

Visualizing Ambiguity: A Grammar of Graphics Approach to Resolving Numerical Ties in Parallel Coordinate Plots

Denise Bradford

2026-01-01

Table of contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 3 |
| 2 | Background and Motivation | 3 |
| 2.1 | Parallel Coordinate Plots | 3 |
| 2.2 | Numerical Ties and Visual Overlap | 3 |
| 2.3 | Existing Solutions for Categorical Ties in ggpcp | 4 |
| 2.4 | The Challenge of Visualizing Mixed-Type Data | 4 |
| 2.4.1 | Step 1: Identify the Parallel Axes | 4 |
| 2.4.2 | Step 2: Locate the Starting Point | 4 |
| 2.4.3 | Step 3: Follow the Line Segment | 6 |
| 2.4.4 | Step 4: Read Values at Intersections | 6 |
| 2.4.5 | Step 5: Continue Across All Axes | 6 |
| 2.4.6 | Current State: ggpcp's Categorical Tie-Breaking | 6 |
| 2.4.7 | Alternative Approach: Hammock Plots | 8 |
| 2.4.8 | Visual Comparison: Triangles vs. Boxes | 9 |
| 2.4.9 | The Line Width Illusion and Perceptual Constraints | 9 |
| 3 | Handling Numerical Ties | 10 |
| 3.1 | The Problem: Overlapping Lines | 11 |
| 3.2 | The Solution: Tie Spreading | 11 |
| 3.3 | Hierarchical Sorting for Minimal Crossings | 11 |
| 3.4 | Theoretical Framework: Perception-Driven Design | 13 |
| 3.4.1 | Stage 1: Pattern Perception and Gestalt Principles | 13 |

| | | |
|----------|---|-----------|
| 3.4.2 | Stage 2: Visual Working Memory and Attention | 14 |
| 3.5 | Preattentive Processing and Visual Search | 14 |
| 3.5.1 | External Cognition and Computational Offloading | 15 |
| 3.5.2 | Gestalt Principles and Line Continuity | 15 |
| 3.5.3 | The Challenge of Area Perception | 16 |
| 3.5.4 | Visual Clutter and Information Density | 16 |
| 3.6 | Problem Statement and Research Question | 17 |
| 3.6.1 | Main Research Question | 17 |
| 3.6.2 | The Specific Conflict: Numerical Ties | 17 |
| 3.6.3 | The Limitations of Current Solutions | 17 |
| 3.6.4 | Gap in Research | 17 |
| 3.6.5 | Suggested Fix | 18 |
| 3.6.6 | Hierarchical Sorting for Line Crossing Minimization | 18 |
| 3.7 | Implementation and Evaluation Plan | 19 |
| 3.7.1 | Phase 1: Algorithm Development (Weeks 1-4) | 19 |
| 3.8 | Phase 2: Perceptual Validation Studies (Weeks 5-7) | 19 |
| 3.9 | Phase 3: Comparative Benchmarking (Weeks 6-10) | 20 |
| 3.10 | Phase 4: Integration and Dissemination (Weeks 8-12) | 20 |
| 4 | Timeline and Milestones | 20 |
| 5 | Expected Contributions | 20 |
| 5.1 | Theoretical Contributions | 20 |
| 5.2 | Practical Contributions | 21 |
| 5.3 | Methodological Contributions | 21 |
| 6 | Broader Impact | 21 |
| 7 | Limitations and Future Directions | 22 |
| 7.1 | Study Limitations | 22 |
| 7.2 | Future Extensions | 22 |
| 8 | Conclusion | 22 |
| | References | 22 |

1 Introduction

This proposal outlines a systematic approach to visually distinguish tied numerical values in multidimensional datasets by employing parallel coordinate plots (PCPs). Parallel coordinates, first popularized by Alfred Inselberg, are a powerful technique for investigating patterns across multiple attributes simultaneously (Inselberg 2009). However, when datasets contain exact numerical ties, the resulting overlapping lines in PCPs can obscure critical distinctions. *how about "remove useful info."?* *ide what this means*

To address this, we propose a uniform method for collecting spacing to tied values. The method will be integrated into the `ggpcp` package in R, ensuring a streamlined workflow for users seeking enhanced clarity in their parallel coordinate visualizations. *is this AI, or are you trying to sound fancy?*

Importantly, our approach complements recent work on generalized parallel coordinate plots (GPCPs), an extension of PCPs that supports categorical variables (VanderPlas et al. 2023). The `ggpcp` package for R implements these GPCPs using a grammar of graphics framework, which seamlessly incorporates both continuous and categorical variables in a single parallel coordinate plot. One of the key contributions of that work is a robust tie-breaking mechanism for categorical variables, implemented through the `pcp_arrange()` function with methods including “from left” and “from right” hierarchical sorting. This ensures that individual observations can be traced across multiple dimensions, even when categories induce identical or “tied” values.

By adding multiple numerical tie-breaking techniques for continuous data—including our three deterministic approaches—we further refine GPCPs’ capacity to handle the visualization of real-world datasets exhibiting many types of ties.

*a similar
ties in numerical and categorical data.
enhance*

2 Background and Motivation

2.1 Parallel Coordinate Plots

Parallel coordinate plots assign each dimension of an n -dimensional dataset to a vertical axis arranged in parallel (Wegman 1990). Each observation is drawn as a polyline connecting its values on these axes, providing a visual representation that can illuminate underlying data structures.

2.2 Numerical Ties and Visual Overlap

When multiple observations share the same value in a given dimension, their polylines perfectly overlap, creating “visual collisions.” This masks information about distribution, density, or potential outliers. The treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg (1985) and Wegman (1990). However, the `ggpcp` implementation has demonstrated that careful tie-handling is essential for both continuous and categorical variables.

Introducing a small offset (“jitter”) to these tied values can mitigate overlap without distorting the overall relationships in the data (Peng, Ward, and Rundensteiner 2004). In the context of generalized parallel coordinate plots, the `ggpcp` package separates data management from visual rendering into three distinct components: variable selection and reshaping, scaling of axes, and treatment of ties in categorical axes (VanderPlas et al. 2023).

by implementing multiple stages of data transformation:

selection (and implicit reshaping)

Scaling, and sorting - a step necessary because of ggpcp's treatment of categorical variables.

2.3 Existing Solutions for Categorical Ties in ggpcl

The ggpcl package currently addresses categorical ties through sophisticated tie-breaking algorithms. The package implements hierarchical sorting through the `pcp_arrange(data, method, space)` function, with two primary methods: "from-left" and "from-right", meaning that tie breaks are determined hierarchically by variables' values from the specified direction. The parameter `space` specifies the amount of the y-axis to use for spacing between levels of categorical variables the axis used for spacing.

This hierarchical sorting approach serves as "external cognition," the additional computational processing reduces the cognitive load required to untangle overlapping lines in the parallel coordinate plot. The categorical tie-breaking creates equispaced tie-breaking that reduces line crossings and allows users to follow individual observations from left to right through the plot even for categorical variables.

This TP should probably go to 2.2?

2.4 The Challenge of Visualizing Mixed-Type Data

Parallel coordinate plots (PCPs) have been established as valuable tools for exploratory data analysis of high-dimensional numerical data since their introduction (Inselberg 1985; Wegman 1990). However, the use of PCPs is fundamentally limited when working with categorical variables or mixed categorical-continuous data. As VanderPlas et al. (2023) note in their introduction to generalized parallel coordinate plots (GPCPs), existing solutions for categorical values become insufficient when attempting to maintain visual continuity across both data types.

The treatment of ties, multiple observations sharing the same value, represents a critical design decision that affects both perceptual effectiveness and analytical utility. As VanderPlas et al. (2023) observe, "The treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg (1985) and Wegman (1990). We have found a need to deal with ties" (p. 6). This observation extends naturally from categorical to numerical variables.

The ability to follow individual observations is central to the analytical power of PCPs, enabling users to identify patterns, outliers, and relationships that span multiple variables simultaneously.

make this a simple list.

2.4.1 Step 1: Identify the Parallel Axes

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.

2.4.2 Step 2: Locate the Starting Point

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.

diagram of when/ how this happens w/ code from pkg
focus less on this more on the general approach
How does it depend on the grammar?
not w/ ggpcl?

You've already quoted this

Step 1: Identify the Parallel Axes

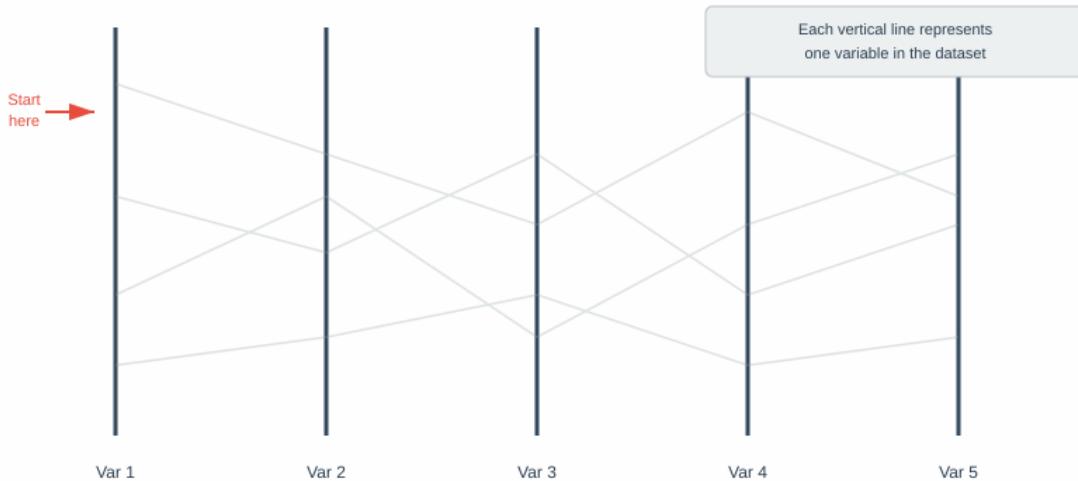


Figure 1: The parallel axes form the structural framework of the visualization.

Step 2: Locate the Starting Point on the First Axis

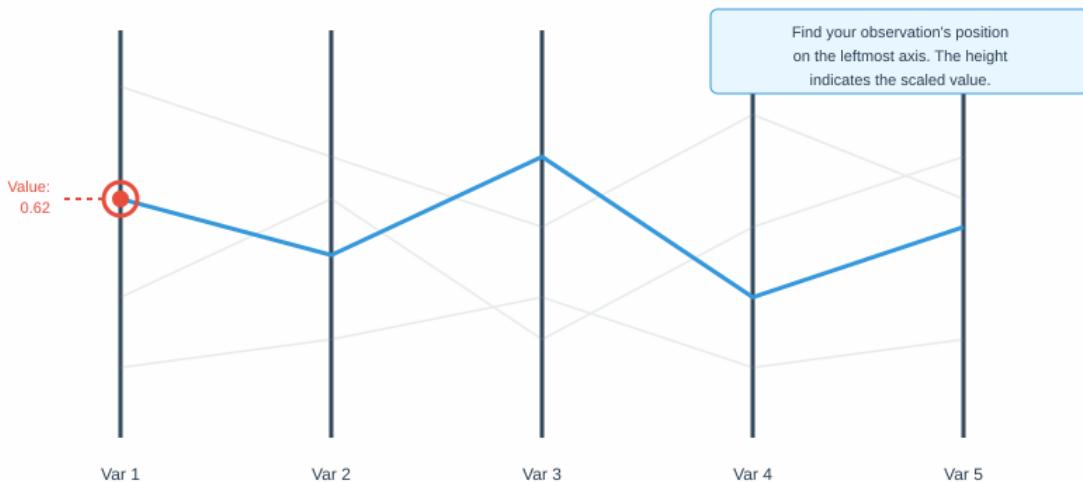


Figure 2: The starting point is identified on the first axis with its corresponding value.

2.4.3 Step 3: Follow the Line Segment

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. This principle allows viewers to perceive connected lines as unified objects, making it easier to track observations across multiple variables.

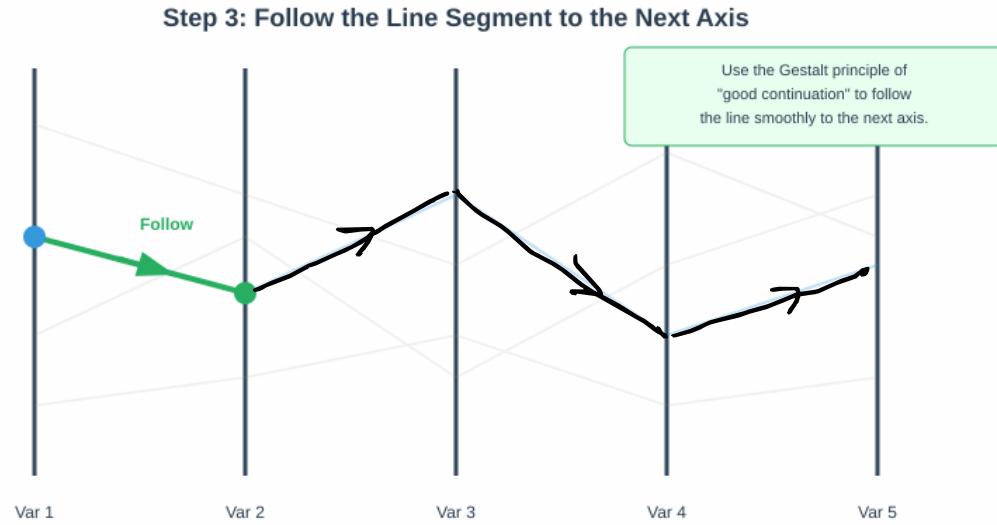


Figure 3: Following the line segment uses the Gestalt principle of good continuation.

2.4.4 Step 4: Read Values at Intersections

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 across different dimensions of the data. The slope of line segments between axes provides information about the relationship between consecutive variables for that specific observation.

2.4.5 Step 5: Continue Across All Axes

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.

2.4.6 Current State: ggpcp's Categorical Tie-Breaking

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:

Step 4: Read the Value at Each Intersection

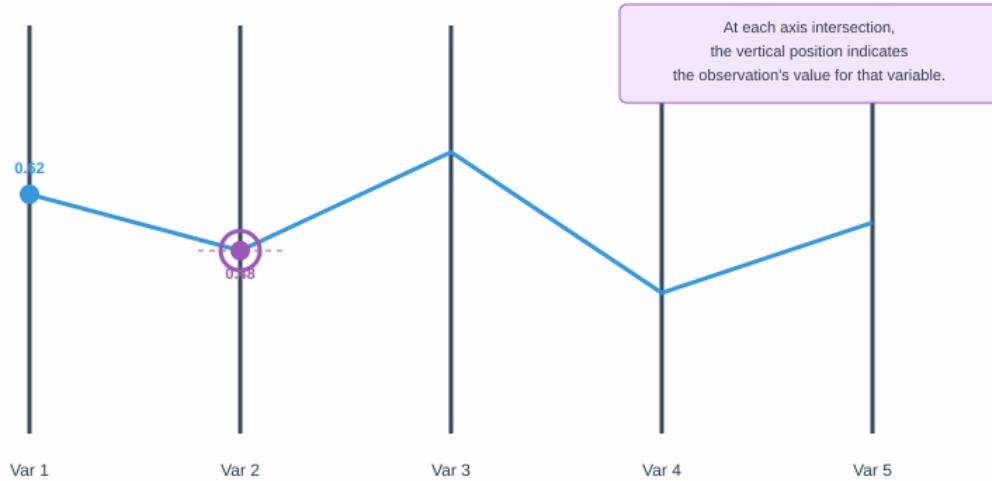


Figure 4: Values are read as approximate scaled value relative to the total range.

Step 5: Continue Across All Axes

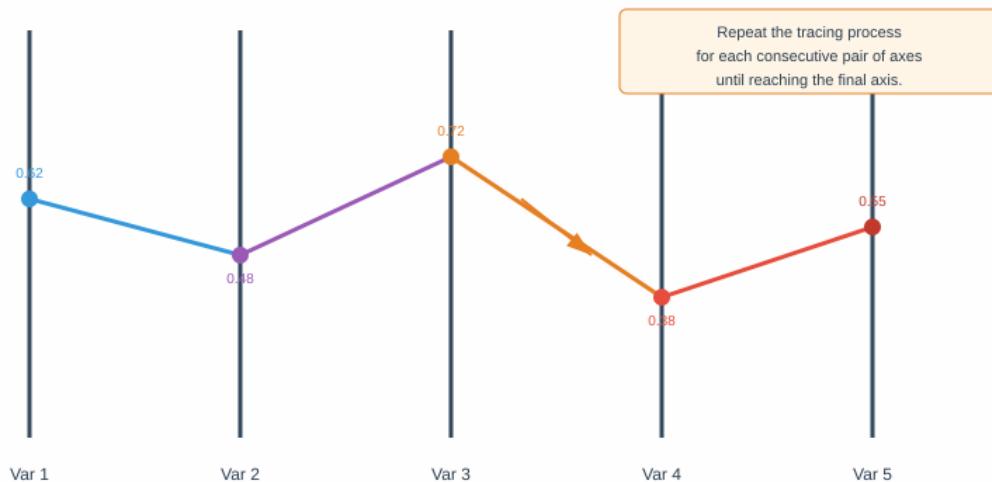


Figure 5: The complete traced path reveals the observation's values across all variables.

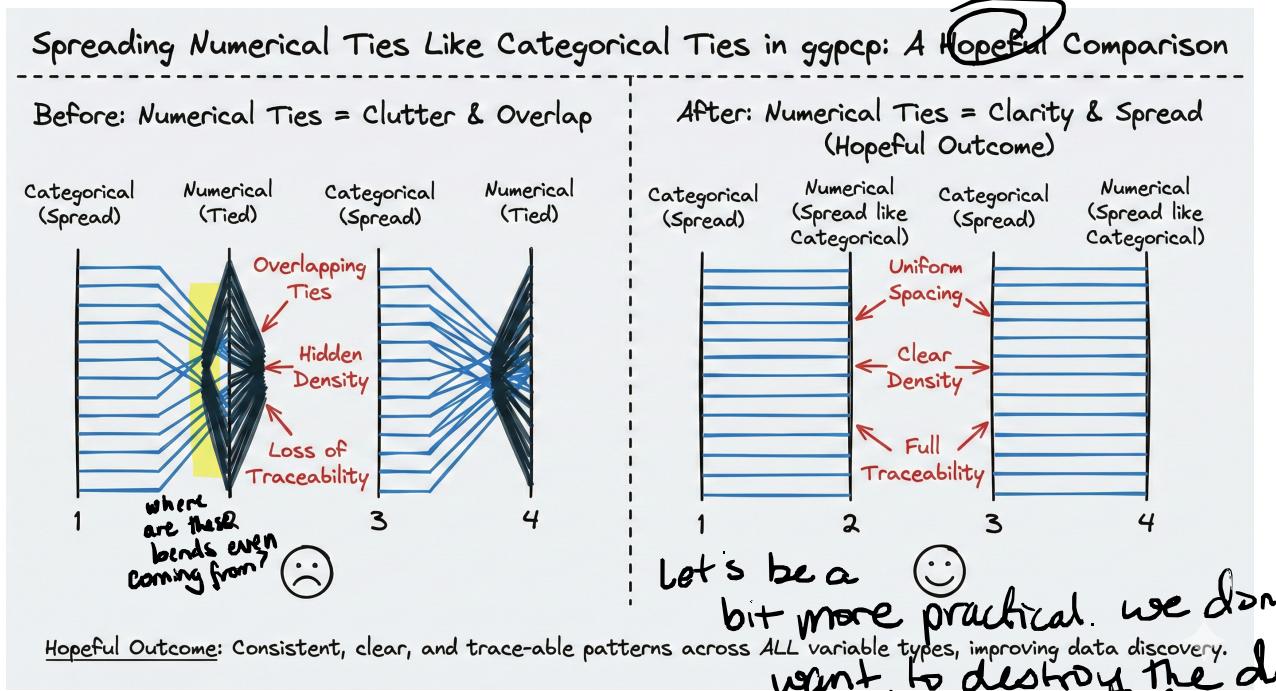


Figure 6: Comparison of standard parallel coordinate plot representation and the Even Tie Spreading Algorithm for resolving numerical ties.

“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 (default 5%). This optimization can be formalized as:

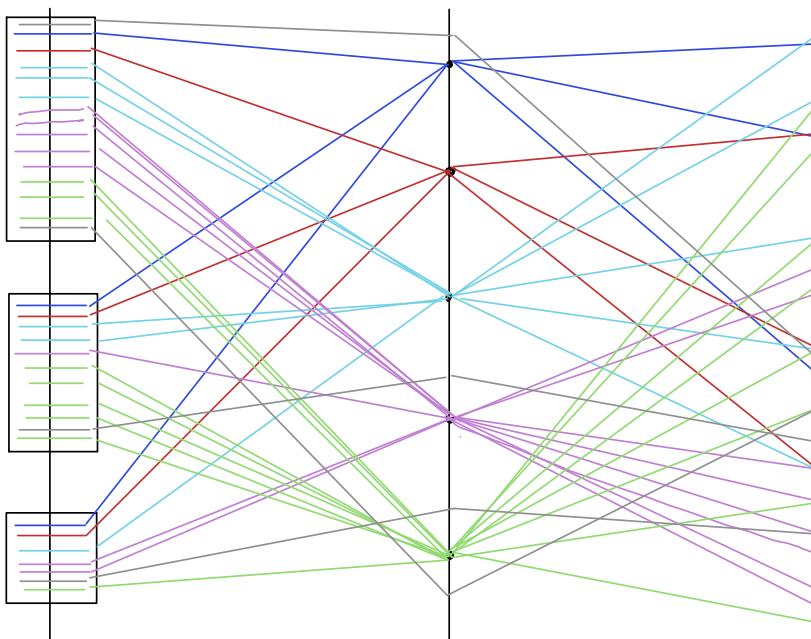
$$d_i = \frac{S_i - S_i^- - S_i^+}{n_i - 1}$$

where:

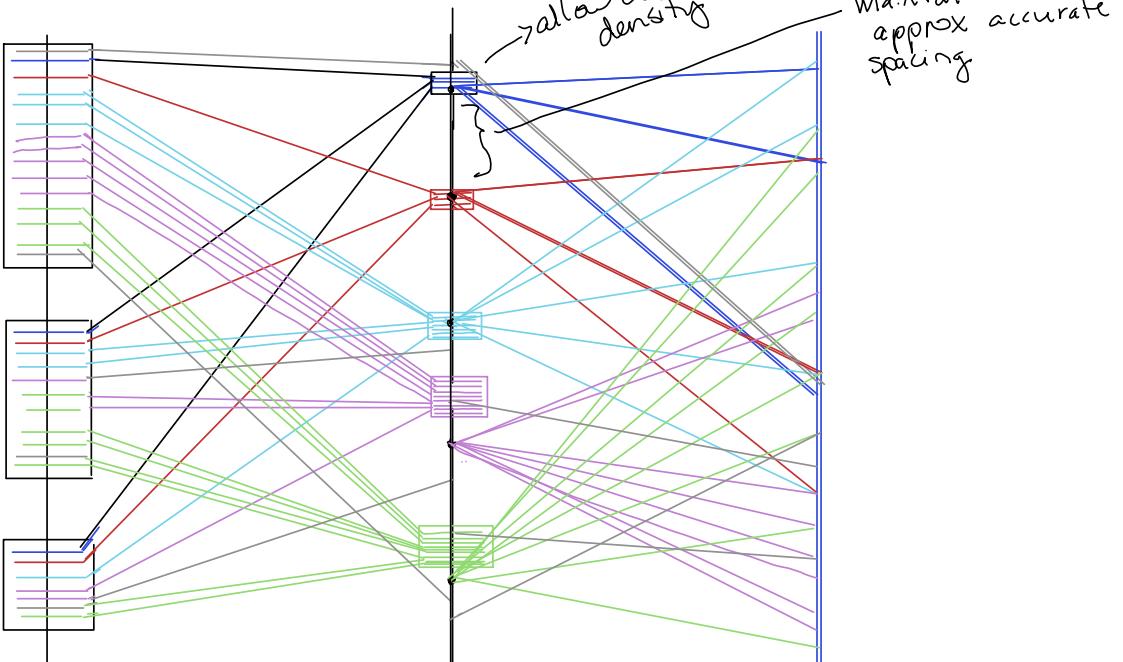
- S_i is the total space allocated to category i
- S_i^- is the spacing below category i
- S_i^+ is the spacing above category i
- n_i is the number of observations in category i
- d_i is the optimal spacing distance between consecutive observations

2.4.7 Alternative Approach: Hammock Plots

Hammock plots, introduced by Schonlau Schonlau and Yang (2024), take a fundamentally different approach to handling both categorical and numerical variables. Rather than using individual



Categorical



Categorical



Spatial distortion
is bad b/c I can't draw

lines, hammock plots employ two-dimensional boxes to connect adjacent axes, with box width proportional to the number of observations.

As Schonlau and Yang (2024) describes:

“Like a parallel coordinate plot, the axes are aligned parallel to one another. Categories of adjacent variables are connected by boxes. (The boxes shown are parallelograms; I use the word boxes for simplicity). The width of boxes is proportional to the number of observations.” (p. 3)

For numerical variables specifically, hammock plots maintain constant-width boxes throughout the visualization. Schonlau and Yang (2024) explains the spatial constraint this imposes:

“When treating this variable as numerical, the range from 0 to 20 leaves 1/21th of the space for each unit length. Consequently, the widths of the boxes have to be more frugal.” (p. 21)

This creates a fundamental trade-off: hammock plots explicitly encode frequency through box width but sacrifice individual observation traceability. As Schonlau and Yang (2024) notes, “For small data sets, GPCP plots beautifully show all individual observations whereas hammock plots require highlighting to feature individual observations” (p. 19).

Now you talk about the fact that maybe we don't have to sacrifice anything.

2.4.8 Visual Comparison: Triangles vs. Boxes

A key visual difference emerges when connecting categorical to numerical variables. VanderPlas et al. (2023) 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)

Show this if it is important enough to mention

This observation suggests that each approach has perceptual advantages in different contexts, motivating rigorous comparative evaluation grounded in perceptual science.

make sure to reference every figure in the relevant text.

Understanding the perceptual and practical challenges of parallel coordinate plots provides essential context for our tie-resolution methods. This section synthesizes empirical evidence from the visualization literature that directly informs our design decisions and evaluation strategy.

2.4.9 The Line Width Illusion and Perceptual Constraints

Both approaches must contend with perceptual illusions that affect visual interpretation. Schonlau and Yang (2024) discusses the line width illusion:

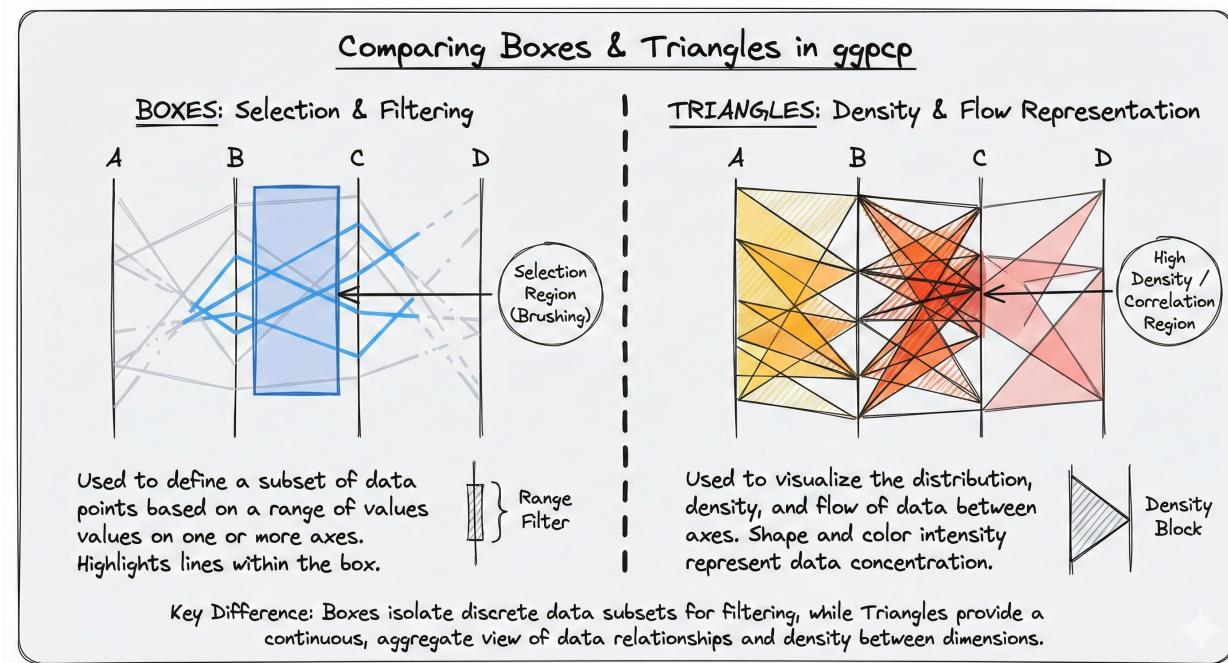


Figure 7: The visual proof of the perceptual trade-offs they describe regarding how data transitions between axes.

Are you doing anything with illusions? if not, omit this,

“The distance between two parallel lines is perceived at a right angle rather than as the vertical distance between the lines (Wallgren et al., 1996; Tufte, 2001)... The line width illusion is part of the family of Müller-Lyer illusions where two lines of same length appear to be of different lengths.” (p. 8)

Hofmann and Vendettuoli (2013) further identify the “reverse line width illusion,” where centering of lines creates a contextual cue encouraging evaluation of line widths using vertical rather than orthogonal measures. As they demonstrate empirically, this illusion can lead to systematic biases in frequency estimation.

The equispaced line approach proposed here may inherently avoid these illusions by maintaining parallel lines at consistent vertical spacing rather than using area-based encoding, though this requires empirical validation.

3 Handling Numerical Ties

A significant challenge arises when multiple observations share identical values on an axis. In traditional PCPs, these observations converge to a single point, creating overlapping lines that make individual tracking impossible. The ggpcp package addresses this through the tie_spread algorithm.

fig. 1

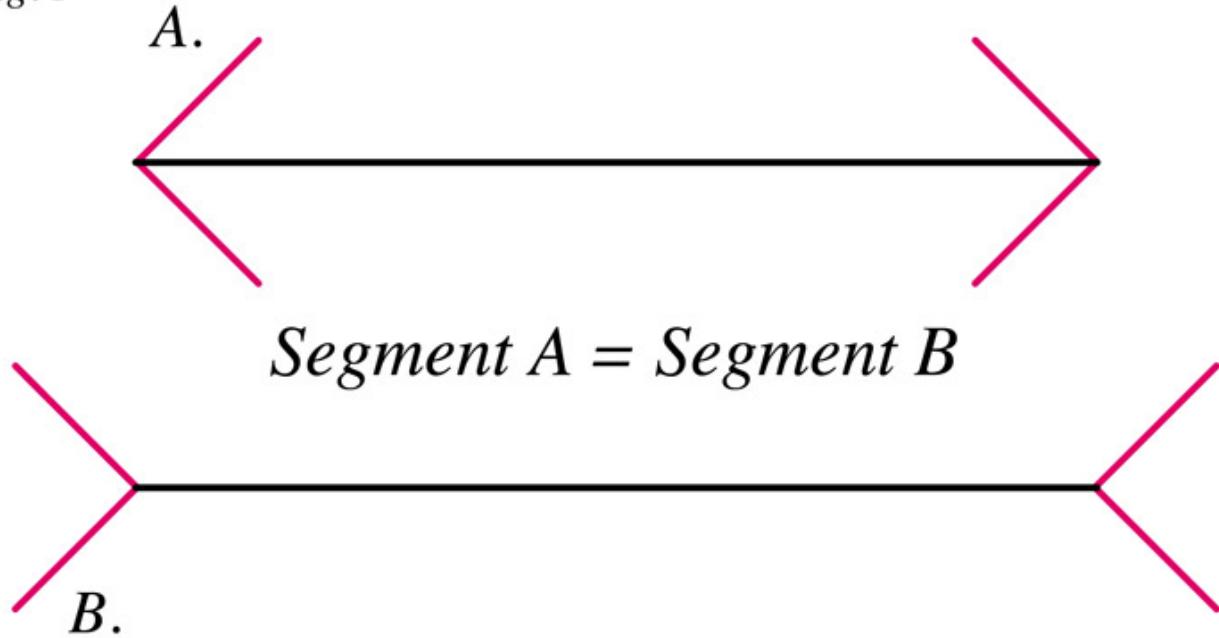


Figure 8: This figure illustrates the line width illusion and its impact on frequency estimation in parallel coordinate plots.

No, it does not. This is Müller-Lyer, not the line width illusion

3.1 The Problem: Overlapping Lines

When multiple observations have the same value for a variable, their lines converge to a single point on that axis. This convergence creates visual clutter and breaks the continuity needed for individual observation tracking.

3.2 The Solution: Tie Spreading

The `tie_spread` algorithm separates tied observations by distributing them evenly within a small range centered on their original value. By default, this range is set to maximize the available space for tiebreaking while preserving the approximate positions of values and maintaining visual separation. *No, it only takes space up to what is necessary to break ties. - goal should be to balance tie breaking*.

3.3 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.

The Problem: Numerical Ties Create Overlapping Lines

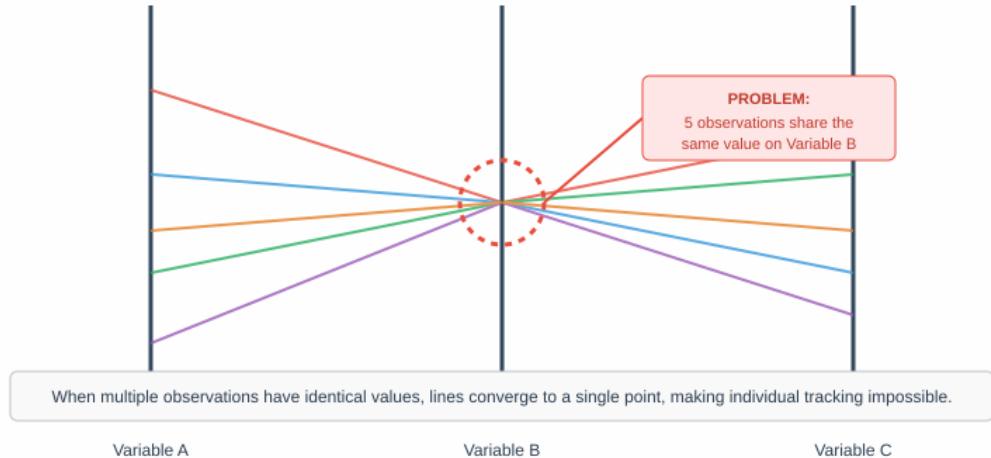


Figure 9: Numerical ties cause lines to converge at a single point.

The Solution: Even Tie Spreading Algorithm

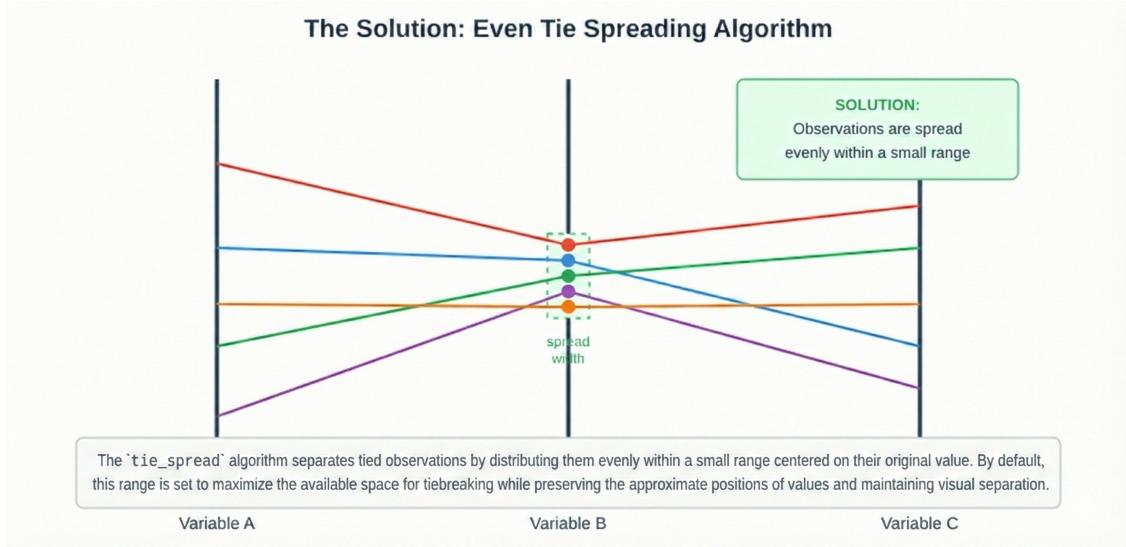


Figure 10: Tie spreading maintains visual separation while preserving approximate positions.

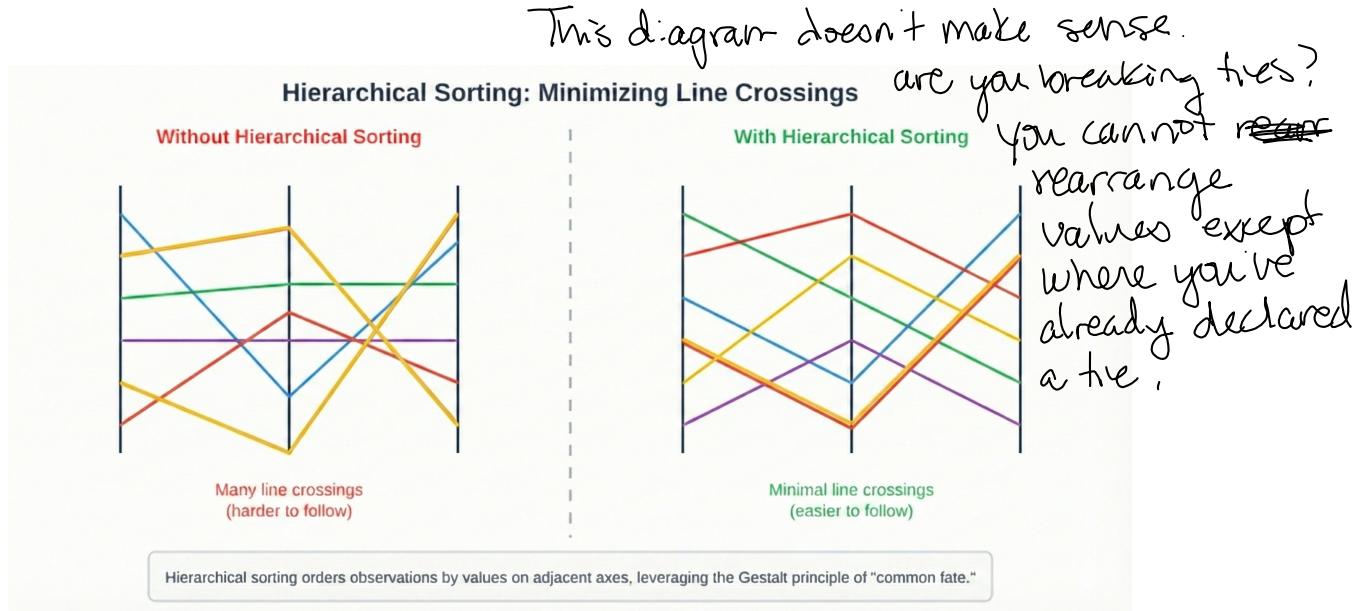


Figure 11: Hierarchical sorting reduces line crossings, making patterns easier to perceive.

3.4 Theoretical Framework: Perception-Driven Design

Effective data visualizations must be grounded in how people actually see and process information. This theoretical framework draws on research from cognitive psychology and visual perception to explain how viewers interpret visual displays. The framework covers four key areas that trace the path from seeing a visualization to understanding its meaning. **Stage 1** explores how Gestalt principles help viewers group visual elements into meaningful patterns. **Stage 2** examines how visual working memory and attention allow people to hold and manipulate information while analyzing a visualization. **Preattentive processing** describes how certain visual features can be detected quickly and automatically, enabling rapid visual search before focused attention is required. Finally, **external cognition and computational offloading** addresses how visualizations act as external aids that extend our mental capabilities by shifting some cognitive work from our minds to the visual display itself. These four components provide a foundation for understanding how design choices affect visualization effectiveness.

3.4.1 Stage 1: Pattern Perception and Gestalt Principles

The first stage involves active pattern perception processes that segment the visual scene into coherent regions. The Gestalt laws of perceptual organization, articulated by German psychologists in 1912, remain fundamental principles for visualization design:

- **Proximity:** Objects near each other are perceived as belonging together
- **Similarity:** Objects sharing visual properties (color, shape, size) are grouped perceptually
- **Connectedness:** Objects connected by lines are seen as related
- **Continuity:** The visual system prefers interpretations with smooth, continuous paths
- **Symmetry:** Symmetric patterns are more easily perceived and remembered
- **Closure:** Incomplete shapes are perceptually completed
- **Common Fate:** Objects moving together are perceived as a group

- **Relative Size:** Smaller elements are perceived as figures against larger backgrounds

VanderPlas et al. (2023) explicitly invoke these principles in their design rationale for ggpcp:

“By reducing the number of line crossings at non-axis points simplifies the plot, reducing the overall cognitive load required to ‘untangle’ (literally and metaphorically) the individual observations and leveraging the Gestalt principle of common fate.” (p. 4)

The principle of **good continuation** is particularly relevant to parallel coordinate plots: can viewers smoothly follow individual lines through numerical ties, or do visual discontinuities disrupt the perceptual flow?

3.4.2 Stage 2: Visual Working Memory and Attention

The second stage involves conscious attention and visual working memory, which holds only a limited number of objects (typically 3-4) for active processing.

This three-stage model has direct implications for evaluating tie-breaking strategies:

Equispaced Lines:

- Leverage preattentive orientation and position
- Support continuous line following (good continuation)
- Implicitly encode frequency through density (requires stage 1 processing)

Constant-Width Boxes:

- Leverage preattentive size and area
- Explicitly encode frequency (reduced cognitive load)
- May disrupt line continuity (challenges good continuation)

3.5 Preattentive Processing and Visual Search

A key distinction between the two approaches concerns the nature of visual search required for different analytical tasks. Ware distinguishes between:

- **Parallel Search (Preattentive):** Target pops out immediately, search time independent of number of distractors
- **Serial Search (Attentive):** Search time increases linearly with number of items, requiring conscious attention

For tracing individual observations through equispaced lines, the task may benefit from preattentive processing of orientation and position. Each observation maintains a consistent visual identity as a continuous line with specific orientation. In contrast, tracing through constant-width boxes requires conscious attention to reconstruct the path, as the area-based encoding does not support preattentive line following.

However, for frequency estimation tasks, constant-width boxes may have an advantage. Equispaced lines require inferring frequency from density, which is a second-order visual property requiring pattern perception (stage 1).

3.5.1 External Cognition and Computational Offloading

Scaife and Rogers (1996) concept of “external cognition”:

“While hierarchical sorting requires additional computations relative to the jittering or equally spaced solutions in Figure 5, this extra processing serves as ‘external cognition’ [Scaife and Rogers, 1996] - the additional computer time reduces the cognitive load required to untangle the lines” (p. 11)

we've already discussed this, unless you're going into every-one who did mem load stuff, ...
This principle aligns with Ware's emphasis on reducing working memory load through careful visual design. By computing optimal line positions algorithmically, the equispaced approach offloads cognitive work from the viewer. Rather than the viewer mentally sorting and organizing lines, the visualization presents pre-organized information that leverages natural perceptual grouping.

Hammock plots achieve a different form of cognitive offloading by aggregating observations into boxes, explicitly computing and displaying frequency information. The trade-off is between individual-level detail (equispaced lines) and aggregate-level clarity (boxes).

3.5.2 Gestalt Principles and Line Continuity

mention Ware. You make him sound way more important than he actually is.
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 and provides an immediate visual representation of bivariate contingency relationships. However, tracing a single observation through a box requires viewers to mentally reconstruct the implied path, engaging conscious attention and working memory. In this sense, the two approaches serve distinct analytical purposes: equispaced lines support individual-level tracing, while boxes support aggregate-level frequency comparison.

This distinction becomes increasingly relevant as dataset size grows. With many observations, equispaced lines may coalesce into ribbon-like bands, yet the continuity principle still applies—viewers can perceive flow patterns even when individual lines are indistinguishable. Boxes, by contrast, maintain their aggregate interpretation but are not designed to preserve individual-level information. Rather than viewing these as competing representations, they can be understood as complementary: the line-based approach is optimal when the analytical task involves following specific cases or detecting outliers, whereas the box-based approach is optimal when the task involves comparing category frequencies or understanding distributional patterns.

3.5.3 The Challenge of Area Perception *Not necessary*

Although area is preattentively processed, humans demonstrate notable difficulty in accurately comparing areas, particularly when aspect ratios vary. This limitation is well documented in Cleveland and McGill (1984)'s hierarchy of elementary perceptual tasks, which ranks position along a common scale as the most accurate encoding, followed by length, angle, slope, area, and volume in decreasing order of precision.

Hammock plots and related box-based visualizations rely on area perception for frequency encoding. Observers must compare box widths (area divided by length) to estimate relative frequencies. The line width illusion and its reverse counterpart, documented extensively by Hofmann and Vendettuoli (2013), further complicate these judgments by introducing systematic biases. Specifically, viewers tend to assess the orthogonal distance between curved or angled lines rather than the vertical distance that actually encodes the data, leading to perceptual distortion when comparing ribbons drawn at different angles.

Equispaced lines rely instead on density estimation, viewers count or estimate the number of lines within a given region. While density estimation also presents perceptual challenges, research suggests it may be less susceptible to geometric illusions that affect area judgment. Furthermore, density provides a more continuous encoding: frequencies can be estimated at any scale, from individual line counts to overall distributional patterns. Common angle plots, which use a fixed angle for all connecting segments between categories, represent one attempt to neutralize line width illusions in box-based displays while preserving aggregate frequency information.

3.5.4 Visual Clutter and Information Density *We aren't addressing this yet. At most, add it to a "to investigate" list*

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 a 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.

3.6 Problem Statement and Research Question

3.6.1 Main Research Question

How can visualization systems implement hybrid representations that utilize full even spacing to resolve numerical ties, thereby preserving individual-level tracing capabilities while simultaneously encoding aggregate frequency information to facilitate seamless transitions between granular and high-level analytical perspectives?

Parallel coordinate plots are a primary tool for multivariate data analysis, allowing researchers to observe relationships across multiple dimensions simultaneously. Their strength lies in individual-level tracing, where a single data point's "path" can be followed through various axes to identify profiles and outliers.

3.6.2 The Specific Conflict: Numerical Ties

A significant breakdown occurs when numerical ties (identical values) are present. Standard implementations cause these lines to overlap perfectly, creating two distinct issues:

- Loss of Visibility: Multiple observations appear as a single line, masking the true density or frequency of the data at that coordinate.
- Traceability Failure: When lines converge and then diverge at a tie, it becomes visually impossible to determine which incoming line corresponds to which outgoing line, a phenomenon known as "ambiguous crossings."

3.6.3 The Limitations of Current Solutions

Existing visualization systems typically force a trade-off between two extremes:

- Individual Spreading: Techniques like equispaced line-spreading (e.g., "tie spreading") separate tied values to restore individual traceability. However, this often distorts the visual representation of density, making it difficult to judge the relative frequency of values at a glance.
- Aggregate Encoding: Approaches like constant-width boxes or frequency bars effectively communicate the "weight" of a tie (the forest) but completely sever the user's ability to trace an individual observation (the tree) through the plot.

3.6.4 Gap in Research

There is currently no established hybrid representation that preserves the mathematical integrity of individual line paths while simultaneously encoding aggregate frequency information. This lack of integration forces users to toggle between different views, creating cognitive friction and preventing a fluid transition between granular and high-level analytical perspectives during the discovery process.

Present implementations show a split:

Repetitive -
You need to streamline →
re-organize the paper
so that
it flows
instead of
echoing.

1. **ggpcp**: This method uses Gestalt principles of continuity and common fate to break ties between categorical variables in a smart way. However, when numbers are tied, the lines overlap.
2. **Hammock plots**: All variable types have boxes with the same width, which makes frequency encoding clear, but you can't follow individual observations without highlighting them.

There is no current research that thoroughly evaluates these methodologies from the perspective of perceptual science or adapts ggpcp's optimization framework to numerical variables.

This gap is important because:

- Mixed-type datasets are common in modern data analysis.
- Different ways of visualizing data may work best for different types of analysis.
- Perceptual trade-offs have not been quantified or empirically validated.
- Design choices are not based on established principles of human visual perception.

3.6.5 Suggested Fix

We suggest that ggpcp's categorical tie-breaking algorithm be expanded to include numerical variables, using the same optimization framework that keeps things evenly spaced.

For a number v with n tied observations:

1. Find out how much space is available by looking at the values and data density of nearby items.
2. Use spacing optimization: $spacing = \frac{available_space}{n-1}$
3. Organize observations in a hierarchical way and spread them out so they are perpendicular to the axis direction.
4. Keep the rules of visual continuity in mind by making sure that lines flow smoothly from one to the next.

This creates a single framework where the same algorithm deals with both categorical and numerical ties. This could make things easier on the brain by keeping the visual grammar consistent and keeping the visual continuity across mixed-type data.

3.6.6 Hierarchical Sorting for Line Crossing Minimization

To minimize line crossings and leverage the Gestalt principle of common fate, observations within each tie group are sorted hierarchically based on values on adjacent axes.

Algorithm (Simplified):

```
For each tied group on axis i:
  Sort observations by (value on axis i+1, value on axis i-1)
  Assign positions y_{i,j} in sorted order
```

This ensures that observations with similar trajectories are positioned near each other, creating visual bands that support the perception of common fate.

3.7 Implementation and Evaluation Plan

3.7.1 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):

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

Study 1: Frequency Perception

- **Duration:** Weeks 5–7
- **Participants:** $n = 60$ (30 per condition)
- **Design:** Between-subjects

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

3.9 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

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

Deliverables:

- To not presume to control ggpcp. Heike will get protective.*
1. R Package Release (`ggpcp v2.0`):
 - 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: March 2026 (IEEE VIS)

3. Supplementary Materials:

- Open Science Framework repository
- All experimental stimuli and data
- Reproducibility package

4 Timeline and Milestones

5 Expected Contributions

5.1 Theoretical Contributions

Unified Tie-Breaking Theory: Formalization of equispaced optimization for both categorical and numerical variables, with key perceptual properties (separability, continuity, space efficiency)

Table 1: Research Timeline (12 Weeks)

| Week | Milestone | Deliverable | Perceptual Focus |
|-------|----------------------------------|--|-----------------------------------|
| 1–2 | Literature synthesis | Annotated bibliography with Ware framework integration | Ware framework integration |
| 3–4 | Core algorithm development | Working R code with perceptual properties verified | Perceptual properties implemented |
| 4 | IRB submission | Approved protocol for human subjects research | Human subjects clearance |
| 5–7 | Study 1: Frequency perception | Raw data and statistical analysis | Magnitude estimation validation |
| 8–10 | Benchmarking | Performance report with computational metrics | Computational metrics |
| 10–12 | Integration | Draft manuscript and R package release | Synthesis and dissemination |

Table 3: Application Domains for Research Impact

| Domain | Application | Data Characteristics |
|-----------------------|----------------------------------|---|
| Healthcare | Patient trajectory visualization | Categorical diagnoses + numerical measurements |
| Manufacturing | Quality control monitoring | Defect categories + continuous sensor readings |
| Social Science | Survey analysis | Demographic categories + Likert scales |
| Finance | Portfolio analysis | Categorical sectors + numerical performance metrics |
| Education | Learning analytics | Course completion + assessment scores |

5.2 Practical Contributions

1. **Production Software:** Open-source R package extension immediately usable by practicing data scientists and researchers
2. **Evidence-Based Guidelines:** Design recommendations grounded in empirical evidence and perceptual theory:
 - When to use equispaced lines vs. constant-width boxes
 - How to set spacing parameters for optimal perceptual clarity
 - Task-specific visualization selection criteria
3. **Benchmark Suite:** Reusable experimental framework and stimuli for evaluating future parallel coordinate plot innovations

5.3 Methodological Contributions

1. **Theory-Driven Evaluation:** Demonstrates how established perceptual theory (Gestalt principles) can generate specific, testable hypothesis
2. **Multi-Method Approach:** Integrates behavioral experiments with computational benchmarking to triangulate findings
3. **Reproducible Research:** Complete open science package enabling replication and exten-

7 Limitations and Future Directions

7.1 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

7.2 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

8 Conclusion

This comprehensive examination proposal addresses a fundamental question in information visualization by grounding design decisions in Colin Ware’s three-stage model of visual information processing. The proposed equispaced line approach extends ggpcp’s proven categorical tie-breaking algorithm to numerical variables, creating a unified framework that maintains visual continuity, leverages preattentive processing, and supports Gestalt principles of good continuation and common fate.

Compared to hammock plots’ constant-width boxes, equispaced lines offer different perceptual trade-offs: enhanced individual observation traceability and smoother visual flow at the potential cost of explicit frequency encoding. This research contributes to theory by demonstrating how established perceptual principles generate testable hypotheses about visualization effectiveness, to practice by delivering open-source software and evidence-based design guidelines, and to methodology by exemplifying theory-driven, multi-method evaluation.

As Ware (2012) emphasizes, effective visualization must be grounded in understanding of human perception. This research takes that principle seriously, asking not just “which visualization looks better?” but rather “which visualization better aligns with the architecture of human visual processing for specific analytical tasks?” The answer, as with most visualization questions, is nuanced and context-dependent—but now it will be grounded in empirical evidence and perceptual theory.

References

- 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>.

- 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*.
- Peng, Wei, Matthew O Ward, and Elke A Rundensteiner. 2004. “Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering.” In *IEEE Symposium on Information Visualization*, 89–96. IEEE.
- Scaife, M., and Y. Rogers. 1996. “External Cognition: How Do Graphical Representations Work?” *International Journal of Human-Computer Studies* 45 (2): 185–213.
- 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>.
- Ware, Colin. 2012. *Information Visualization: Perception for Design*. 3rd ed. Waltham, MA: Morgan Kaufmann.
- Wegman, Edward J. 1990. “Hyperdimensional Data Analysis Using Parallel Coordinates.” *Journal of the American Statistical Association* 85: 664–75.