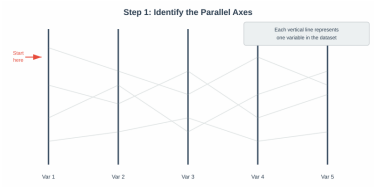


Figure 2: The parallel axes form the structural framework of the visualization.

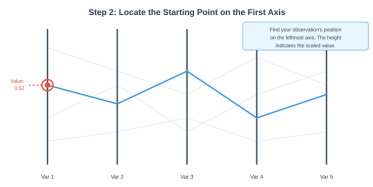


Figure 3: The starting point is identified on the first axis with its corresponding value.

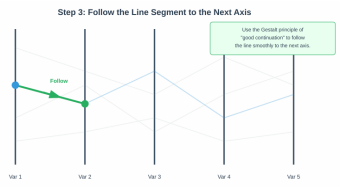


Figure 4: Following the line segment uses the Gestalt principle of good continuation.

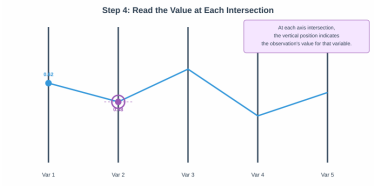


Figure 5: Values are read as approximate scaled value relative to the total range.

Figure 6: The complete traced path reveals the observation's values across all variables.

- Step 1: Identify the Parallel Axes (Figure 2).
  Begin by identifying each vertical axis in the plot. Each axis represents one variable from the dataset. The axes are typically arranged from left to right, and the order may be determined by the data analyst to highlight specific relationships or minimize visual clutter.

- Step 2: Locate the Starting Point (Figure 3).
  Find the observation of interest on the leftmost axis. The vertical position indicates the scaled value of that observation for the first variable. If you are examining a highlighted or color-coded observation, look for its distinctive marker at this starting position.

- Step 3: Follow the Line Segment (Figure 4).
  Trace the line segment from the starting point to its intersection with the next axis. The human visual system naturally follows smooth, continuous paths due to the Gestalt principle of good continuation, which allows viewers to perceive connected line segments as a single, piecewise continuous line, making it easier to track observations across multiple parallel axes.

- Step 4: Read Values at Intersections (Figure 5).
  At each axis intersection, the vertical position of the line indicates the observation's value for that variable. Read these values to understand how the observation changes (typically, relative to the other observations) across different dimensions of the data. The slope of

[2] A crossing is necessary if it is induced by the order of x-axis variables and scaled values, rather than by the treatment of categorical variables.

line segments between axes provides information about the relationship between consecutive variables for that specific observation.

- Step 5: Continue Across All Axes (Figure 6).
  Repeat the tracing process for each consecutive pair of axes until reaching the rightmost axis. By following the complete path, you obtain a comprehensive view of how that particular observation behaves across all measured variables. The vertical positions along each axis represent scaled values that can be interpreted as quantiles when the data are appropriately transformed. This enables identification of unique characteristics, cluster membership, or outlier status.

- Step 6: Numerical Ties – a fork in the road.
  Describe why this happens and what the problem is, showing a different line on the same chart.

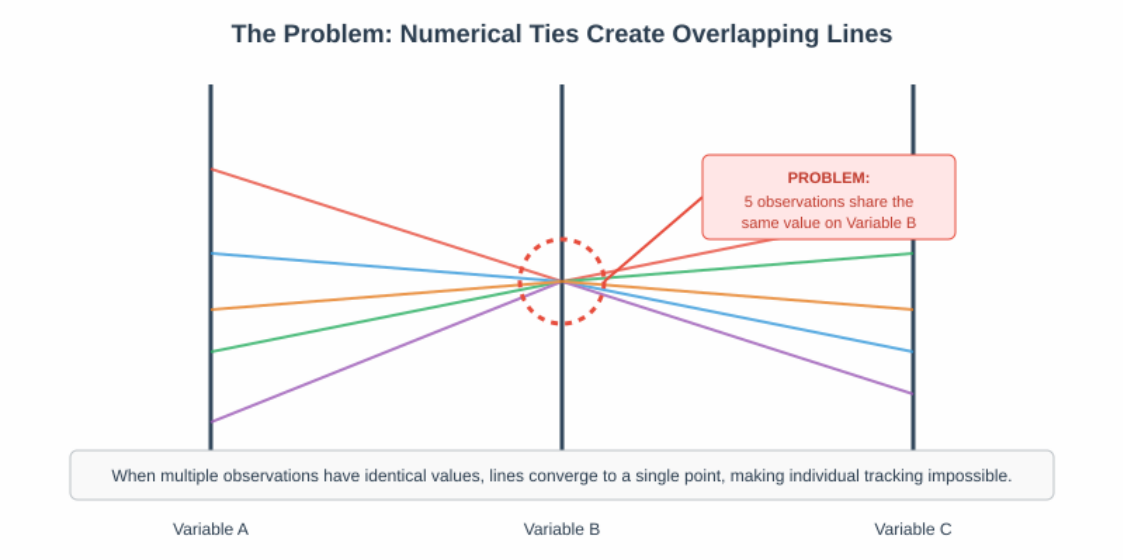


Figure 7: Numerical ties cause lines to converge at a single point.

- Step 6: Ties in Numbers — A Fork in the Road (Figure 7).

What was once a single point now becomes the starting point for two separate line segments that go in different directions. This split is a tie in numbers, which makes it hard to understand. The Gestalt principle of good continuation doesn't work anymore when two or more lines come from the same point on an axis. This makes it hard for the viewer to figure out which outgoing segment belongs to which incoming path. The visual continuity that made tracing easy in earlier steps breaks down exactly where it is needed most. The observer cannot tell if the original observation follows the upper branch or the lower branch of the fork unless there are more visual cues, like color coding, clear labels, or a systematic tie-breaking arrangement. This lack of clarity is not just an aesthetic problem; it also makes it harder to do the main analytical job of the parallel coordinate plot, which is to show patterns that would be hard to see in tabular data by following individual observations through high-dimensional space.

1.2 Addressing Numerical Ties

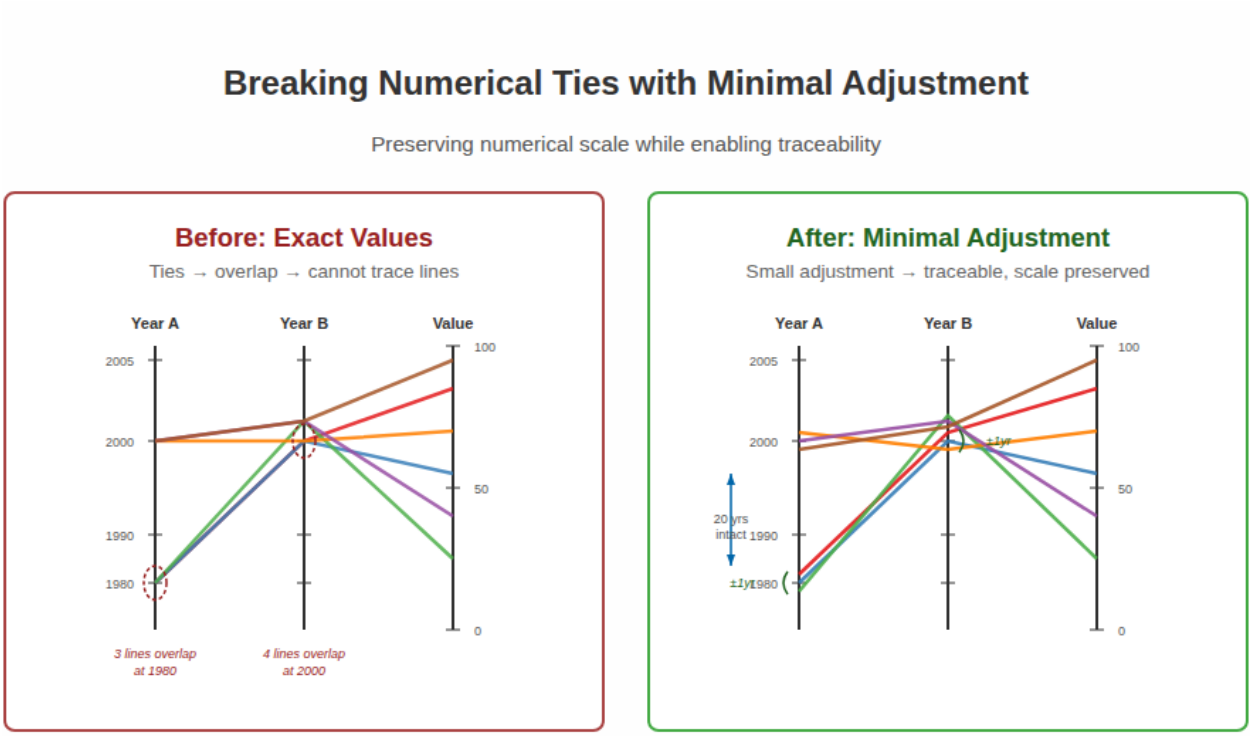


Figure 8: Comparison of standard parallel coordinate plot representation and the Even Tie Spreading Algorithm for resolving numerical ties.

The ggpcp package implements a sophisticated tie-breaking algorithm for categorical variables that maintains individual observation traceability. The approach spaces observations evenly within categorical levels:

> "All observations are spaced out evenly. This results in a natural visualization of the marginal frequencies along each axis (additionally enhanced by the light gray boxes grouping observations in the same category) that is not as prominent in the previous three panels. The ordering of the observations within the level is such that a minimal number of line crossings occurs between the axes." (p. 11)

The algorithm achieves this through hierarchical sorting implemented in `pcp_arrange(data, method, space)`, where the `space` parameter specifies the proportion of the y-axis used for spacing between categorical levels. This optimization can be formalized as:

$$d_i = \frac{S_i - S_i^- - S_i^+}{n_i - 1}$$

where:

- S_i is the total space allocated to category i
- S_i^- is the spacing below category i

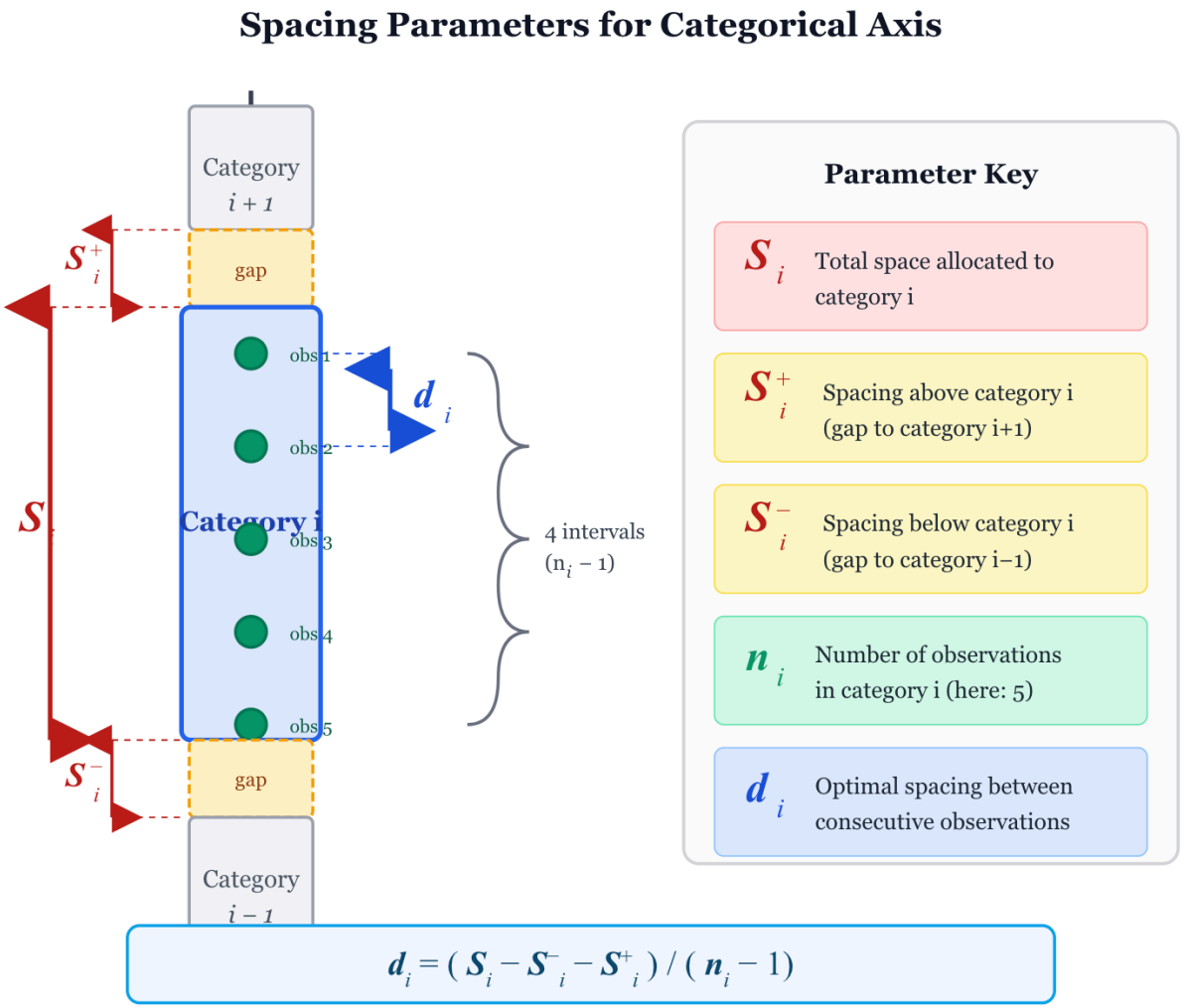


Figure: The optimal inter-observation spacing d_i is computed by subtracting the inter-category gaps from the total allocated space, then dividing by the number of intervals between observations.

Figure 9: Spacing parameters for Categorical Axis

- S_i^+ is the spacing above category i
- n_i is the number of observations in category i
- d_i is the optimal spacing distance between consecutive observations

---

A key visual difference emerges when connecting categorical to numerical variables. Schonlau and Yang (2024) observe:

> "When many observations have the same value for a categorical and an adjacent numerical variable, the corresponding area looks like a triangle… Notice the lines/boxes between the variables hospitalizations and comorbidities in the GPCP (Figure 13) and hammock plots (Figure 2). Most of the observations are in the boxes leading from hospitalizations=0 to either comorbidities=0 or comorbidities=1. This is far more obvious in the hammock plot than in the GPCP plot." (p. 19)

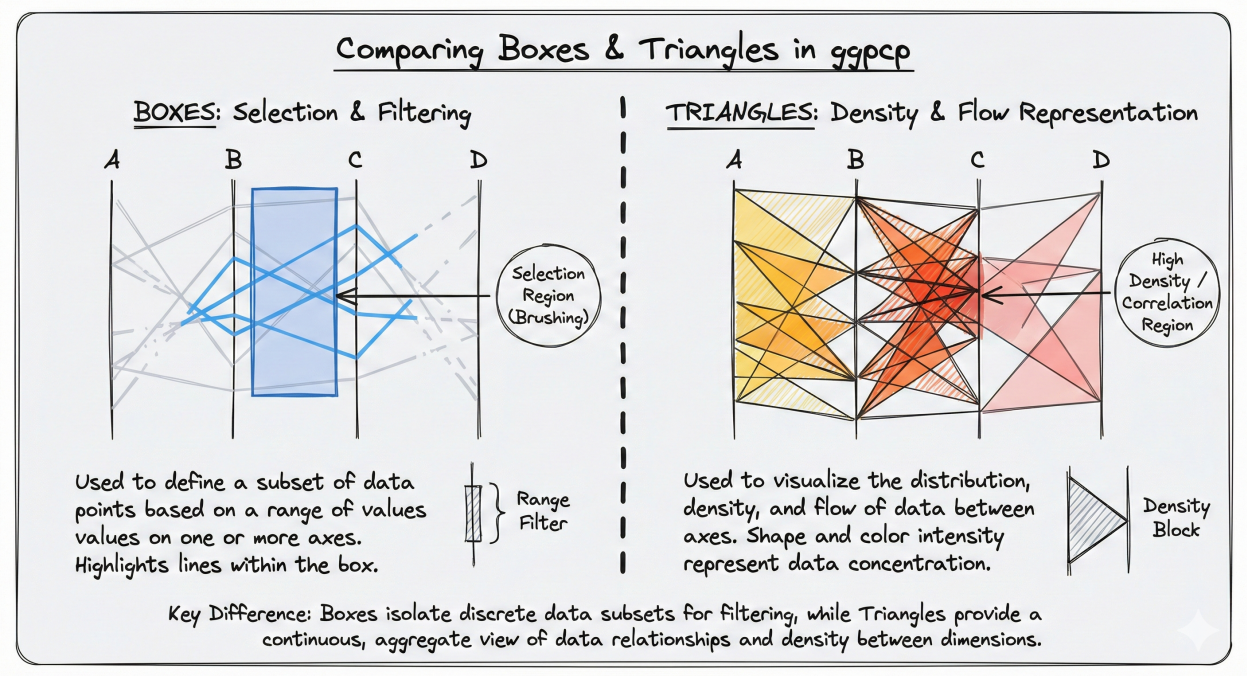


Figure 10: The visual proof of the perceptual trade-offs as describe regarding how data transitions between axes.

When ggpcp connects categorical and numerical variables, the tie-breaking algorithms keep the triangular patterns clear (Figure 10).

Schonlau and Yang (2024) observes that concentration patterns—such as the flow from zero hospitalizations to low comorbidity values—are "far more obvious in the hammock plot than in the GPCP plot" precisely because hammock plots maintain constant-width boxes that make marginal densities apparent. The triangular narrowing that occurs in GPCPs when categorical variables meet numerical ones obscures these density patterns and makes it difficult to distinguish individual

observations. The `ggpcp` package can address both limitations through intelligent tie resolution. The `space` parameter controls vertical spacing, while the `method` parameter determines systematic ordering of tied values. When implemented effectively, these parameters transform the vertical axis into something functionally equivalent to a rug plot, allowing users to visually estimate density while simultaneously preserving the ability to trace individual observation paths across the plot. This dual benefit—clear marginal density representation and resolvable individual lines—emerges from the same underlying solution.

This observation suggests an opportunity to incorporate conventions from both GPCPs and hammock plots to improve the utility of generalized parallel coordinate visualizations. Rather than treating these as competing approaches with distinct perceptual trade-offs, thoughtful tie-breaking implementation can combine the strengths of each method.

To understand how our tie-resolution methods achieve this synthesis, it is important to recognize the perceptual and practical challenges inherent in parallel coordinate plots. The following section draws on evidence from the visualization literature that directly informs our design choices and evaluation plan.

2 An Approach to Numerical Ties

Currently, numerical variables disrupt some of the best innovations in the GPCP approach by making it impossible to track an observation across the plot and by obscuring marginal density information through overplotting. The remaining two chapters of the dissertation will address the handling of ties on numerical axes and assess the utility of visual cues that can be paired with the tie breaking method to indicate that values are approximately accurate spatially and equivalent numerically.

The `ggpcp` package does not currently provide a mechanism for resolving these tied values. We propose `tie_spread`, an algorithm that distributes observations with identical values along the vertical axis, transforming overlapping lines into a visually resolvable spread. Consistent with tidyverse conventions and the tidyselect grammar, the implementation allows users to specify tie-breaking behavior selectively across axes—applying spreading to some variables while preserving exact positioning on others. Sensible defaults balance visual clarity against faithful representation of the underlying data, ensuring that users can immediately benefit from improved legibility without extensive parameter tuning. The following section details the design and implementation of this approach.

2.1 The Problem: Overlapping Lines

2.2 The Solution: Tie Spreading

The `tie_spread` algorithm splits up tied observations by evenly spreading them out over a small range around their original value. This range is set by default to balance the need for effective tie-breaking with spatial perception. This makes sure that separated observations stay visually grouped near their shared value while still being different enough to trace individual paths through the plot.

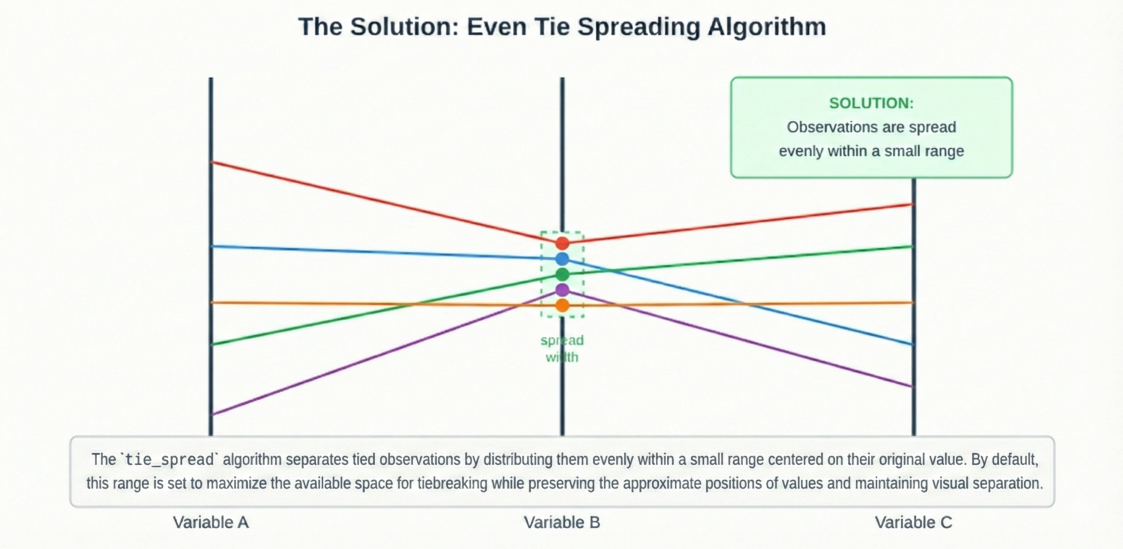
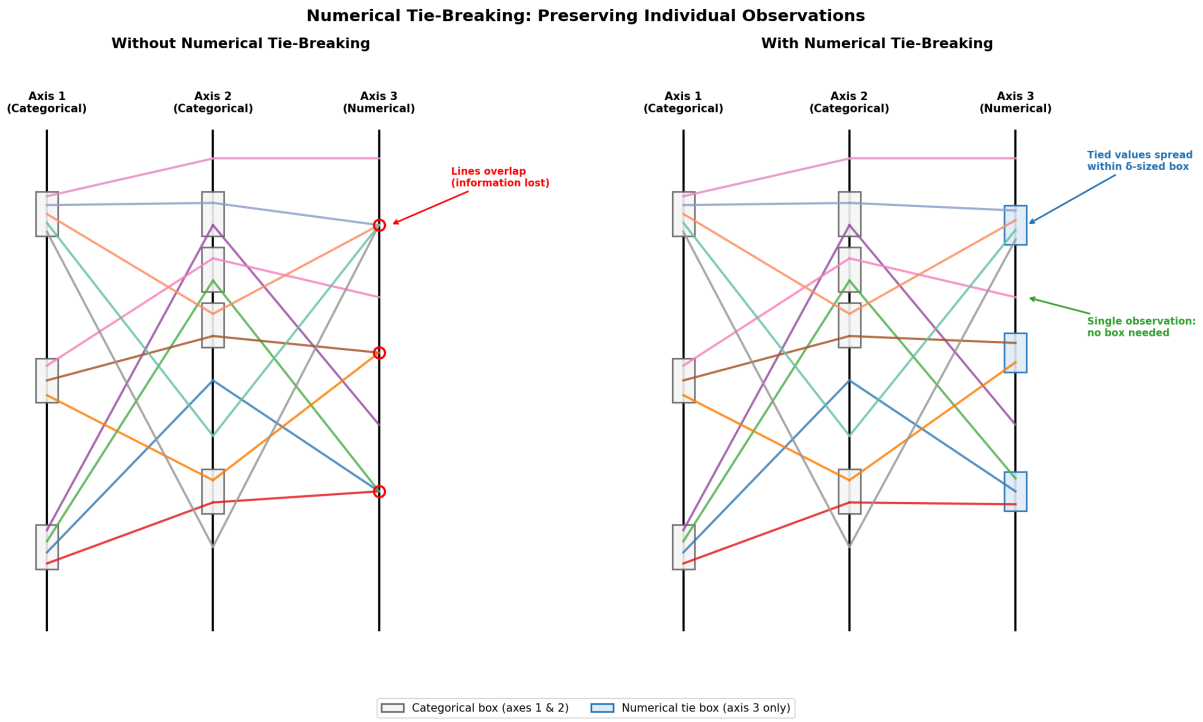


Figure 11: Tie spreading maintains visual separation while preserving approximate positions.

2.3 Hierarchical Sorting for Minimal Crossings

Beyond simply spreading tied values, the ggpcp package implements hierarchical sorting to minimize line crossings as in categorical tie breaking.

By ordering observations within a tie group based on their values on adjacent axes, the package leverages the Gestalt principle of common fate. This strategy ensures that observations with similar trajectories are positioned in close proximity, creating cohesive visual bands. These bands facilitate the perception of distinct groups as they move together through high-dimensional space, reducing visual noise and highlighting underlying patterns.



2.4 Gestalt Principles and Line Continuity

The principle of good continuation states that the human visual system preferentially perceives smooth, continuous contours over interpretations requiring abrupt changes in direction. This principle directly impacts the effectiveness of parallel coordinate plots for tracing individual observations.

Equispaced lines maintain continuity by representing each observation as a continuous polyline extending from the leftmost to the rightmost axis. When observations share identical numerical values (ties), the equispacing mechanism distributes lines within each category so that they remain visually distinct without introducing discontinuities. The ggpcp framework implements this through hierarchical sorting, which minimizes line crossings at axis intersections and leverages both the Gestalt principle of good continuation and the principle of common fate, wherein lines with similar values across multiple axes appear to move together through the display.

Constant-width boxes represent observations as aggregated area segments rather than individual lines. This encoding communicates aggregate quantities effectively: the area of each segment is proportional to the number of observations sharing that combination of values across adjacent axes, making the joint frequencies of category pairs immediately visible. In contrast, equispaced lines preserve individual observations as distinct visual elements. When lines are continuous and uninterrupted, viewers can leverage preattentive processes such as the Gestalt principle of good continuation to trace a single case across all axes (Healey and Enns 2012). The two approaches thus serve distinct purposes: lines support individual-level tracing through low-level perceptual grouping, while boxes support aggregate-level frequency comparison.

This distinction becomes increasingly relevant as dataset size grows, and it is important to distinguish between two different uses of box-like visual elements in parallel coordinate displays. Hammock plots use parallelograms to represent aggregated bivariate relationships: the area of each

parallelogram encodes the joint frequency of category pairs across adjacent axes, replacing individual observations with area segments. This approach scales well with large datasets because the visual representation remains stable regardless of sample size, but individual observations cannot be traced.

GPCPs take a different approach: when multiple observations share the same categorical value, their vertical positions are spread out within a frame to avoid overplotting. The frame signals that the precise positions within that region reflect tie-breaking rather than meaningful data variation, but the individual lines remain continuous and traceable across the full plot. With many observations, these equispaced lines may coalesce into ribbon-like bands; however, the principle of good continuation still applies, allowing viewers to perceive flow patterns even when individual lines are indistinguishable.

Rather than viewing these as competing representations, they can be understood as complementary: the GPCP approach is optimal when the analytical task involves following specific cases or detecting outliers, whereas the hammock approach is optimal when the task involves comparing category frequencies or understanding distributional patterns without regard to individual observations.

3 Mathematical Framework

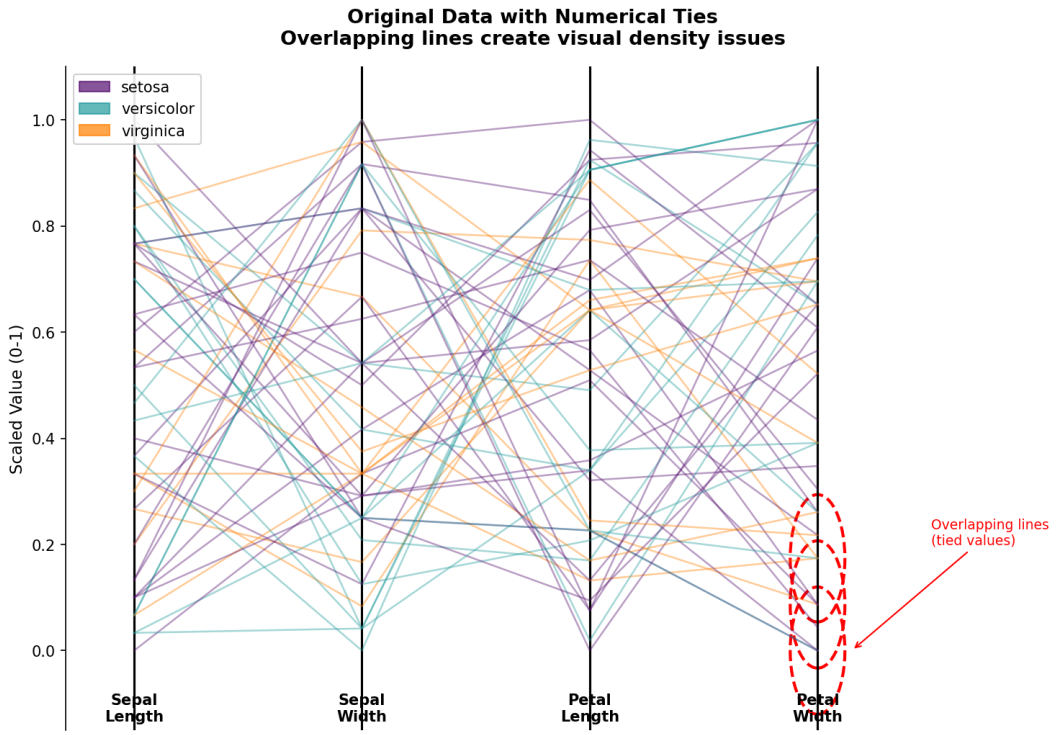


Figure 12: The tie problem in parallel coordinate plots. When observations share values, lines overlap and create visual collisions that mask the underlying data distribution.

3.1 Notation

Let $\mathcal{D}$ be a dataset with n observations and p variables displayed on parallel axes $X_1, X_2, \ldots, X_p$.

Table 1: Core notation

Symbol	Definition
x_{ij}	Raw value of observation i on variable j
y_{ij}	Scaled value (normalized to $[0, 1]$)
$\tilde{y}_{ij}$	Adjusted value after tie-breaking
$T_j(v)$	Set of observations tied at value v on axis j
k	Number of tied observations: $k = \lvert T_j(v)\rvert$
δ_j	Resolution of axis j: minimum distance between distinct values

3.2 Scaling

Variables are normalized to $[0, 1]$ using min-max scaling:

$$y_{ij} = \frac{x_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{1}$$

3.3 The Tie-Breaking Algorithm

3.3.1 Step 1: Compute Axis Resolution

To preserve density information and maintain consistent spacing, we first compute the global resolution of axis j—the smallest difference between any two distinct values:

$$\delta_j = \min\{|y_1 - y_2| : y_1 \neq y_2,\ y_1, y_2 \in Y_j\} \tag{2}$$

where Y_j is the set of unique values on axis j.

Using a consistent jitter range across all values ensures that the visual representation preserves both the density of observations at each value and the gaps between distinct values. An approach that computed jitter ranges individually for each tied value would risk collapsing visually meaningful gaps, destroying distance information in the display.

3.3.2 Step 2: Identify Ties

A value v on axis j is a tie if multiple observations share it:

$$T_j(v) = \{i : y_{ij} = v\}, \quad |T_j(v)| > 1 \tag{3}$$

3.3.3 Step 3: Compute Available Space

Tied observations at value v are spread within an interval defined by the axis resolution:

$$L(v) = v - \frac{\delta_j}{2}, \quad U(v) = v + \frac{\delta_j}{2} \tag{4}$$

If v is at the axis boundary, apply a buffer β (default 0.05):

- $L(v) = \max(0, v - \delta_j/2)$ if $v - \delta_j/2 < 0$
- $U(v) = \min(1, v + \delta_j/2)$ if $v + \delta_j/2 > 1$

The available space is:

$$S(v) = U(v) - L(v) \tag{5}$$

Step 4: Compute Optimal Spacing

The core principle is identical for categorical and numerical ties:

$$\boxed{\Delta = \frac{S(v)}{k-1}} \tag{6}$$

This distributes k observations evenly across the available space $S(v)$.

3.3.4 Step 5: Order Observations Hierarchically

To minimize line crossings, tied observations are sorted by their positions on adjacent axes. For axis j, let A_j be the adjacent axis (left or right depending on processing direction).

For each observation $i \in T_j(v)$, the sorting key is:

$$\kappa_i = \tilde{y}_{i,A_j} \tag{7}$$

Observations are ordered so that $\kappa_{i_1} \leq \kappa_{i_2} \leq \cdots \leq \kappa_{i_k}$.

Key insight: This requires *sequential* axis processing—the adjusted positions from previously processed axes inform the ordering on the current axis.

3.3.5 Step 6: Assign Positions

Given ordered observations $(i_1, i_2, \ldots, i_k)$:

$$\tilde{y}_{i_m,j} = L(v) + (m-1) \cdot \Delta, \quad m = 1, 2, \ldots, k \tag{8}$$

3.3.6 Position Functions

Two variants of position functions distribute observations within the available space:

The edge-inclusive version places observations at:

$$\text{positions} = L(v) + \frac{i-1}{k-1} \cdot S(v), \quad i = 1, ..., k$$

3.4 Unified Framework

The algorithm treats numerical ties identically to categorical levels:

Table 2: Unified spacing principle

Variable Type	Available Space	Spacing
Categorical	Box height h_ℓ for level ℓ	$\Delta = h_\ell/(n_\ell - 1)$
Numerical	Resolution-based interval $S(v) = \delta_j$	$\Delta = S(v)/(k-1)$

Both apply the same formula: **spacing = available_space / (n − 1)**.

3.5 Visual Results

3.5.1 Uniform Spacing Applied

After applying the tie-breaking algorithm, observations that previously overlapped are now visually distinguishable:

3.5.1.1 Annotated Tie Regions

Tie regions can be annotated with boxes showing the original value and the spread region:

3.5.2 Tie Box Detection

The `get_tie_box()` function identifies and bounds tie regions for visualization:

3.6 Algorithm Summary

Algorithm: Unified Tie-Breaking

Input: Data D, variables $X_1 ... X_p$, method $\in$ {from-left, from-right, from-both} Output: Adjusted positions $_{ij}$

1. Scale all variables to $[0, 1]$
2. For each axis j, compute resolution: $\delta_i = min|y_1 - y_2|$ for distinct $y_1, y_2 \in Y_j$
3. Set processing order based on method:

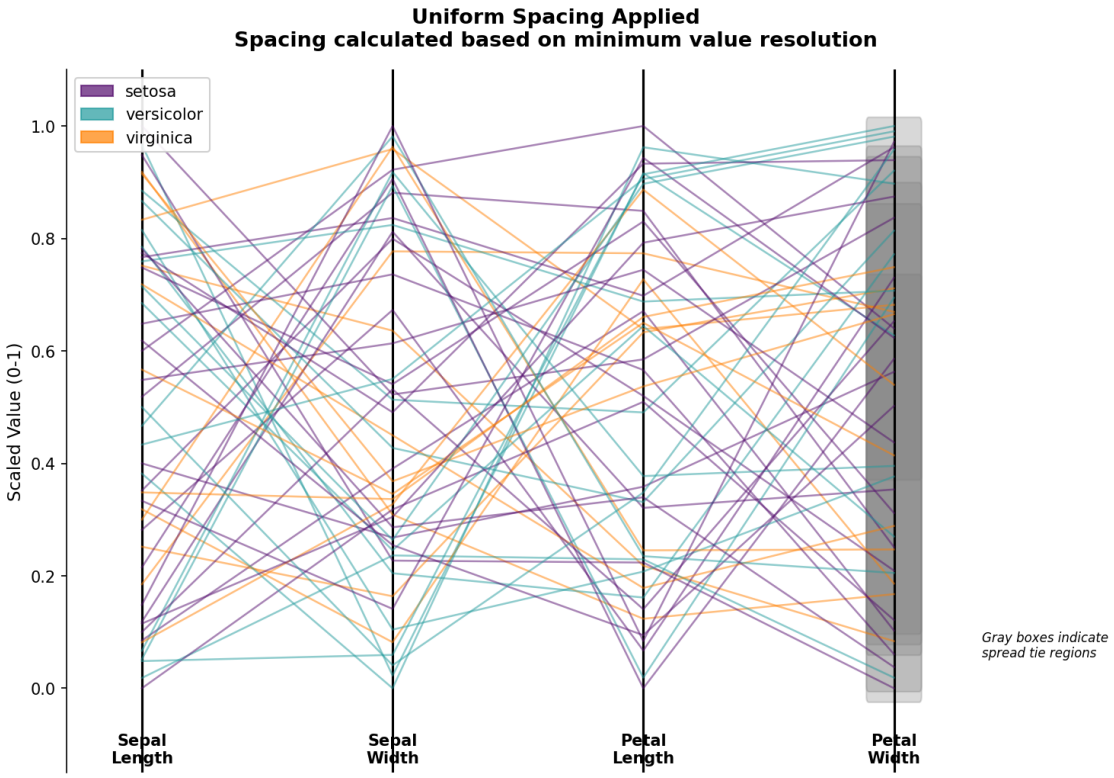


Figure 13: Parallel coordinate plot after applying uniform spacing. Gray boxes indicate regions where tied values have been spread, making individual observations traceable.

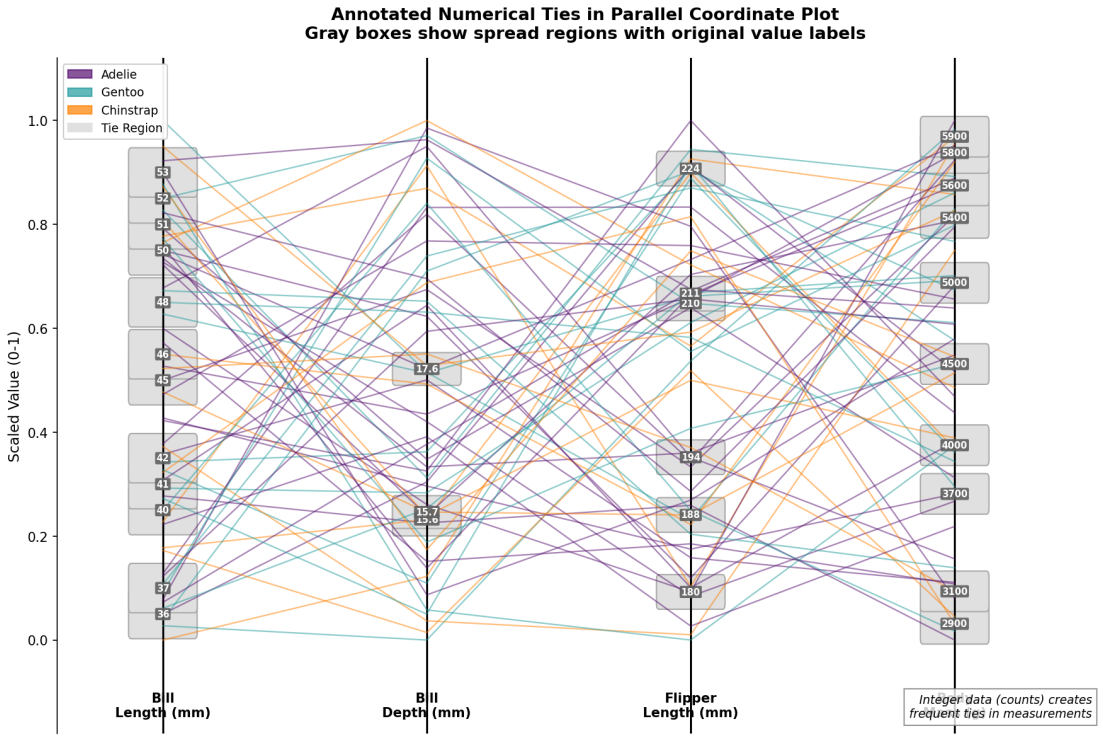


Figure 14: Annotated tie regions with labels. Gray boxes highlight where ties occurred, with labels showing the original tied value. This is particularly useful for integer data where ties are frequent.

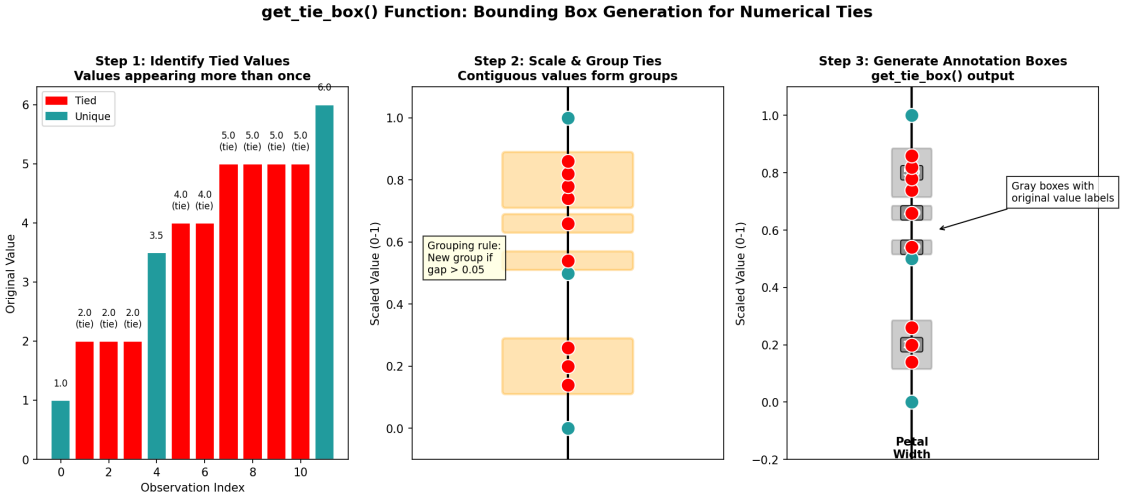


Figure 15: The tie box detection process. (1) Identify duplicated values, (2) scale and group contiguous spreads, (3) generate bounding box coordinates for annotation.

- from-left: process axes $1, 2, \ldots, p$; reference left neighbor
- from-right: process axes $p, p-1, \ldots, 1$; reference right neighbor
- from-both: process outward from center axis; reference previously processed neighbor

4. For each axis j in processing order:
    a. Identify tied values: $V = v : |T_j(v)| > 1$
    b. For each $v \in V$:
        i. Set available space: $S(v) = \delta_j$
        ii. Compute bounds: $L(v) = v - \frac{\delta_i}{2} 2, U(v) = v + \frac{\delta_i}{2}$
        iii. Order $T_i(v)$ by positions on adjacent axis
        iv. Compute spacing: $\Delta = \frac{S(v)}{(k-1)}$
        v. Assign positions: $_{i_m,j} = L(v) + (m-1) * \Delta \; for \; m = 1, ..., k$
5. Return adjusted data

3.7 Parameters

Table 3: Algorithm parameters

Parameter	Symbol	Default	Description
Boundary buffer	β	0.05	Fallback when no neighbor exists
Minimum spacing	δ	0.01	Ensures minimum separation
Tie tolerance	ε	10^{-10}	For floating-point comparison

4 Implementation and Evaluation Plan

4.1 Phase 1: Algorithm Development (Weeks 1-4)

Deliverable: Functional R package extension with comprehensive documentation Tasks:

1. Extend `pcp_arrange()` to detect and handle numerical ties
2. Implement `optimize_spacing_numerical()` or `tie_spacing()` function with perceptually-motivated parameters
3. Handle edge cases while preserving perceptual properties:

- Single unique value (centered position)
- Extreme skew (adaptive buffer sizing)
- Missing values (explicit separation region)

4. Integration testing with existing ggpcp workflow

Code Structure (Simplified):

4.2 Phase 2: Perceptual Validation Studies (Weeks 5–7)

Study 1: Frequency Perception

- Duration: Weeks 5–7
- Participants: n = 128 (64 per condition)
- Design: Between-subjects comparison of GPCP vs. hammock plot representations

Sample Size Justification:

A power analysis was conducted using G*Power 3.1 (Faul et al. 2007) to determine the minimum sample size required to detect a meaningful difference between visualization conditions. Based on prior graphical perception studies—including Heer and Bostock's (2010) crowdsourced replication of Cleveland and McGill (1984), which used approximately 50 participants per condition, and Hofmann and Vendettuoli's (2013) common angle plot study with 46 participants, we anticipated a medium effect size (Cohen's d = 0.5). For a two-tailed independent samples t-test with = 0.05 and power = 0.80, the required sample size is n = 64 per group (total N = 128). This sample size also provides 80% power to detect effects as small as d = 0.50, which represents a practically meaningful difference in task accuracy between visualization types (Brysbaert 2019). To account for potential exclusions due to failed attention checks or incomplete responses (estimated at 10–15% based on crowdsourcing norms), we will recruit n = 75 per condition.

Procedure:

1. Magnitude estimation (12 trials): "What percentage have value X?"
2. Ordinal comparison (12 trials): "Which value is more frequent?"
3. Ratio judgment (12 trials): "A is how many times B?"
4. Confidence ratings (7-point scale) for each response

Analysis:

- Absolute percentage error for magnitude estimation
- Accuracy rates for ordinal and ratio tasks
- Bias analysis: systematic over/under-estimation
- Confidence calibration: accuracy vs. subjective confidence
- Effect sizes reported as Cohen's d with 95% confidence intervals

4.3 Phase 3: Comparative Benchmarking (Weeks 6–10)

Computational Metrics:

1. Rendering performance (time, memory, scalability n = 2 to 100)
2. Visual quality metrics (Dennig et al. 2021): line crossing count, ribbon overlap, visual clutter index
3. Perceptual quality estimates: modeled eye movements, predicted visual search time

Case Studies:

- Palmer Penguins: Mixed categorical-numerical data with known correlation structure
- Iris Dataset: Multiple numerical variables with natural ties
- Asthma Data (Schonlau and Yang 2024): Direct comparison with published hammock plot

4.4 Phase 4: Integration and Dissemination (Weeks 8–12)

Deliverables:

1. R Package Update:

- CRAN submission with full documentation
- Vignette: "Handling Numerical Ties in Parallel Coordinate Plots"
- Unit tests achieving $> 95\%$ coverage

2. Academic Paper:

- Target: IEEE Transactions on Visualization and Computer Graphics
- Submission deadline: May 2026 (IEEE VIS)

3. Supplementary Materials:

- Open Science Framework repository with pre-registration
- All experimental stimuli and data
- Reproducibility package with power analysis scripts

5 Limitations and Future Directions

5.1 Study Limitations

- **Sample Characteristics:** University student population may not generalize to domain experts
- **Experimental Constraints:** Laboratory tasks may lack ecological validity
- **Scope:** Limited to static visualizations; interactive features not evaluated

5.2 Future Extensions

- Longitudinal study with domain experts in real analytical workflows
- Investigation of interaction effects between individual differences and visualization method
- Hybrid approaches: smooth interpolation between lines and boxes based on dataset properties
- Integration with animation and interactive highlighting techniques

6 Conclusion

This proposal addresses a fundamental challenge in information visualization: grounding design decisions in perceptual science rather than convention or intuition. The proposed equispaced line method extends ggpcp's established categorical tie-breaking algorithm to numerical variables, creating a unified framework that preserves visual continuity and leverages preattentive processing through the Gestalt principles of good continuation and common fate.

It is important to distinguish between two different uses of box-like visual elements in parallel coordinate displays. Hammock plots use parallelograms to represent aggregated bivariate relationships, where the area of each segment encodes the joint frequency of category pairs across adjacent axes. This approach scales well with large datasets but does not permit tracing of individual observations. GPCPs take a different approach: when multiple observations share the same value, their vertical positions are spread within a frame defined by the global resolution of the axis—the smallest difference between any two distinct values. The frame signals that positions within it reflect tie-breaking rather than meaningful data variation, but individual lines remain continuous and traceable. Using a consistent jitter range across all values ensures that the visual representation preserves both the density of observations at each value and the gaps between distinct values; an individualized approach would risk destroying this distance information.

Rather than viewing these as competing representations, they can be understood as complementary. The GPCP approach is optimal when the analytical task involves following specific cases or detecting outliers, whereas the hammock approach is optimal when the task involves comparing category frequencies or understanding distributional patterns without regard to individual observations. With many observations, equispaced lines may coalesce into ribbon-like bands; however, the principle of good continuation still applies, allowing viewers to perceive flow patterns even when individual lines are indistinguishable.

This research contributes to theory by demonstrating how Gestalt principles and models of preattentive processing generate testable hypotheses about visualization effectiveness. It contributes to practice by providing open-source software and evidence-based design guidelines. And it contributes to methodology by illustrating theory-driven, multi-method evaluation—computational benchmarking to verify visual quality before perceptual testing, followed by appropriately powered user studies grounded in the graphical perception literature.

Effective visualization must be rooted in how people actually perceive: how they use principles of proximity, similarity, continuity, and common fate to organize and interpret visual features. This study takes that principle seriously by asking not merely "which visualization looks better?" but "which visualization aligns more effectively with perceptual organization principles for specific analytical tasks?" The answer, as with most questions in visualization, will depend on context. But it will now rest on empirical evidence and perceptual theory rather than assumption.

7 Appendix

7.1 Visual Clutter and Information Density (To investigate)

Visual clutter constitutes a fundamental limitation on visualization effectiveness. Clutter occurs when too many visual elements compete for attention, overwhelming the viewer's capacity to extract meaningful patterns from the display.

The two approaches manage clutter through fundamentally different mechanisms. For equispaced lines, clutter increases with the number of observations, but hierarchical sorting minimizes line crossings and thereby reduces visual complexity. The approach exhibits graceful degradation: as density increases, individual lines transition perceptually into ribbons and eventually into filled areas, yet individual observations remain theoretically accessible for highlighting or interactive selection. For constant-width boxes, clutter depends primarily on the number of unique value combinations rather than on raw observation counts. Aggregation inherently reduces clutter, though at the cost of individual-level accessibility. The resulting display provides a clear representation of bivariate contingency relationships.

Dennig et al. (2021) formalize clutter metrics for parallel sets visualizations, measuring ribbon overlap, crossing angles, and ribbon width variance. These metrics can be adapted to compare the two approaches quantitatively and to guide the selection of dimension and category orderings that minimize visual complexity.

References

Brysbaert, Marc. 2019. "How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables." *Journal of Cognition* 2 (1): 16. https://doi.org/10.5334/joc.72.

Cleveland, William S, and Robert McGill. 1984. "Graphical Methods for Data Presentation: Full Scale Breaks, Dot Charts, and Multibased Logging." *The American Statistician* 38 (4): 270–80. https://doi.org/10.1080/00031305.1984.10483223.

Dennig, Frederik L., Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. 2021. "ParSetgnostics: Quality Metrics for Parallel Sets." *Computer Graphics Forum* 40 (3): 375–86. https://doi.org/10.1111/cgf.14314.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. https://doi.org/10.3758/BF03193146.

Healey, Christopher G., and James T. Enns. 2012. "Attention and Visual Memory in Visualization and Computer Graphics." *IEEE Transactions on Visualization and Computer Graphics* 18 (7): 1170–88. https://doi.org/10.1109/TVCG.2011.127.

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. https://doi.org/10.1145/1753326.1753357.

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.

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

———. 2009. "Parallel Coordinates: Visual Multidimensional Geometry and Its Applications." *Springer Science & Business Media*.

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. https://doi.org/10.1109/TVCG.2006.76.

Pilhöfer, Alexander, and Antony Unwin. 2013. "New Approaches in Visualization of Categorical Data: R Package `extracat`." *Journal of Statistical Software* 53 (i07): 1–25. https://doi.org/

10.18637/jss.v053.i07.

Schonlau, Matthias. 2003. "The Hammock Plot: Visualizing Mixed Categorical and Numerical Data." In *Proceedings of the American Statistical Association*.

Schonlau, Matthias, and Rosie Yuyan Yang. 2024. "Hammock Plots: Visualizing Categorical Data Beyond Parallel Coordinates." *Journal of Computational and Graphical Statistics*.

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): 1405–20. https://doi.org/10.1080/10618600.2023.2181762.

Wegman, Edward J. 1990. "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association* 85: 664–75.