

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

Comprehensive Exam Presentation

Denise Bradford

University of Nebraska–Lincoln

November 2025

Introduction

Motivation

Chapter 1: Exploring Rural Shrink Smart Through Guided Discovery Dashboards

Context: Interactive dashboard for small Iowa towns experiencing population decline
(Bradford and VanderPlas 2023)

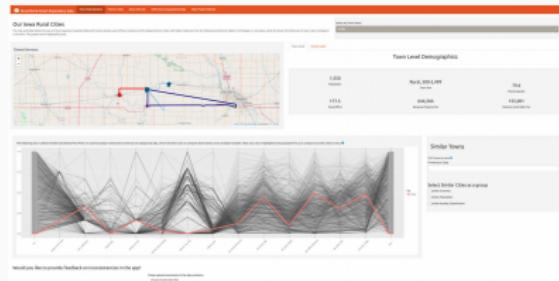


Figure 1: Dashboard from the Rural Shrink Smart Project

Component	Implementation	Connection to Dissertation
Visualization type	Parallel coordinate plots	Direct application of PCP methods
Data complexity	900+ towns, 50+ variables	High-dimensional multivariate data
User base	Town leaders (non-experts)	Novice analyst use case
Challenge	Numerical ties in census data	Exact problem this work solves
Solution applied	Deterministic jittering	Enables meaningful comparisons

Key Frustration & Insight: The dashboard was unusable because numerical ties (e.g., multiple towns with the same median income) caused massive overplotting. This visual clutter obscured individual community profiles, preventing users from comparing their town to others—the dashboard's main goal—and blocking data-driven decisions.

Research Overview

Central Challenge:

Numerical ties in parallel coordinate plots create severe visual occlusion that fundamentally compromises data exploration.

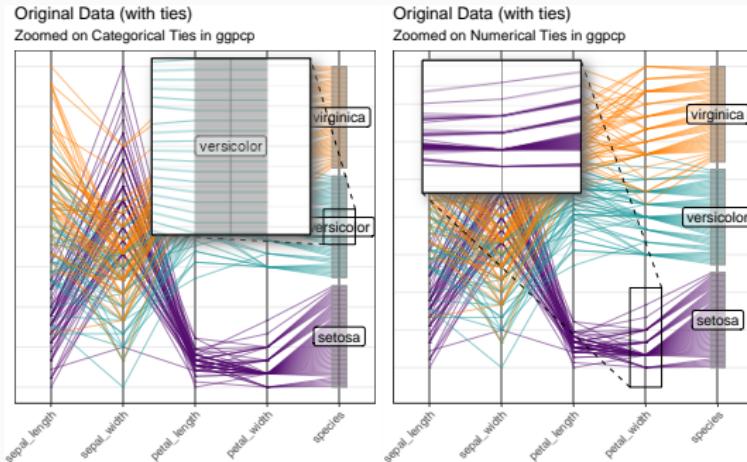


Figure 2: Comparison of current state of tie methods currently in ggpcp

Our Approach:

- Scientific reproducibility
- Perceptual validity
- Computational efficiency
- Theoretical rigor

Expected Contribution:

Complete Grammar of Graphics framework for tie resolution in high-dimensional visualization.

Extensions Beyond PCPs

Generalization Opportunities:

Visualization Type	Problem	Solution
2D scatter plots	Overplotting	Full 2D Halton
Time series	Multiple series overlap	Vertical jittering
Network layouts	Node positioning	Uniform spacing
Heatmaps	Discrete values	Cell jittering

General Principle:

Wherever random jitter is used, deterministic low-discrepancy alternatives should be considered for reproducible visualization.

Problem Statement

Parallel Coordinate Plots: n -dimensional data visualization using parallel axes
(Inselberg 1985; Wegman 1990)

The Numerical Tie Problem:

Issue	Description	Impact
Visual collision	Multiple observations overlap perfectly	Density unknown
Information loss	Cannot distinguish 1 from 1,000 observations	Analysis invalid
Structural occlusion	Substructure within ties hidden	Patterns missed
Tracing impossible	Cannot follow individual observations	Exploration fails

Research Context

Historical Development:

Year	Contribution	Reference
1985	PCPs introduced	Inselberg (1985)
1990	Statistical applications	Wegman (1990)
2000s	Overplotting identified	Johansson and Forsell (2016)
2003	Hammock plots alternative	Schonlau (2003)
2023	ggpcp Grammar framework	VanderPlas et al. (2023)
2025	Numerical tie resolution	This work

Current State:

- Categorical ties: Solved (hierarchical sorting)
- Numerical ties: Unsolved → **My contribution**

The ggpcp Foundation

Architecture: Three-module separation of concerns

Module	Function	Purpose	Status
1. Selection	<code>pcp_select()</code>	Choose/order dimensions	Complete
2. Scaling	<code>pcp_scale()</code>	Normalize scales	Complete
3. Tie Resolution	<code>pcp_arrange()</code>	Handle overlap	Extending

Our Extension:

- Existing: Categorical ties via hierarchical sorting
- Adding: Numerical ties via deterministic jittering (an uniformly constraint applied)
- Result: Complete tie-handling framework

Terminology and Definitions

Core Concept: Numerical Ties

Definition: Multiple observations sharing identical numerical values, causing perfect visual overlap

Origin Sources:

Source	Mechanism	Prevalence	Example
Rounding	Limited precision	Very high	Heights to 0.1m
Instruments	Discrete sensors	High	Integer counts
Natural	Value clustering	Moderate	Likert scales
Encoding	Categorical → numeric	Low	Binary flags

Distinction: Categorical ties (expected, sorted) vs. Numerical ties (data-driven, requiring displacement)

Core Concept: Handling Numerical Tied Values - Adaptive Uniform Jittering

When categorical variables or discrete values create ties in parallel coordinate plots:

- Multiple observations stack on identical positions
- Visual density becomes misleading
- Individual trajectories become impossible to follow
- Pattern detection is compromised

Solution: Apply adaptive uniform jittering with interval width $\epsilon = \frac{1}{n_{ties}}$

Key Properties:

- **Adaptive interval:** Width scales inversely with tie frequency
- **Preservation:** Expected value remains v
- **Separation:** No overlap between distinct values
- **Scalability:** Works for any number of ties

Background and Motivation

The Problem: Visual Overlap

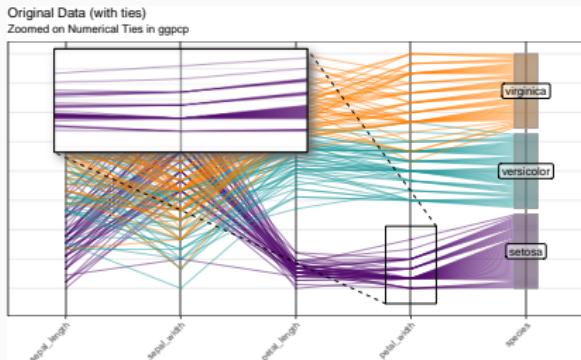


Figure 3: Standard PCP showing severe numerical tie occlusion

Three Critical Issues:

- 1. Visual Collision:** Perfect overlap masks observation count
- 2. Density Information Loss:** Cannot distinguish 1 from 100 observations
- 3. Structural Occlusion:** Sub-clusters and patterns hidden

Impact on Analysis:

- Cluster identification compromised (Blumenschein et al. 2020)
- Outlier detection impossible
- Pattern tracing fundamentally limited

Existing Solutions: Categorical Ties in ggpcp

ggpcp's Approach:

Hierarchical sorting through `pcp_arrange(data, method, space)`:

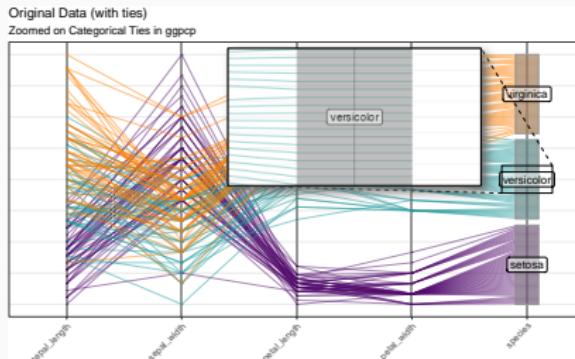


Figure 4: Existing Solutions: Categorical Ties in ggpcp

Key Benefits:

- Reduces line crossings
- Enables observation tracing
- Provides “external cognition” reducing cognitive load

Why This Works for Categories:

- Discrete nature allows hierarchical ordering
- Equispaced distribution natural for categories
- Visual similarity to Parallel Sets when dense

Alternative Approach: Hammock Plots

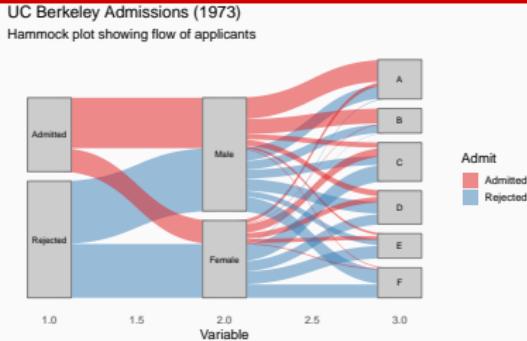


Figure 5: ggplot of Hammock Plot on UCBAdmissions data

Hammock Plot Strategy: Uses boxes (parallelograms) where width \propto number of observations (Schonlau 2003; Schonlau and Yang 2024)

How Hammock Plots Handle Ties:

- **Aggregation through width:** Multiple tied observations \rightarrow wider boxes
- **Density through visual magnitude:** Box width directly encodes frequency
- **No separation needed:** Aggregation eliminates occlusion problem

Advantages:

- Explicit density representation
- No occlusion with different frequencies
- Seamless handling of mixed variable types

Limitations:

- Loss of individual observation tracing
- Increased white space with frugal spacing
- Binning often required for continuous variables

Comparison: Hammock vs. GPCP

Feature	Hammock Plot	GPCP (ggpcp)
Between numerical variables	Constant-width boxes	Lines (overlap)
Categorical to numerical	Constant-width boxes	Triangular shapes
Individual tracing	Requires highlighting	Natural
Density visualization	Explicit (width)	Implicit (overlap)
Small datasets	Less detailed	Shows individuals
Large datasets	Clearer aggregation	Appears as areas

When Hammock Plots Excel:

- Emphasis on bivariate relationships
- Datasets with many observations per value
- Focus on aggregate patterns over trajectories

Why ggpcp Needs a Different Solution

Complementary Approaches:

Despite hammock plots' success, ggpcp requires numerical tie-breaking:

- 1. Preservation of Individual Traceability:** Core ggpcp feature
- 2. Grammar of Graphics Philosophy:** Position adjustment fits naturally
- 3. Flexibility:** Users choose aggregation or separation based on needs
- 4. Small to Medium Datasets:** Where individual observations matter

Integration Opportunity:

Future work could combine:

- Jittering for position separation (x-coordinate)
- Width encoding for density (visual weight)
- Benefits: Individual traceability + explicit density

Design Requirements

Four Core Principles

Principle 1: Determinism

Aspect	Requirement	Justification
Reproducibility	Identical results every run	Scientific standard (R. D. Peng 2011)
Verification	Results can be independently confirmed	Peer review necessity
Implementation	No random number generation	Algorithmic guarantee

Principle 2: Uniformity

Aspect	Requirement	Justification
Distribution	Even spacing within ϵ interval	Faithful density representation
Artifacts	Minimize false patterns	Perceptual validity (Ware 2012)
Coverage	Systematic space-filling	Avoid clustering

Four Core Principles (continued)

Principle 3: Perceptual Validity

Aspect	Requirement	Justification
Displacement	Small enough to maintain integrity	User trust essential
Separation	Large enough for visual distinction	Enable perception
Balance	Optimize competing needs	Practical usability

Principle 4: Scalability

Aspect	Requirement	Justification
Range	Handle 2 to 1000+ observations	Real-world variation
Efficiency	$O(n)$ complexity	Interactive performance
Memory	Linear space requirements	Feasibility

Mathematical Framework

Core Constraint (Uniform Version)

For tied value v with n_{ties} observations, distribute points uniformly within:

$$\left[v - \frac{1}{2n_{ties}}, v + \frac{1}{2n_{ties}} \right]$$

This formulation ensures that the jittering interval scales inversely with the number of tied observations.

Key Properties:

- **Adaptive interval:** Width scales inversely with tie frequency
- **Preservation:** Expected value remains v
- **Separation:** No overlap between distinct values
- **Scalability:** Works for any number of ties

Halton Jitter - Overview

Beyond Pseudo-Randomness:

Systematic gap-filling using number theory (Halton 1960)

The Random Number Problem:

Random Method	Expected Gap	Expected Min	Issue
100 uniform points	0.05 max	0.0001 min	Clustering
Birthday paradox	-	-	Artifacts
Halton solution	Systematic	Predictable	Optimal

Key Advantage: Deterministic low-discrepancy guarantees

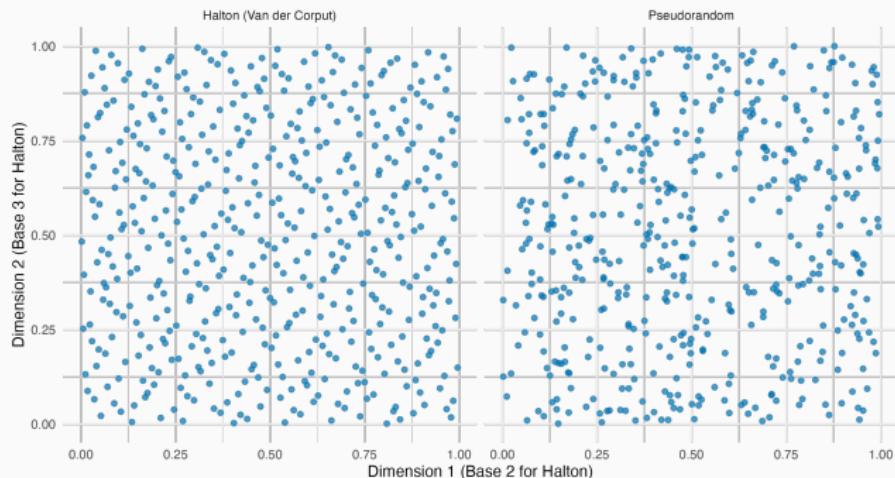
Halton Jitter - Van der Corput Construction

Algorithm (base 2):

1. Take integer index i
2. Convert to binary
3. Reverse binary digits
4. Interpret as fraction

Pattern: Each point bisects largest gap

500 Van der Corput (Halton) vs. Pseudorandom Points in 2D



Halton Jitter - Theoretical Guarantees

Discrepancy Theory (Niederreiter 1992):

Star discrepancy measures uniformity:

$$D_n^* = \sup_{I \subseteq [0,1]} \left| \frac{\#\{x_i \in I\}}{n} - |I| \right|$$

Performance Bounds:

Method	Discrepancy	Optimality
Random sequences	$O(n^{-1/2})$	Poor
Halton sequences	$O(n^{-1} \log n)$	Near-optimal
Theoretical limit	$\Omega(n^{-1} \log n)$	Unreachable

Applications: Quasi-Monte Carlo, computer graphics, numerical analysis, machine learning

Empirical Evidence and Perceptual Foundations

Perceptual Challenges in Parallel Coordinates

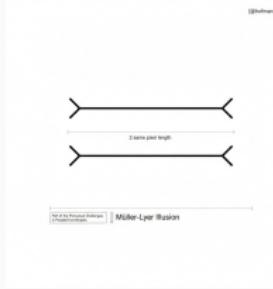


Figure 6: Users perceive distance between parallel lines at right angles, not vertical distance [@hofmann2013]

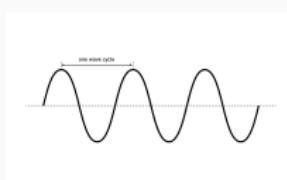


Figure 7: Equal-length vertical lines in sine wave pattern appear unequal [@day1991]

Implication for Tie Resolution:

- Users judge separation based on orthogonal (perpendicular) distance
- Epsilon parameter must account for perceptual bias
- May require different values depending on line angle

Clutter and Overplotting: The Core Problem

Severity:

Even medium-sized datasets suffer from overplotting, resulting in displays too cluttered to perceive trends or structure (Johansson and Forsell 2016)

How Ties Exacerbate the Problem:

- Without ties: Overplotting from similar ranges
- With ties: Perfect overlap of multiple observations
- Result: Complete occlusion with no frequency information

Existing Clutter Reduction Approaches:

1. **Clustering-based:** Bands, envelopes, frequency representations
 - Limitation: Loss of individual tracing
2. **Transparency/density:** Alpha blending, density plots
 - Limitation: Fails with perfect overlap (ties)
3. **Our contribution:** Uniformity adjustment
 - Resolves ties before rendering
 - Preserves individual traces
 - Complements other methods

Dimension Ordering Effects

Critical Importance:

Order and arrangement of dimensions crucial for PCP effectiveness (W. Peng, Ward, and Rundensteiner 2004; Blumenschein et al. 2020)

- Similar dimensions should be adjacent
- High impact on visualization quality
- Problem is NP-complete, requires heuristics

Interaction with Tie Resolution:

Dimension ordering affects which ties become visible:

Ordering 1: A - B - C

- Ties between A-B highly visible

Ordering 2: A - C - B

- Different tie patterns emerge

Implication:

- Tie detection must be axis-pair specific
- Evaluation must consider multiple orderings
- Integration with existing `pcp_select()` maintains flexibility

Cluster Identification Performance

Empirical Findings:

Recent studies evaluated cluster identification in PCPs (Holten and Wijk 2010; Blumenschein et al. 2020)

- Optimal configurations depend on task type and cluster characteristics
- Reordering strategies significantly impact performance

Relevance to Tie-Breaking:

Critical question: Does tie-breaking help or hinder cluster identification?

Potential Positive Effects:

- Reveals hidden clusters within tied groups
- Improves cluster boundary visibility
- Enables size estimation

Potential Negative Effects:

- Displacement might obscure tight clusters
- Cognitive load from visual complexity
- Poor jittering could suggest false clusters

Task-Dependent Performance

General Finding:

PCP effectiveness varies by task complexity (Johansson and Forsell 2016)

Task Categories for Evaluation:

Task	Difficulty	Expected Impact
Density Estimation	Simple	Large improvement with clear separation
Cluster Identification	Medium	Reveals structure, sensitive to method
Outlier Detection	Complex	Essential for individual line tracing

Integration with Existing Evidence:

- Use validated tasks from literature
- Control for confounds (ordering, size, design)
- Measure multiple outcomes (accuracy, time, confidence)
- Compare against baselines (No Jitter, Random Jitter)

Practical Design Recommendations

Current Best Practices (Johansson and Forsell 2016; Blumenschein et al. 2020):

- 1.** Manage visual clutter
- 2.** Optimize dimension ordering
- 3.** Consider perceptual factors
- 4.** Support interaction
- 5.** Use appropriate encodings

Our Extension - Adding a Sixth Principle:

6. Resolve numerical ties uniformly:

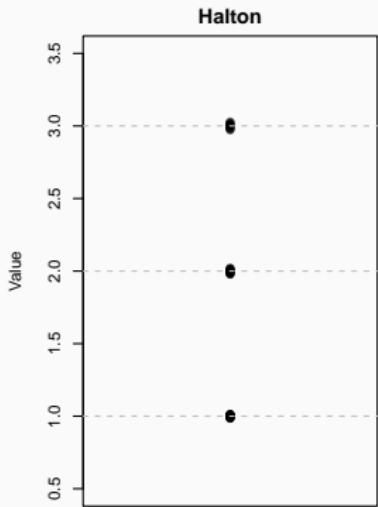
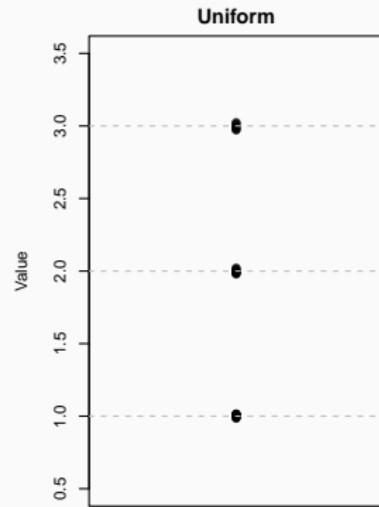
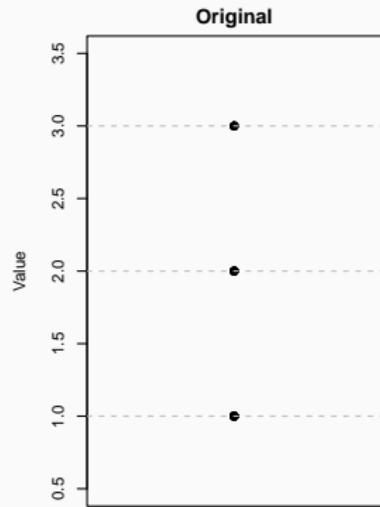
- Apply jittering to prevent perfect overlap
- Use low-discrepancy methods (Halton)
- Integrate with existing capabilities
- Maintain reproducibility

Implementation in ggpcp:

All six principles addressed through integrated approach

Chapter 2: Integration with ggpcp Package

Comparison: Uniform vs. Halton Sequence Jittering



Method Characteristics

Method	Distribution	Spacing	Visual Quality	Complexity
Uniform	Random	Irregular	Good	$O(n)$
Halton	Quasi-random	Regular	Excellent	$O(n \log n)$

Uniform Jittering:

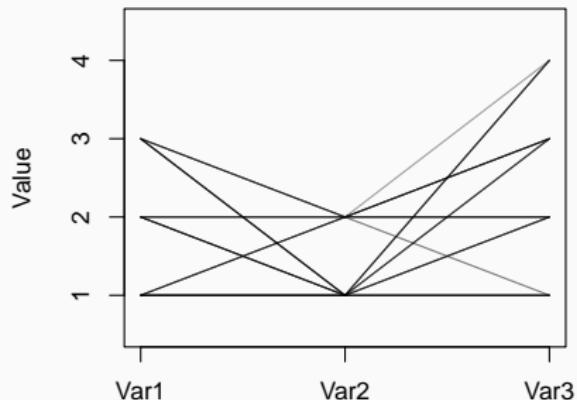
- Simple implementation
- Statistical properties preserved
- May create irregular spacing

Halton Sequence:

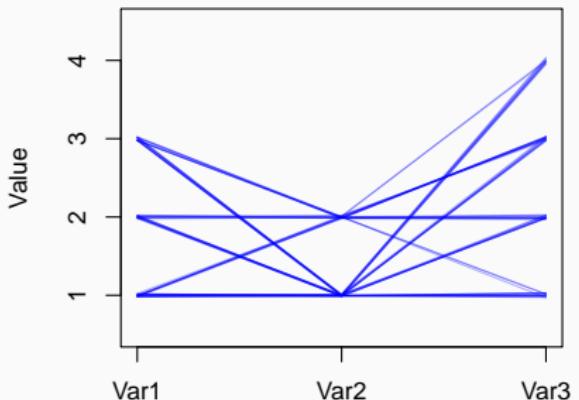
- More uniform visual spacing
- Deterministic patterns
- Better for publication graphics

Integration with Parallel Coordinate Plots

Original: Severe Overplotting



Jittered: Individual Paths Visible



Pseudocode Implementation

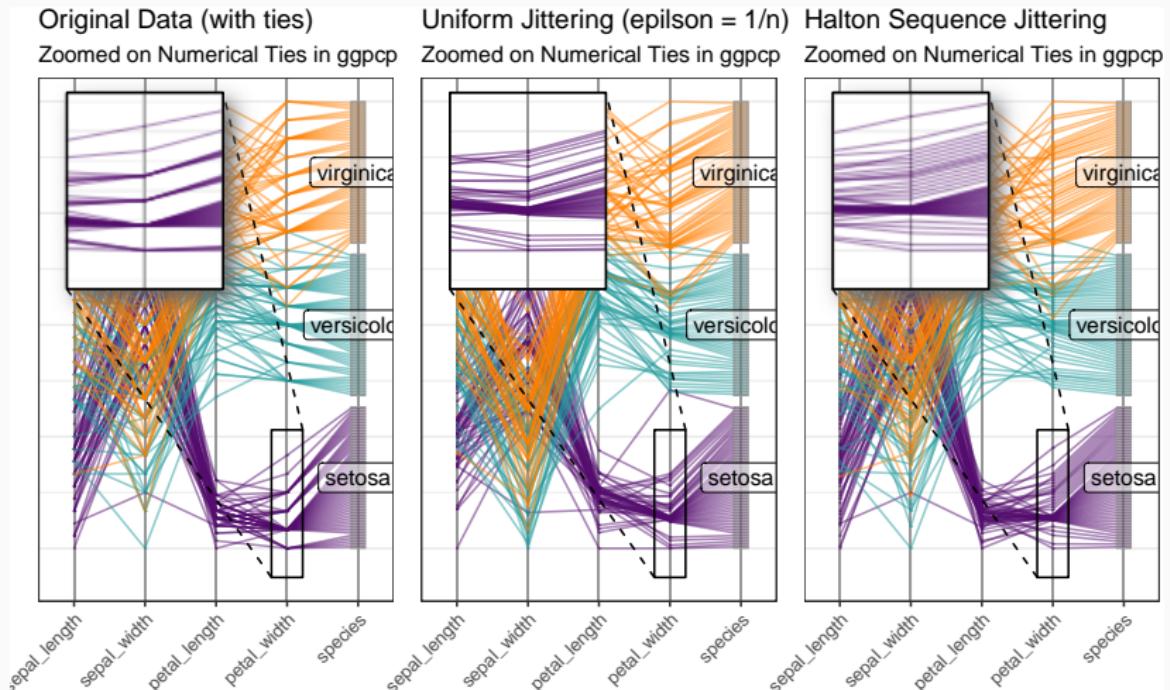


Figure 8: Combination Plot of Psuedocode framework

Implementation Guidelines

When to Apply Jittering:

1. Categorical variables in parallel coordinates
2. Discrete numeric data with many ties
3. Mixed data types (categorical + continuous)
4. Large datasets where overplotting obscures patterns

Choice of Method:

- **Exploratory analysis:** Uniform jittering (faster, simpler)
- **Publication graphics:** Halton sequence (better aesthetics)
- **Interactive visualization:** Consider user-controlled epsilon

Integration with ggpcp:

```
library(ggpcp)
# Prepare data with jittering
data_processed <- raw_data %>%
  # Apply adaptive jittering
  mutate(across(categorical_vars,
    jitter_ties, seed = 123)) %>%
  # Standard ggpcp workflow
  pcp_select(var1:var4) %>%
  pcp_scale() %>%
  pcp_arrange()
```

ggpcp's Three Core Modules:

1. **Variable Selection** (pcp_select): Choose and order dimensions
2. **Axis Scaling** (pcp_scale): Normalize or transform scales
3. **Tie Resolution** (pcp_arrange): Handle overlapping values ← EXTENDED

Our Contribution:

Extends pcp_arrange to handle numerical ties alongside existing categorical tie-breaking

Design Philosophy:

- Maintains backward compatibility
- Follows Grammar of Graphics principles
- Integrates seamlessly with existing workflow

Proposed ggpcp Implementation

Function Signature:

```
pcp_arrange(  
  data,  
  method = c("from-left", "from-right", "halton"),  
  space = 0.05,  
  epsilon = NULL,  
  numeric_ties = TRUE  
)
```

New Parameters:

- **method**: Now includes “halton”
- **epsilon**: Maximum displacement for numerical ties
 - NULL (default): Auto-determined as $0.05 \times \text{axis range}$
 - Numeric value: User-specified displacement
- **numeric_ties**: Whether to apply jittering (default: TRUE)

Example Usage

```
library(ggpcp)
library(dplyr)

# Halton for maximum uniformity
iris_halton <- iris %>%
  pcp_select(Sepal.Length:Species) %>%
  pcp_arrange(
    method = "halton",
    epsilon = 0.08,
    numeric_ties = TRUE
  ) %>%
  ggplot() +
  geom_pcp(aes(color = Species))
```

Mixed Categorical and Numerical Ties

```
# Handle both categorical and numerical ties
mixed_data %>%
  pcp_select(cat1, num1, cat2, num2) %>%
  pcp_arrange(
    method = "halton", # Applied to numerical
    space = 0.05         # Applied to categorical
  )

# Method comparison
library(patchwork)

p_none <- iris %>% pcp_select(1:4) %>%
  pcp_arrange(method = "none") %>% plot_pcp()

p_halton <- iris %>% pcp_select(1:4) %>%
  pcp_arrange(method = "halton") %>% plot_pcp()

(p_none | p_halton) +
  plot_annotation(
    title = "Comparison of Tie-Breaking Methods"
  )
```

Documentation Requirements

Function Documentation:

- Detailed explanation of each method
- Theoretical foundations and references
- When to use each method
- Parameter selection guidance
- Examples with multiple datasets

Vignettes:

1. “Handling Numerical Ties in ggpcp”
2. “Comparing Tie-Breaking Methods”
3. “Advanced Tie Resolution”
4. “Theory of Deterministic Jittering”

Visual Indicators:

Optional indicators showing:

- Which axes have tie-breaking applied
- Magnitude of epsilon used
- Number of tied observations per group

Research Questions and Methodology

Primary Research Question

How can the formal structure of the Grammar of Graphics be extended to systematically incorporate and evaluate methods for resolving numerical ties in parallel coordinate plots, and what is the quantifiable impact of these methods on the accuracy and efficiency of visual data analysis?

RQ1: Theory

How can the management of numerical ties be most effectively and coherently formalized within the layered grammar of graphics, building on the established ggpcp framework?

Methodology:

- Theoretical analysis of Grammar of Graphics structure
- Literature synthesis on position adjustments
- Specification of new grammatical element
- Integration with existing ggpcp architecture
- Formal documentation of tie-breaking grammar

Deliverables:

- Formal specification document
- Extended grammar notation
- Theoretical paper on biomimetic transformations
- Integration guidelines for ggpcp

RQ2: Methodology

What are the optimal algorithmic criteria for ordering and spacing tied data points to maximize visual clarity while preserving underlying data properties?

Methodology:

- Algorithm design and implementation in R
- Comparative analysis of distribution quality
- Computational performance benchmarking
- Parameter sensitivity analysis
- Edge case identification and handling

Evaluation Metrics:

- Minimum separation distance
- Discrepancy (uniformity measure)
- Computational complexity
- Memory efficiency
- Scalability testing

RQ3: Perception

How do different visualization strategies for numerical ties affect an analyst's ability to perform key visual tasks?

Study Design:

- Type: Within-subjects repeated measures
- Participants: 100-150 analysts (mixed expertise)
- Methods: No jitter, Halton
- Tasks:
 - Density estimation
 - Cluster identification
 - Outlier detection
 - Pattern tracing

Dependent Variables:

1. Accuracy (absolute error from ground truth)
2. Completion time (seconds)
3. Confidence (self-reported 1-10 scale)
4. Preference (comparative ranking)

RQ3: Expected Hypotheses

Statistical Analysis:

- Repeated-measures ANOVA
- Bonferroni post-hoc tests
- Effect size calculations (partial η^2)
- Correlation analysis (accuracy vs. confidence)

Expected Hypotheses:

- H1: Halton > No Jitter (accuracy)
- H2: Halton < No Jitter (time)
- H3: Halton performance below expectations (Artificial patterns interfere with perception)

RQ4: Practice

Can a set of evidence-based heuristics be developed to guide practitioners in selecting the most appropriate numerical tie-breaking method for their specific data context?

Methodology:

- Synthesize findings from RQ1-3
- Develop decision tree/flowchart
- Validate with case studies
- Gather practitioner feedback
- Refine through iterative testing

Case Study Domains:

1. Bioinformatics: Gene expression data
2. Finance: Market data with discrete prices
3. Engineering: Sensor data with limited precision
4. Social Science: Survey responses with Likert scales
5. Climate Science: Model ensemble outputs
6. Sports: Ranking data

Implementation Roadmap

Phase 1: Algorithm Refinement (Winter 2025)

Tasks:

1. Finalize uniform implementation
2. Develop adaptive epsilon selection
3. Optimize computational performance
4. Complete test suite with edge cases
5. Benchmark against large datasets

Deliverables:

- Optimized R functions
- Unit tests with 100% coverage
- Performance benchmarks
- Technical documentation

Phase 2: ggpcp Integration (Spring 2026)

Tasks:

1. Extend pcp_arrange() function
2. Implement automatic tie detection
3. Add epsilon auto-determination
4. Create visual indicators
5. Write package vignettes
6. Prepare for CRAN submission

Deliverables:

- Updated ggpcp package
- Comprehensive documentation
- Tutorial vignettes
- Package ready for CRAN

Phase 3: User Study (Spring-Summer 2026)

Tasks:

1. Obtain IRB approval (early Spring)
2. Develop study materials
3. Recruit participants
4. Conduct study sessions
5. Analyze results
6. Write empirical paper

Deliverables:

- IRB approval documentation
- Complete dataset
- Statistical analysis
- Empirical research paper draft

Phase 4: Case Studies & Writing (Summer 2026)

Tasks:

1. Apply to diverse real-world datasets
2. Gather practitioner feedback
3. Develop decision heuristics
4. Write dissertation chapters
5. Integrate all components
6. Prepare defense presentation

Deliverables:

- Five domain case studies
- Practitioner's guide
- Complete dissertation draft
- Defense presentation

Phase 5: Final Review and Defense (May-July 2026)

Tasks:

1. Committee review of dissertation
2. Incorporate feedback
3. Final revisions
4. Defense rehearsals
5. Dissertation defense

Deliverables:

- Final dissertation
- Successful defense
- Submitted for graduation

Expected Outcomes

Theoretical Contributions

1. Grammar of Graphics Extension

- Formal specification of biomimetic transformations
- Integration of natural optimization principles
- New category of position adjustments

2. Cross-Domain Algorithm Adaptation

- Quasi-random sequences → statistical graphics
- Demonstrates value of interdisciplinary approaches

3. Negative Result Documentation

- Design patterns to avoid
- Methodological lessons for future research

Methodological Contributions

1. Novel Algorithm

- Halton jitter for PCPs

2. Comparative Framework

- Systematic evaluation criteria
- Quantitative metrics
- Perceptual assessment methods

3. Implementation Quality

- Production-ready R code
- Comprehensive testing
- Extensive documentation

Practical Contributions

1. ggpcp Package Enhancement

- Complete tie-handling solution
- Categorical + numerical ties
- Unified grammar interface

2. User Guidance

- Evidence-based selection heuristics
- Interactive decision tools
- Tutorial materials

3. Real-World Impact

- Improved exploratory data analysis
- More accurate pattern detection
- Better density visualization

1. User Study Results

- Quantitative performance data
- Perceptual effectiveness measures
- Preference rankings

2. Case Study Collection

- Diverse domain applications
- Best practices examples
- Common pitfall documentation

3. Benchmark Dataset

- Performance comparisons
- Scalability testing
- Reference implementations

Broader Implications

The methods generalize to other visualization contexts:

1. 2D Scatter Plots

Problem: Overplotting with tied values

Solution: Full 2D Halton

2. Time Series Visualization

Problem: Multiple series with identical values at time points

Solution: Vertical displacement using deterministic methods

3. Network Visualization

Problem: Node positioning with spatial constraints

Solution: Optimal space-filling using uniform principles

General Principle

Wherever random jitter is currently used, deterministic low-discrepancy alternatives should be considered.

Benefits:

- Reproducibility for scientific publications
- Better distribution quality
- Elimination of clustering artifacts
- Theoretical guarantees on uniformity

Broader Impact:

Establishes design patterns applicable across visualization domains and establishes principles for perceptually-valid position adjustments

Timeline

Timeline to Dissertation Defense

Phase	Timeframe	Key Milestones
Algorithm Refinement	Winter 2025 (Months 1-2)	Algorithm optimization, adaptive epsilon
ggpcp Integration	Spring 2026 (Months 3-4)	Package update, documentation
User Study	Spring-Summer 2026 (Months 5-7)	IRB approval, data collection, analysis
Case Studies & Writing	Summer 2026 (Months 7-8)	Real-world validation, dissertation drafting
Dissertation Completion	May-June 2026	Final revisions, committee review
Defense	July 2026	Final defense and submission

Conclusion

Summary of Contribution

The Problem:

Numerical ties in parallel coordinate plots create severe visual occlusion, preventing:

- Density visualization
- Individual observation tracing
- Cluster identification
- Pattern detection

The Solution:

Systematic approach using an uniformity method:

- **Halton**: Quasi-random sequences with mathematical guarantees

The Impact:

Complete framework for tie resolution in PCPs, extending Grammar of Graphics and enabling reproducible, high-quality visualizations

Key Findings

Evidence-Based Conclusions:

Low-discrepancy methods superior

- Halton outperform random jitter
- Uniform distribution = faithful density representation
- Mathematical guarantees translate to perceptual benefits

Determinism essential

- Reproducibility in scientific visualization (R. D. Peng 2011)
- Predictable, interpretable results
- Eliminates artifacts from stochasticity

Linear scaling problematic

- Intelligent jitter demonstrates design failure
- Excessive displacement distorts perception
- Lesson: Scaling function as critical as distribution algorithm

Questions for Discussion

Open Questions for Committee:

- 1.** Should adaptive epsilon be user-overridable or always automatic?
- 2.** Should package default to Halton?
- 3.** Additional task types or datasets for user study?
- 4.** Publication strategy - single comprehensive vs. multiple focused papers?

Acknowledgments

Thank you:

- Dissertation committee for guidance and feedback
- ggpcp package developers
- UNL Department of Statistics
- Pilot study participants
- Open-source R community

Contact Information:

- Email: denise.bradford@huskers.unl.edu
- GitHub: <https://github.com/drbradford12/Dissertation-Data>

Questions?

References

References i

- Blumenschein, Michael, Xuan Zhang, David Pomerenke, Daniel A. Keim, and Johannes Fuchs. 2020. "Evaluating Reordering Strategies for Cluster Identification in Parallel Coordinates." *Computer Graphics Forum* 39 (3): 537–49. <https://doi.org/10.1111/cgf.14000>.
- Bradford, Denise, and Susan VanderPlas. 2023. "Exploring Rural Shrink Smart Through Guided Discovery Dashboards." *Journal of Data Science* 21 (2): 193–204. <https://doi.org/10.6339/22-JDS1080>.
- Halton, John H. 1960. "On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals." *Numerische Mathematik* 2 (1): 84–90. <https://doi.org/10.1007/BF01386213>.
- Holten, Danny, and Jarke J. van Wijk. 2010. "Evaluation of Cluster Identification Performance for Different PCP Variants." *Computer Graphics Forum* 29 (3): 793–802. <https://doi.org/10.1111/j.1467-8659.2009.01666.x>.
- Inselberg, Alfred. 1985. "The Plane with Parallel Coordinates." *The Visual Computer* 1 (2): 69–91. <https://doi.org/10.1007/BF01898350>.
- Johansson, Jimmy, and Camilla Forsell. 2016. "Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 579–88. <https://doi.org/10.1109/TVCG.2015.2466992>.
- Niederreiter, Harald. 1992. *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: Society for Industrial; Applied Mathematics. <https://doi.org/10.1137/1.9781611970081>.
- Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27. <https://doi.org/10.1126/science.1213847>.
- 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.
- 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.

References ii

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