Understanding Visual Clustering Factors in Scatterplots: An interactive study

Abhigyan Sharma
12340050

Abstract

Scatterplots are essential tools in data visualization, enabling users to discern relationships, detect outliers, and identify clusters in multidimensional data. While their basic structure is simple, the efficacy of scatterplots heavily depends on how visual elements are configured. This research investigates how human perception of cluster patterns in scatterplots is influenced by visual design parameters such as point density, opacity, color contrast, and point size. Through systematic experiments involving 18 plot variations rated by users, we analyze how preferences shift with different settings. The insights from this work aim to provide best practices for scatterplot design and foster intelligent visualization systems that adapt to user perception.

1 Introduction

In the realm of data analysis, visual representation is not merely aesthetic; it fundamentally affects comprehension and decision-making. Scatterplots are a favored visualization method because they succinctly portray data distributions and relationships in two or more dimensions. However, the capacity of a scatterplot to communicate information effectively is contingent on numerous design choices, especially in contexts where multiple clusters exist.

A key challenge arises from the fact that human perception does not always align with algorithmic clustering. While metrics like silhouette score quantify structural separation in data, visual interpretability involves perceptual phenomena such as contrast, saturation, and spatial arrangement. This discrepancy creates a gap between machine-recognized clusters and human-identified patterns.

This study seeks to bridge that gap by conducting controlled experiments that quantify how specific visual properties affect the interpretability of clusters in scatterplots. Our goal is to develop an empirically backed framework for optimal scatterplot design based on human perceptual feedback.

2 Problem Motivation

The problem begins with a simple observation: scatterplots that are statistically similar can have vastly different levels of human readability depending on their visual configuration. This discrepancy can mislead users, especially in exploratory data analysis, where intuition often guides hypothesis formation.

Several critical challenges emerge:

- **Visual Clutter**: High point density may obscure natural clusters, particularly when combined with high point size or opacity.
- **Perceptual Thresholds**: Balancing opacity and color contrast is necessary to avoid visual confusion.
- **Interpretive Bias**: User experience and visual acuity can alter perception significantly.

Addressing these challenges requires structured analysis of user preferences across systematically varied scatterplot configurations.

3 Literature Review and Related Work

Numerous studies have explored perceptual and algorithmic evaluations of scatterplots. Key works include:

- **ClustMe (Abbas et al., 2019)** – Automated ranking using image processing.
- **SEPMe (Aupetit & Sedlmair, 2016)** – Quantification of visual separation without subjective feedback.
- **Perceptual Studies (Anobile et al., 2016)** – Investigating fundamental perceptual mechanisms.
- **Interactive Exploration (Alexander et al., 2014)** – Emphasizing the value of user-driven tools.

Our work integrates subjective human responses with these approaches to evaluate perceptual clarity directly.

4 Data Collection Strategy

We created 18 scatterplots by varying:

- **Point Density (PD)**: {10, 50, 100, 200, 500}
- **Color Contrast (CC)**: {10, 20, 30, 50, 100, 200, 500}

- **Opacity (O)**: {0.1, 0.3, 0.5, 0.7, 0.8, 1.0}
- **Point Size (PS)**: {1, 10, 20, 30, 50}

Participants rated each plot from 1 to 10 based on cluster clarity. Data was stored in CSV format.

5 Feature Engineering and Modeling

The dataset included: Plot ID, PD, CC, O, PS, Rating.

Additional features:

- **Interaction Terms**: PD×O, CC×PS, PS/O
- **Binned Ranges**: Low/medium/high categories
- **Visual Noise Index**: Composite measure from PD, O, PS

Models used:

- Linear Regression
- Decision Trees
- User Behavior Clustering

6 Visualizations and Analysis

Visual techniques included:

- **Heatmaps** – Ratings vs. two-factor combinations
- **Violin plots** – Rating distributions per factor level
- **Pairwise scatter plots** – Visual factor vs. rating
- **Line graphs** – Trend analysis

7 Observations and Empirical Results

7.1 Opacity

- Optimal range: 0.3–0.7
- Below 0.2: points too faint
- Above 0.8: overplotting reduces clarity

7.2 Point Size

- Small/medium sizes (1–10) improve readability
- Large sizes (20+) decrease clarity, especially with high PD

7.3 Point Density

- Low PD: readable but sparse
- High PD: requires high contrast and low opacity to remain interpretable

7.4 Color Contrast

- High CC (200–500): best ratings
- Low CC (10–20): poor clarity

7.5 Combined Insights

Multifactor tuning is essential:

- Low opacity + high CC can balance high PD
- Small PS mitigates overlap in dense plots

8 Insights and Contributions

Key contributions:

- Empirical perceptual threshold validation
- Subjective clarity tied to identifiable design ranges
- Proposal of the **Perceptual Clarity Index (PCI)**
- Framework for human-centric scatterplot evaluation

9 Additional Reflections and Ideas

1. Colorblind-friendly alternatives (patterns, shapes)
2. Interactive plot adjustments (opacity sliders, size toggles)
3. ML-based configuration recommender
4. Adaptive visualization based on data shape
5. Library plugin for matplotlib/seaborn

10 Future Work

- Broaden participant base
- Integrate eye-tracking to quantify attention
- Evaluate plots in live analysis environments
- Build interactive dashboard to aid design

11 Conclusion

This comprehensive study illuminates the nuanced interplay between visual parameters in scatterplots and their impact on cluster interpretability. By collecting and analyzing user ratings, we've provided empirically grounded insights into optimal design choices. Our results advocate for user-centric visualization design, emphasize perceptual clarity, and lay groundwork for future tools that assist analysts in crafting effective scatterplots.

References

1. Abbas, M. M., et al. *ClustMe: A visual quality measure for ranking monochrome scatterplots based on cluster patterns.* Computer Graphics Forum, 38(3):225–236, 2019.
2. Alexander, E., et al. *Serendip: Topic model-driven visual exploration of text corpora.* IEEE VAST, 2014.
3. Amar, R., et al. *Low-level components of analytic activity in information visualization.* IEEE InfoVis, 2005.
4. Anobile, G., et al. *Number as a primary perceptual attribute: A review.* Perception, 45(1–2):5–31, 2016.
5. Aupetit, M., and Sedlmair, M. *SEPME: 2002 new visual separation measures.* IEEE PacificVis, 2016.
6. Bertini, E., and Santucci, G. *Give chance a chance: Modeling density to enhance scatter plot quality.* Information Visualization, 2006.
7. Best, L. A., et al. *Perceiving relationships: A physiological examination of the perception of scatterplots.* Theory and Application of Diagrams, 2006.