

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

Denise Bradford

2025-11-01

Table of contents

1	1. Introduction and Background	3
1.1	The Challenge of Visualizing Mixed-Type Data	3
1.2	Current State: ggpcp’s Categorical Tie-Breaking	3
1.3	Alternative Approach: Hammock Plots	4
1.4	Visual Comparison: Triangles vs. Boxes	4
1.5	The Line Width Illusion and Perceptual Constraints	4
2	2. Theoretical Framework: Perception-Driven Design	5
2.1	Colin Ware’s Model of Visual Information Processing	5
2.1.1	Stage 1: Parallel Processing of Low-Level Properties	5
2.1.2	Stage 2: Pattern Perception and Gestalt Principles	6
2.1.3	Stage 3: Visual Working Memory and Attention	6
2.2	Preattentive Processing and Visual Search	7
2.3	External Cognition and Computational Offloading	7
2.4	Gestalt Principles and Line Continuity	8
2.5	The Challenge of Area Perception	8
2.6	Visual Clutter and Information Density	8
3	3. Problem Statement and Research Questions	9
3.1	Gap in Research	9
3.2	Suggested Fix	9
3.3	Main Research Question	10

4	4. Hypotheses Grounded in Perceptual Theory	10
4.1	Hypothesis 1: Preattentive Line Tracing	10
4.2	Hypothesis 2: Gestalt Principles and Visual Continuity	11
4.3	Hypothesis 3: Working Memory Load	12
4.4	Hypothesis 4: Frequency Estimation Accuracy	13
4.5	Hypothesis 5: Categorical-Numerical Consistency	14
4.6	Hypothesis 6: Scalability and Graceful Degradation	14
5	5. Mathematical Formalization	15
5.1	Optimization Problem	15
5.2	Properties	16
5.3	Hierarchical Sorting for Line Crossing Minimization	17
6	6. Implementation and Evaluation Plan	17
6.1	Phase 1: Algorithm Development (Weeks 1-4)	17
6.2	Phase 2: Perceptual Validation Studies (Weeks 5-14)	18
6.3	Phase 3: Comparative Benchmarking (Weeks 13-15)	20
6.4	Phase 4: Integration and Dissemination (Weeks 15-16)	21
7	7. Expected Contributions	22
7.1	Theoretical Contributions	22
7.2	Practical Contributions	23
7.3	Methodological Contributions	23
7.4	Broader Impact	23
8	8. Timeline and Milestones	24
9	9. Limitations and Future Directions	24
9.1	Study Limitations	24
9.2	Theoretical Extensions	25
9.3	Design Space Exploration	25
9.4	Application Domains	26
10	10. Conclusion	26
11	References	26

1 1. Introduction and Background

1.1 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.

1.2 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:

“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

1.3 Alternative Approach: Hammock Plots

Hammock plots, introduced by Schonlau (2003, 2024), take a fundamentally different approach to handling both categorical and numerical variables. Rather than using individual 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) describe:

“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).

1.4 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)

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

1.5 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:

“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.

2 2. Theoretical Framework: Perception-Driven Design

2.1 Colin Ware’s Model of Visual Information Processing

Colin Ware’s seminal work “Information Visualization: Perception for Design” provides a comprehensive framework for understanding how humans process visual information (Ware 2012). Ware describes visual perception as occurring in three distinct stages, each with specific implications for visualization design.

2.1.1 Stage 1: Parallel Processing of Low-Level Properties

The first stage involves rapid, parallel extraction of basic visual features across the entire visual field. As Ware describes, this preattentive processing occurs in under 500 milliseconds and requires no conscious effort. Four fundamental categories of preattentive visual properties exist:

1. Form Attributes:

- Orientation
- Length
- Width
- Size
- Curvature
- Spatial grouping
- Blur/focus
- Added marks (enclosure)

2. Color Attributes:

- Hue
- Intensity (brightness)
- Saturation

3. Spatial Position:

- 2D position (x, y coordinates)
- Stereoscopic depth
- Concave/convex shading

4. Motion:

- Flicker
- Direction of movement

These preattentive properties are critical for visualization design because they allow information to be perceived “at a glance” without requiring serial visual search. As Ware notes, it is easy to spot a single hawk in a sky full of pigeons, but if the sky contains many types of birds, the hawk becomes much harder to see—the distinctiveness of preattentive properties diminishes as visual variety increases.

2.1.2 Stage 2: Pattern Perception and Gestalt Principles

The second 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?

2.1.3 Stage 3: Visual Working Memory and Attention

The third stage involves conscious attention and visual working memory, which holds only a limited number of objects (typically 3-4) for active processing. As Ware emphasizes, the severe capacity limitations of visual working memory mean that effective visualizations must reduce cognitive load by leveraging the powerful parallel processing of stages 1 and 2.

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 2 processing)

Constant-Width Boxes:

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

2.2 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. Ware notes that area is preattentively processed, allowing immediate perception of relative quantities. Equispaced lines require inferring frequency from density, which is a second-order visual property requiring pattern perception (stage 2).

2.3 External Cognition and Computational Offloading

VanderPlas et al. (2023) invoke (scaife1996?) 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)

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).

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 perfect continuity—each observation is represented by a continuous polyline from the leftmost to rightmost axis. When observations are tied at a numerical value, the spacing ensures that lines remain visually distinct without requiring discontinuities.

Constant-width boxes, in contrast, represent observations as area segments rather than lines. While this encoding effectively communicates aggregate quantities, it disrupts the continuity principle. To trace an individual observation through a box, viewers must mentally reconstruct the implied path, engaging conscious attention and working memory.

This distinction becomes critical as dataset size increases. With many observations, equispaced lines may appear as ribbon-like bands, but the continuity principle still applies—viewers can perceive flow patterns even when individual lines are indistinguishable. Boxes maintain their aggregate interpretation but lose any individual-level information.

2.5 The Challenge of Area Perception

While area is preattentively processed, humans are notoriously poor at accurately comparing areas, especially when aspect ratios vary. This limitation is well-documented in Cleveland and McGill (1984) hierarchy of elementary perceptual tasks, which ranks position along a common scale as most accurate, followed by length, angle, slope, area, and volume.

Hammock plots 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 reverse line width illusion Hofmann and Vendettuoli (2013) further complicate this judgment by creating systematic biases.

Equispaced lines instead rely on density estimation—counting or estimating the number of lines within a region. While density estimation is also challenging, it may be less susceptible to geometric illusions. Additionally, density provides a more continuous encoding: frequencies can be estimated at any scale, from individual line counts to overall patterns.

2.6 Visual Clutter and Information Density

Ware discusses the concept of “visual clutter” as a fundamental limitation on visualization effectiveness. Clutter occurs when too many visual elements compete for attention, overwhelming the viewer’s ability to extract meaningful patterns.

The two approaches differ in how they manage clutter:

Equispaced Lines:

- Clutter increases with number of observations
- Hierarchical sorting minimizes line crossings (reduces clutter)
- Graceful degradation: lines \rightarrow ribbons \rightarrow filled areas as density increases

- Individual observations remain theoretically accessible

Constant-Width Boxes:

- Clutter depends on number of unique value combinations
- Aggregation inherently reduces clutter
- Information loss: individual observations not accessible
- Clear representation of bivariate contingency tables

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.

3. Problem Statement and Research Questions

3.1 Gap in Research

Present implementations show a split:

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.

3.2 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.

3.3 Main Research Question

Is the equispaced line method better for numerical ties than the constant-width box method in hammock plots, based on how people see things?

This question breaks down into smaller, more specific questions based on perceptual theory:

Q1 (Preattentive Processing): Do equispaced lines facilitate preattentive detection and visual search more effectively than constant-width boxes for individual observation tracking?

Q2 (Pattern Perception): How do the two methods use Gestalt principles in different ways, and which one works better for seeing relationships in multivariate data?

Q3 (Memory and Cognition): Which method is better at lowering the cognitive load and visual working memory needs for everyday analytical tasks?

Q4 (How Often You Think): How well can viewers guess frequencies when using implicit density (equispaced lines) instead of explicit area encoding (boxes)?

Q5 (Scalability): How do the pros and cons of perception change when the size of the dataset and the distribution of ties change?

4 4. Hypotheses Grounded in Perceptual Theory

4.1 Hypothesis 1: Preattentive Line Tracing

H1: Equispaced lines support faster and more accurate individual observation tracing than constant-width boxes due to preattentive processing of orientation and position.

Perceptual Basis:

- Orientation is a preattentive visual property (Ware 2012)
- The principle of good continuation supports smooth path following
- Position along a common scale is Cleveland and McGill’s most accurate perceptual task

Operationalization:

- **Task:** “Trace observation X (highlighted in red) from axis A through axis B to axis C. What is its final value?”
- **Independent Variables:**
 - Visualization method (equispaced lines vs. constant-width boxes)

- Dataset size (50, 200, 1000 observations)
- Tie density (10%, 30%, 50% of observations tied)

- **Dependent Variables:**

- Completion time (milliseconds)
- Error rate (% incorrect responses)
- Eye-tracking metrics (fixation count, saccade length, time to first fixation on target)

Expected Results:

- Equispaced lines will show significantly faster completion times (H1a)
- Error rates will be lower for equispaced lines (H1b)
- Eye-tracking data will show fewer fixations and more direct saccades (H1c)
- Advantage will diminish as dataset size increases (interaction effect)

Experimental Design:

- Within-subjects factorial design
- 40 participants
- Counterbalanced presentation order
- Practice trials to control for learning effects
- Analysis: Mixed-effects ANOVA with subject as random effect

4.2 Hypothesis 2: Gestalt Principles and Visual Continuity

H2: Equispaced lines better support perceptual grouping and pattern detection through Gestalt principles of continuity and common fate.

Perceptual Basis:

- Good continuation: Visual system prefers smooth, continuous paths
- Common fate: Objects moving or flowing together are grouped perceptually
- Proximity: Lines close together are perceived as related

Operationalization:

- **Task A** (Pattern Detection): “Identify clusters or groups in the data”
- **Task B** (Relationship Assessment): “Are variables X and Y positively correlated, negatively correlated, or independent?”
- **Task C** (Subgroup Tracing): “Follow all observations in group A from left to right. Do they tend to have high or low values on the final variable?”

Dependent Variables:

- Pattern identification accuracy

- Confidence ratings (7-point Likert scale)
- Time to first correct response
- Qualitative descriptions of perceived patterns

Expected Results:

- Equispaced lines will lead to more accurate pattern identification
- Higher confidence ratings for equispaced lines
- Richer qualitative descriptions demonstrating deeper understanding
- Particularly strong advantage for relationship assessment tasks

Experimental Design:

- Between-subjects with 60 participants (30 per condition)
- Mixed categorical-numerical datasets with known correlation structure
- Think-aloud protocol during task completion
- Content analysis of verbal descriptions

4.3 Hypothesis 3: Working Memory Load

H3: Equispaced lines reduce visual working memory demands compared to constant-width boxes for tasks requiring integration across multiple axes.

Perceptual Basis:

- Visual working memory capacity limited to 3-4 objects (Ware 2012)
- Continuous lines can be treated as single objects (chunking)
- Box-based representations require reconstructing paths across discontinuities

Operationalization:

- **Primary Task:** Multi-axis comparison requiring integration of information
- **Secondary Task:** Concurrent memory load (remember random digit sequence)
- **Measure:** Dual-task performance decrement

Dependent Variables:

- Primary task accuracy under dual-task conditions
- Secondary task performance (digit recall accuracy)
- NASA Task Load Index (TLX) subjective workload assessment
- Pupil dilation (physiological measure of cognitive load)

Expected Results:

- Smaller performance decrement for equispaced lines under dual-task conditions

- Better secondary task performance with equispaced lines
- Lower subjective workload ratings
- Smaller pupil dilation changes

Experimental Design:

- Within-subjects, 40 participants
- Counterbalanced order
- Baseline single-task performance measured first
- Analysis: Compare dual-task decrement between conditions

4.4 Hypothesis 4: Frequency Estimation Accuracy

H4: Constant-width boxes enable more accurate frequency estimation than equispaced lines, but both approaches support adequate ordinal comparisons (rank ordering by frequency).

Perceptual Basis:

- Area is preattentively processed but poorly judged quantitatively (Cleveland and McGill 1984)
- Density estimation is challenging but supports relative comparisons
- Explicit encoding (boxes) vs. implicit encoding (line density)

Operationalization:

- **Task A** (Magnitude Estimation): “What percentage of observations have value v ?”
- **Task B** (Ordinal Comparison): “Which value has more observations: v or v' ?”
- **Task C** (Ratio Judgment): “Value v has how many times as many observations as v' ?”

Dependent Variables:

- Absolute percentage error for Task A
- Accuracy (% correct) for Task B
- Ratio estimation error for Task C
- Response time for each task

Expected Results:

- Hammock plots will show lower absolute error for magnitude estimation (H4a)
- No significant difference for ordinal comparisons (H4b)
- Hammock plots advantage for ratio judgments (H4c)
- Faster response times for hammock plots on frequency tasks

Experimental Design:

- Between-subjects, 60 participants
- Stimuli controlled for actual frequency distributions
- Range of frequencies tested (10% to 50% of observations)
- Analysis: ANOVA with planned contrasts

4.5 Hypothesis 5: Categorical-Numerical Consistency

H5: Using consistent visual encoding (equispaced lines for both categorical and numerical variables) reduces cognitive load compared to mixed encoding strategies.

Perceptual Basis:

- Consistent visual grammar reduces need for strategy switching
- Unified mental model reduces working memory demands
- External cognition principle: computational optimization reduces viewer effort

Operationalization:

- **Conditions:**
 1. ggpcp: Consistent equispaced approach
 2. Hybrid: Equispaced categorical, boxes for numerical
 3. Hammock: Consistent box approach
- **Tasks:** Mixed tasks requiring both tracing and frequency estimation
- **Measure:** Task-switching costs and overall performance

Dependent Variables:

- Learning curve (performance over time)
- Task-switching cost (RT increase after switching task types)
- Transfer effects (performance on novel task combinations)
- Subjective preference ratings

Experimental Design:

- Between-subjects, 60 participants (20 per condition)
- Training phase (10 minutes)
- Practice phase (5 minutes)
- Test phase (15 minutes, 12 mixed tasks)
- Post-test preference questionnaire
- Analysis: Growth curve modeling for learning effects

4.6 Hypothesis 6: Scalability and Graceful Degradation

H6: As dataset size increases, equispaced lines degrade gracefully from individual-level to aggregate-level representation, maintaining perceptual advantages throughout the transition.

Perceptual Basis:

- Perceptual aggregation: Dense lines perceived as unified regions

- Texture perception: Line density patterns processed preattentively
- Gestalt grouping: Proximity and similarity create emergent bands

Operationalization:

- **Dataset Sizes:** \$n = 50, 100, 500, 1000, 5000\$
- **Tie Distributions:**
 - Uniform (equal frequencies)
 - Zipfian (power-law distribution)
 - Extreme (80% in 20% of values)
- **Tasks:** Same tasks as H1-H4 administered at different scales

Dependent Variables:

- Performance metrics from H1-H4 measured across scales
- Subjective clarity ratings
- Point at which individual lines become indistinguishable (perceptual threshold)

Expected Results:

- Smooth performance degradation for equispaced lines
- Discontinuous transition for boxes (individual \rightarrow aggregate)
- Equispaced lines maintain relative advantage even at large scales
- Advantage most pronounced for extreme tie distributions

Experimental Design:

- Mixed design: visualization method (between) \times dataset size (within)
- 40 participants per between-subjects condition
- Subset of tasks from H1-H4 at each scale
- Analysis: Mixed-effects models with scale as continuous predictor

5 Mathematical Formalization

5.1 Optimization Problem

For a numerical axis with values $V = \{v_1, v_2, \dots, v_m\}$ ordered such that $v_1 < v_2 < \dots < v_m$, let n_i denote the number of observations with value v_i .

Objective: Assign positions $\{y_{i,j}\}$ for $j = 1, \dots, n_i$ such that:

1. **Non-intersection constraint:** $y_{i,j} \neq y_{i,k}$ for $j \neq k$ (maintain visual distinctiveness)

2. **Monotonicity constraint:** Within value v_i , hierarchical ordering is preserved (support common fate Gestalt)
3. **Continuity constraint:** Lines maintain smooth transitions between axes (leverage good continuation)
4. **Space efficiency:** Utilize available space proportionally to frequency (balanced visual density)

Space allocation:

For each value v_i , allocate proportional space:

$$S_i = \frac{n_i}{\sum_{k=1}^m n_k} \cdot (1 - space \cdot (m - 1))$$

where $space \in [0, 0.5]$ is the spacing parameter controlling separation between value regions.

Position assignment:

For observation j within value v_i (after hierarchical sorting):

$$y_{i,j} = start_i + \frac{S_i - 2 \cdot buffer}{n_i - 1} \cdot (j - 1) + buffer$$

where: - $start_i = \sum_{k=1}^{i-1} (S_k + space)$ is the starting position for value v_i - $buffer$ provides margin space at boundaries (typically $0.02 \times S_i$) - The formula ensures even distribution across available space

5.2 Properties

Theorem 1 (Non-intersection): For any two observations $j \neq k$ with the same value v_i :

$$y_{i,j} \neq y_{i,k}$$

Proof: By construction, $y_{i,j} = start_i + d_i \cdot (j - 1) + buffer$ where $d_i = \frac{S_i - 2 \cdot buffer}{n_i - 1}$. For $j \neq k$, we have $(j - 1) \neq (k - 1)$. When $n_i > 1$, $d_i > 0$, thus $y_{i,j} \neq y_{i,k}$. When $n_i = 1$, the property holds trivially. \square

Theorem 2 (Perceptual Separability): Any two observations with different values occupy non-overlapping vertical regions:

$$\max_j(y_{i,j}) + \frac{space}{2} < \min_j(y_{i+1,j}) - \frac{space}{2}$$

Proof: The maximum position for value v_i is $start_i + S_i$. The minimum position for value v_{i+1} is $start_{i+1} = start_i + S_i + space$. The spacing parameter $space$ creates explicit separation. \square

This property ensures preattentive separability: observations with different values can be distinguished without serial search.

Theorem 3 (Space Utilization): The algorithm achieves optimal space utilization:

$$\sum_{i=1}^m S_i = 1 - space \cdot (m - 1)$$

Proof: By construction, $S_i = \frac{n_i}{\sum_k n_k} \cdot (1 - space \cdot (m - 1))$. Summing over all values:

$$\sum_{i=1}^m S_i = \sum_{i=1}^m \frac{n_i}{\sum_k n_k} \cdot (1 - space \cdot (m - 1)) = \frac{\sum_i n_i}{\sum_k n_k} \cdot (1 - space \cdot (m - 1)) = 1 - space \cdot (m - 1)$$

□

Corollary (Visual Density Proportionality): Visual density within each region is proportional to observation count, supporting density-based frequency inference.

5.3 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.

6. Implementation and Evaluation Plan

6.1 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()` 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):

```
optimize_spacing_numerical <- function(values, n_obs, space = 0.05, buffer = 0.02) {  
  # Calculate space allocation for each unique value  
  unique_vals <- sort(unique(values))  
  m <- length(unique_vals)  
  
  # Frequency table  
  freq <- table(values)  
  
  # Total available space (excluding spacing between regions)  
  available <- 1 - space * (m - 1)  
  
  # Proportional allocation  
  proportions <- freq / sum(freq) * available  
  
  # Calculate positions with perceptually-motivated spacing  
  positions <- numeric(length(values))  
  cumulative <- 0  
  
  for (i in seq_along(unique_vals)) {  
    val <- unique_vals[i]  
    idx <- which(values == val)  
    n_i <- length(idx)  
  
    if (n_i == 1) {  
      # Center single observations  
      positions[idx] <- cumulative + proportions[i]/2  
    } else {  
      # Even distribution with buffer margins  
      usable_space <- proportions[i] - 2*buffer*proportions[i]  
      d_i <- usable_space / (n_i - 1)  
      positions[idx] <- cumulative + buffer*proportions[i] +  
        seq(0, usable_space, length.out = n_i)  
    }  
  
    cumulative <- cumulative + proportions[i] + space  
  }  
  
  return(positions)  
}
```

6.2 Phase 2: Perceptual Validation Studies (Weeks 5-14)

Study 1: Preattentive Processing and Visual Search (H1)

- **Duration:** Weeks 5-7

- **Participants:** $n = 40$ recruited via university participant pool
- **Compensation:** Course credit or \$10
- **Apparatus:** Eye-tracking system (EyeLink 1000)
- **IRB:** Submitted week 4

Procedure:

1. Calibration and practice (5 minutes)
2. 48 trials: trace observation from axis 1 to axis 3
3. Counterbalanced design: half trials with equispaced lines, half with boxes
4. Record: response time, accuracy, eye movements

Analysis:

- Mixed-effects regression: $RT \sim \text{method} \times \text{dataset_size} + (1|\text{participant})$
- Eye-tracking metrics: fixation count, saccade amplitude
- Compare to predictions from preattentive processing theory

Study 2: Gestalt Principles and Pattern Perception (H2)

- **Duration:** Weeks 8-10
- **Participants:** $n = 60$ (30 per condition)
- **Compensation:** \$12
- **Design:** Between-subjects with think-aloud protocol

Procedure:

1. Training on visualization type (10 minutes)
2. Pattern detection task (6 trials)
3. Relationship assessment task (6 trials)
4. Subgroup tracing task (6 trials)
5. Post-test questionnaire on strategies used

Analysis:

- Accuracy rates compared between conditions
- Content analysis of think-aloud transcripts
- Thematic coding of perceived patterns
- Correlation with participant visualization experience

Study 3: Working Memory and Cognitive Load (H3)

- **Duration:** Weeks 11-12
- **Participants:** $n = 40$

- **Compensation:** \$15
- **Apparatus:** Eye-tracking with pupillometry

Procedure:

1. Single-task baseline (12 trials each visualization)
2. Dual-task condition: visualization task + digit span (24 trials)
3. NASA-TLX workload assessment
4. Post-test preference and strategy questionnaire

Analysis:

- Dual-task decrement: $(\text{single_task_acc} - \text{dual_task_acc}) / \text{single_task_acc}$
- Pupil dilation change from baseline
- Repeated measures ANOVA on workload subscales
- Correlational analysis: working memory capacity (OSPAN) \times visualization method

Study 4: Frequency Perception (H4)

- **Duration:** Weeks 12-14
- **Participants:** $n = 60$ (30 per condition)
- **Compensation:** \$10
- **Design:** Between-subjects

Procedure:

1. Magnitude estimation (12 trials): “What % have value X?”
2. Ordinal comparison (12 trials): “Which value 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

6.3 Phase 3: Comparative Benchmarking (Weeks 13-15)

Computational Metrics:

1. **Rendering Performance:**
 - Time to generate visualization (milliseconds)

- Memory usage (MB)
- Scalability analysis (n = 100 to 100,000)

2. Visual Quality Metrics (Dennig et al., 2021):

- Line crossing count (algorithmic)
- Ribbon overlap proportion
- Visual clutter index
- Information density (observations per screen area)

3. Perceptual Quality Estimates:

- Modeled eye movements based on saliency maps
- Predicted visual search time using Fitts' Law
- Gestalt quality score (proximity + continuity + common fate)

Case Studies:

Case Study 1: Palmer Penguins

- Mixed categorical-numerical data
- Known correlation structure
- Expert validation of insights discovered

Case Study 2: Auto MPG Dataset

- Multiple numerical variables with natural ties
- Replication of classic visualization examples
- Comparison to historical analyses

Case Study 3: Asthma Data (Schonlau, 2024)

- Direct comparison with published hammock plot
- Same data, same research questions
- Evaluate relative effectiveness for medical decision-making

6.4 Phase 4: Integration and Dissemination (Weeks 15-16)

Deliverables:

1. R Package Release (ggpcp v2.0):

- CRAN submission with full documentation
- Vignette: “Handling Numerical Ties in Parallel Coordinate Plots”
- Tutorial examples with perceptual design rationale
- Unit tests achieving >95% coverage

2. Academic Paper:

- Target: IEEE Transactions on Visualization and Computer Graphics
- ~10 pages (double column format)
- Structure:
 - Abstract (150 words)
 - Introduction (1 page): Problem motivation
 - Related Work (1.5 pages): Perceptual theory + existing approaches
 - Method (2 pages): Algorithm + perceptual rationale
 - Evaluation (3 pages): Studies 1-4 results
 - Discussion (1.5 pages): Implications + limitations
 - Conclusion (0.5 pages): Contributions + future work
- Submission deadline: March 2026 (IEEE VIS)

3. Supplementary Materials:

- Open Science Framework repository
- All experimental stimuli and data
- Analysis scripts (R Markdown)
- Preregistration documents
- Reproducibility package

4. Presentation Materials:

- Conference talk (20 minutes)
- Poster (36" × 48")
- Video demo (3 minutes)
- Blog post for broader audience

7. Expected Contributions

7.1 Theoretical Contributions

1. **Perceptually-Grounded Framework:** First comprehensive application of Ware’s three-stage model to evaluate parallel coordinate plot design decisions, establishing methodology for theory-driven visualization evaluation
2. **Unified Tie-Breaking Theory:** Formalization of equispaced optimization for both categorical and numerical variables, with mathematical proofs of key perceptual properties (separability, continuity, space efficiency)
3. **Empirical Characterization:** Rigorous experimental evidence quantifying perceptual trade-offs between individual-level and aggregate-level encoding strategies across multiple task types
4. **Scalability Model:** Theory of graceful degradation describing how equispaced lines transition from discrete to continuous representation as density increases

7.2 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
4. **Teaching Resources:** Documented case studies demonstrating application of perceptual principles to visualization design decisions

7.3 Methodological Contributions

1. **Theory-Driven Evaluation:** Demonstrates how established perceptual theory (Ware, Gestalt principles, Cleveland & McGill) can generate specific, testable hypotheses about visualization effectiveness
2. **Multi-Method Approach:** Integrates eye-tracking, dual-task paradigm, think-aloud protocol, and algorithmic analysis to triangulate findings
3. **Reproducible Research:** Complete open science package enabling replication and extension by other researchers

7.4 Broader Impact

This research addresses visualization challenges across numerous domains:

Healthcare: Patient trajectory visualization with mixed clinical variables (categorical diagnoses + numerical measurements)

Manufacturing: Quality control monitoring combining defect categories with continuous sensor readings

Social Science: Survey analysis integrating demographic categories with Likert scales and open-ended responses

Finance: Portfolio analysis mixing categorical sectors with numerical performance metrics

Education: Learning analytics combining course completion (categorical) with assessment scores (numerical)

By providing both theoretical understanding and practical tools, this work enables more effective exploratory data analysis across these domains, ultimately supporting better data-driven decision making.

8 8. Timeline and Milestones

Week	Milestone	Deliverable	Perceptual Focus
1-2	Literature synthesis	Annotated bibliography	Ware framework integration
3-4	Core algorithm	Working R code	Perceptual properties implemented
4	IRB submission	Approved protocol	Human subjects clearance
5-7	Study 1: Visual search	Raw data + analysis	Preattentive processing
8-10	Study 2: Pattern perception	Raw data + analysis	Gestalt principles
11-12	Study 3: Cognitive load	Raw data + analysis	Working memory
12-14	Study 4: Frequency	Raw data + analysis	Magnitude estimation
13-15	Benchmarking	Performance report	Computational metrics
15-16	Integration	Draft manuscript	Synthesis

Total Duration: 16 weeks (4 months)

Key Dependencies: - IRB approval required before any human subjects research (Week 4) - Algorithm implementation must be stable before Study 1 (Week 5) - Studies 1-4 can run partially in parallel (Weeks 8-14) - Benchmarking overlaps with final studies (Weeks 13-15)

9 9. Limitations and Future Directions

9.1 Study Limitations

Sample Characteristics:

- University student population may not generalize to domain experts
- Limited representation of colorblind individuals
- Computer-based tasks may not reflect real-world analytical workflows

Mitigation Strategies:

- Include visualization experience as measured covariate
- Supplementary expert validation study (case studies)
- Colorblind-friendly palette testing in pilot phase

Experimental Constraints:

- Laboratory tasks may lack ecological validity
- Limited number of trial repetitions per condition
- Practice effects despite counterbalancing

Future Extensions:

- Longitudinal study with domain experts
- In-situ evaluation in real analytical workflows
- Cross-cultural validation of perceptual effects

9.2 Theoretical Extensions

Interaction Effects:

- How do individual differences in perceptual abilities moderate effects?
- Does domain expertise change relative advantages?
- What role does color encoding play in combination with spacing?

Cognitive Modeling:

- Develop computational model predicting performance from perceptual properties
- Formal ACT-R or EPIC cognitive architecture implementation
- Quantitative predictions for novel visualization variations

Perceptual Phenomena:

- Investigation of emergent Gestalt effects at varying scales
- Characterization of density-to-area perception mapping
- Study of attention allocation patterns using entropy measures

9.3 Design Space Exploration

Hybrid Approaches:

- Smooth interpolation between lines and boxes based on dataset properties
- Adaptive visualization that adjusts encoding based on detected user task
- Level-of-detail rendering that transitions strategically

Novel Encodings:

- Combining color intensity with spacing to encode frequency
- Using texture patterns within equispaced regions
- Incorporating animation to support tracing tasks

Interactive Techniques:

- Dynamic highlighting on hover to support box-based frequency perception
- Lens-based magnification for dense regions
- Semantic zooming that transitions between individual and aggregate views

9.4 Application Domains

Domain-Specific Validation:

- Medical informatics: Clinical decision support visualizations
- Financial analytics: Portfolio risk assessment tools
- Manufacturing: Quality control dashboards
- Education: Learning analytics platforms

Specialized Datasets:

- Time series with categorical events
- Spatial data with attribute tables
- Network data with node attributes
- Text data with metadata

10 10. Conclusion

This comprehensive exam addresses a fundamental question in information visualization: how should we visually encode multiple observations sharing the same numerical value in parallel coordinate plots? By grounding this design decision in Colin Ware’s three-stage model of visual information processing and conducting rigorous perceptual experiments, this research will establish evidence-based guidelines for visualization practitioners.

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.

The research contributes to theory by demonstrating how established perceptual principles generate testable hypotheses about visualization effectiveness. It contributes to practice by delivering open-source software and evidence-based design guidelines. And it contributes to methodology by exemplifying theory-driven, multi-method evaluation combining eye-tracking, dual-task paradigms, and think-aloud protocols with computational benchmarking.

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.

11 References

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 531-554.

- Dennig, F. L., Bögl, M., & Keim, D. A. (2021). ParSetgnostics: Quality metrics for parallel sets. *Computer Graphics Forum*, 40(3), 375-386. <https://doi.org/10.1111/cgf.14314>
- Hofmann, H., & Vendettuoli, M. (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, A. (1985). The plane with parallel coordinates. *The Visual Computer*, 1, 69-91. <https://doi.org/10.1007/BF01898350>
- Kosara, R., Bendix, F., & Hauser, H. (2006). Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4), 558-568. <https://doi.org/10.1109/TVCG.2006.76>
- Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45(2), 185-213.
- Schonlau, M. (2003). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics*, 19(1), 95-111.
- Schonlau, M. (2024). The hammock plot: Where categorical and numerical data relax together. *The Stata Journal*, 24(1), 1-21. <https://doi.org/10.1177/1536867X241234567>
- Schonlau, M., & Yang, T. (2024). The hammock plot: Where categorical and numerical data relax together. arXiv:2506.13630v1 [stat.AP].
- Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.
- VanderPlas, S., Ge, Y., Unwin, A., & Hofmann, H. (2023). Penguins go parallel: A grammar of graphics framework for generalized parallel coordinate plots. *Journal of Computational and Graphical Statistics*. Advance online publication.
- Wallgren, A., Wallgren, B., Persson, R., Jorner, U., & Haaland, J. (1996). *Graphing Statistics & Data: Creating Better Charts*. Sage Publications.
- Ware, C. (2004). *Information Visualization: Perception for Design* (2nd ed.). Morgan Kaufmann.
- Ware, C. (2012). *Information Visualization: Perception for Design* (3rd ed.). Morgan Kaufmann.
- Ware, C., Purchase, H. C., Colpoys, L., & McGill, M. (2002). Cognitive measurements of graph aesthetics. *Information Visualization*, 1(2), 103-110.
- Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411), 664-675.
- 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>.

- 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.