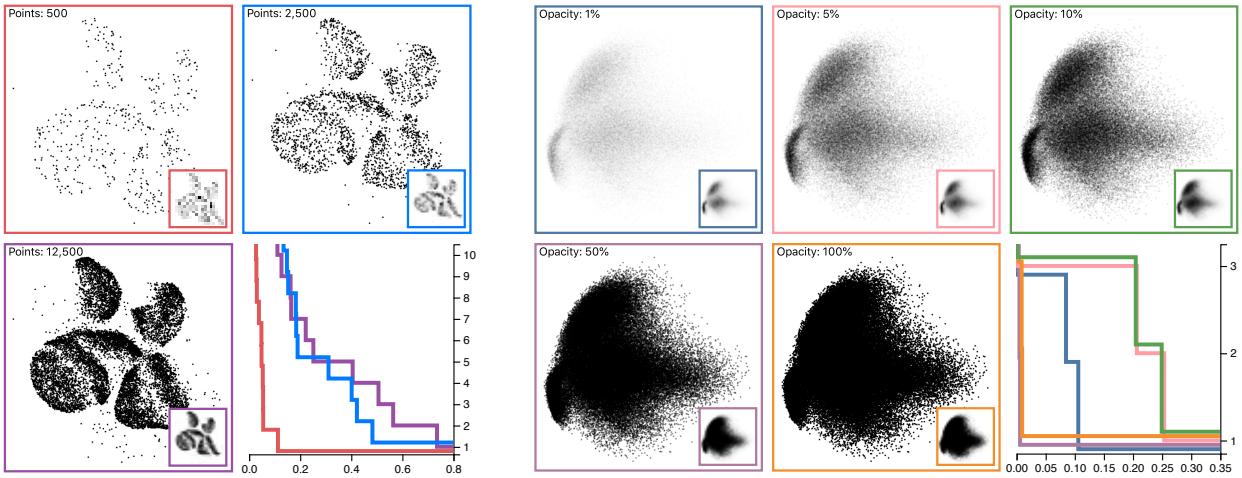
Modeling the Influence of Visual Density on Cluster Perception in Scatterplots Using Topology

Ghulam Jilani Quadri and Paul Rosen



(a) Varying the Number of Points

(b) Varying the Opacity of Points

Fig. 1. Demonstration of our density-based model being used to evaluate how differentiable clusters are in scatterplots that vary (a) the number of points shown (500, 2500, and 12500) or (b) the opacity of points (1%, 5%, 10%, 50%, and 100%). The threshold plots (lower right) show how easy it is to visually identify (horizontally, bigger is better) a certain number of clusters (vertically). For example, when varying the opacity (b), the threshold plot shows that 3 clusters are most clearly visible in the pink (5% opacity) and green (10% opacity) scatterplots, significantly less visible in the blue (1% opacity) scatterplot, and not visible at all in the purple (50% opacity) or orange (100% opacity) scatterplots. Designers can use these threshold plots to select the visual encodings that maximize the clarity of data.

Abstract— Scatterplots are used for a variety of visual analytics tasks, including cluster identification, and the visual encodings used on a scatterplot play a deciding role on the level of visual separation of clusters. For visualization designers, optimizing the visual encodings is crucial to maximizing the clarity of data. This requires accurately modeling human perception of cluster separation, which remains challenging. We present a multi-stage user study focusing on four factors—*distribution size of clusters*, *number of points*, *size of points*, and *opacity of points*—that influence cluster identification in scatterplots. From these parameters, we have constructed two models, a distance-based model, and a density-based model, using the merge tree data structure from Topological Data Analysis. Our analysis demonstrates that these factors play an important role in the number of clusters perceived, and it verifies that the distance-based and density-based models can reasonably estimate the number of clusters a user observes. Finally, we demonstrate how these models can be used to optimize visual encodings on real-world data.

Index Terms—Scatterplot, clustering, perception, empirical evaluation, visual encoding, crowdsourcing, topological data analysis

1 Introduction

Scatterplots are commonly used to reveal several types of relationships between quantitative variables [37]. Numerous perceptual studies have evaluated the effectiveness of scatterplots in low-level tasks that include assessing trends [22, 62], measuring correlation [7, 42, 66], and average and relative mean judgments [38]. Clustering, in particular, is an aggregate-level task [53, 61, 70] that has been utilized in a variety of applications, e.g., weather forecasting, text analysis, and large-scale data analysis [52, 78, 83, 87]. Clustering occurs when patterns in the data form distinct groups [3, 71]. However, at its core, clustering is an ill-posed problem, as the "correct" clustering depends upon multiple factors [23, 35].

When considering clustering in scatterplots, several factors play a role in how they are perceived. Data aspects, such as the data distribution size/type and the number of data points, can influence the visual presentation. On the other hand, visual encoding properties, such as mark type, size, and opacity, have the potential to influence perceptual judgments [15]. What is not well understood is how these various factors, *many of which are under the control of the visualization designer*, influence the perception of clusters. The presentation of data is particularly important when considering that a biased representation of the data may provide an inaccurate summary, leading to invalid conclusions [39, 46, 50, 77].

In this paper, we explore the multi-factor judgments used in identifying clusters in scatterplots through a crowdsourced user study. Based upon this study, we develop 2 models for the perception of clusters in scatterplots, using a data structure from Topological Data Analysis, called the *merge tree* [84]. We validate the models on a variety of variables—*the number of points*, *cluster distribution size*, *size of data*

- *Ghulam Jilani Quadri is with the University of South Florida. E-mail: ghulamjilani@usf.edu.*
- *Paul Rosen is with the University of South Florida. E-mail: prosen@usf.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

points, and *opacity of data points*—to verify the accuracy of the models and analyze their effects. Our results show that the perception of the number of clusters does indeed depend upon all 4 factors. Moreover, we show that the merge tree-based models do match an average user's perception of the clusters in a given scatterplot.

Finally, we demonstrate how the models can be used to optimize visualization designs. While some variables, such as distribution size, are difficult to control in a visualization, designers can use our models and findings as a guideline to balance the design factors that they do have control over—the number of points shown (e.g., via subsampling [12, 45]), data point size [78, 80], or opacity [56, 60]—to optimize the saliency of the clusters in a visualization. A demo of the approach can be viewed at <https://usfdatavisualization.github.io/TopoClusterPerceptionDemo>.

2 Prior Work

We provide brief coverage of clustering in scatterplots and perception of the visual factors evaluated in our study.

2.1 Clustering in Scatterplots

Clustering plays an important role in exploring and understanding many types of data [70, 71]. A design factor survey defined clustering as a high-level data characterization—the ability to identify groups of similar items [71]. Amar et al. presented a set of tasks for visual analytics that defined clusters as having "similar attribute values in a given set of data" [3].

Taxonomies of Clustering Factors Identifying clusters is directly influenced by the perception of cluster separation, and much of our understanding has come from studying dimension reduction (DR) techniques. Lewis et al. compared the effectiveness of DR techniques using human judgments and concluded that T-SNE performs better than other commonly used methods when expecting clusters in the data [49]. Etemadpour et al. showed, however, the performance of DR techniques also depends on data characteristics [33], e.g., the separability of clusters, and later created a user-centric taxonomy of visual tasks related to clustering in DR techniques [32]. A taxonomy of visual cluster separation in scatterplots used a qualitative evaluation to identify 4 important factors—scale, point distance, shape, and position [73]. The taxonomy gives a context to our visual factor selection. Sedlmair and Aupetit later evaluated 15 class separation measures for assessing the quality of DR using human input for building a machine learning framework [72] and later extended the framework to include an even greater number of measures [5].

Perceptual Models of Clustering Several works have considered how to model the perception of clusters. For example, a recent study that used eye-tracking to analyze user perception in cluster identification, highlighted the role of Gestalt principles, especially proximity and closure [34]. Matute et al. provided a method to quantify and represent scatterplots through skeleton-based descriptors that measured scatterplot similarity [57]. However, their approach does not consider visual encodings in the evaluation. ScatterNet, a deep learning model, captures perceptual similarities between scatterplots to emulate human clustering decisions but lacks explainability in the choices [54]. The scagnostics technique focused on identifying the patterns in scatterplots, including clusters [20]. However, a study by Pandey et al. showed that they do not reliably reproduce human judgments [65]. Recently, ClustMe used visual quality measures to model human judgments to rank scatterplots [1]. ClustMe performed well in reproducing human decisions for clustering patterns. In contrast, we are studying the extent to which various factors influence the perception of clusters and building explainable models of how humans perceive cluster separation using the merge tree data structure.

Clustering in Non-Scatterplot Contexts Clustering has been studied in other types of visualization, such as text [2], maps [52, 83], and bubble charts [78]. A task-based evaluation found that on small data, bar and pie charts outperformed tables, scatterplots, and line charts in clustering tasks [69]. The performance in cluster perception in pie charts is traced back to its effectiveness in facilitating proportional judgments through a part-whole relationship [26, 74]. Similarly, we hypothesize that the relative distance between clusters and the relative density of the image influence cluster identification.

2.2 Factor Selection on Scatterplots

Several prior perceptual studies have demonstrated the effect of visual encodings on analysis tasks [15, 39, 77]. A variety of factors influence group or separation perception [88], including color, size, shape [73], orientation [16], texture [4], opacity [60], density [86], motion and animation [11, 31, 81], chart size [44], and others. Other studies have demonstrated a perceptual effect in scatterplots when changing factors in the data, including data distribution types, number of points, the proximity of concentrations of points, data point opacity, and relative density [14, 18, 38, 39, 46, 68, 77]. Overdraw in scatterplots, in particular, has been addressed with a variety of techniques, e.g., splatterplot [59], recursive sampling [12], set cover optimization [45], feature-preserving visual abstraction [10], or by applying various clutter reduction techniques [29], e.g., sampling [21, 27, 28, 85] or changing opacity [56].

From this collection of possible factors, *we focus our study specifically on the factors that most influence visual density*, including the distribution of and distance between concentrations of points, the number and size of data points, and data point opacity in the visualization.

Point Distribution Several prior studies have investigated the influence of the distribution of data points on cluster perception. An early study of 8 participants on 24 homogeneous dot patterns studied the impact of varying densities and gaps between 2 square-shaped clusters [63]. Sadahiro later developed a mathematical model to represent cluster perception in point distributions based on 3 factors—proximity, concentration, and density change—and suggested perception is significantly influenced by the concentration and density change [68]. Similarly, the scagnostics density property identifies concentrations of points directly influenced by the distribution of points [86].

Number of Data Points Sadahiro also showed that the higher the number of points in a given area, the higher the chances are that they would be perceived as a cluster, due to increased density [68]. Gleicher et al.'s empirical study asked participants to compare and identify average values in multi-class scatterplots [38]. It demonstrated that judgments are improved with a higher number of points.

Size of Data Points The size of symbols is an important factor in visual aggregation tasks in scatterplots [78]. As the size of data points increases, so does the density, which directly influences cluster perception [68]. Symbol size also has a direct influence on discriminability in certain tasks [50], e.g., in color perception tasks [76]. Szafir's study on color-difference perception found that perceived color difference varies by the size of marks [77]. Size also influences search task effectiveness. Gramazio et al.'s study on target search demonstrated that while the quantity of data points has little effect on searching for a target, increasing symbol size reduces search time in a display of random points [39].

Opacity of Data Points As the number of data points increases, scatterplots suffer from overplotting, which obscures the data distribution. Reducing mark opacity can alleviate overplotting to aid various visual analytics tasks [70], e.g., spike detection in dot plots [18]. Furthermore, different opacity levels aid in different visual tasks—while low opacity benefits density estimation for large data, it also makes locating outliers more difficult [60]. Matejka et al. defined an opacity scaling model for scatterplots that is based on the data distribution and crowdsourced responses to opacity scaling tasks [56]. Still, their study did not evaluate how a scatterplot design based on data symbol opacity can affect user performance on visual analysis tasks. Somewhat related to opacity is luminance, which can be modeled using extreme end lightness [51], creating a popout effect [41].

3 Study Methodology

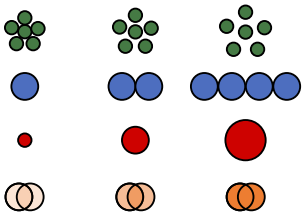
We investigate how visual factors affect subject responses in the task of counting the number of clusters in a scatterplot. From this, we build and analyze two models to estimate the number of clusters an average

user would perceive. One model is based on the separation distance between distributions, and the other uses the visual density of points.

3.1 Factors

Data are presented as point marks (i.e., circles ●) on the scatterplots and groups of similar objects form clusters. Based on our review of prior work, we chose to use a normal distribution to generate clusters, and we selected the following experimental factors:

1. Distribution size (S)
2. Number of data points (N)
3. Size of data points (P)
4. Data point opacity (O)



3.2 Experiments Setup

We designed our experimental study in 3 stages: (1) a preliminary experiment to calibrate the experimental factors; (2) a crowdsourced Amazon's Mechanical Turk (AMT) experiment to validate our models; and (3) a follow-up AMT study to elaborate upon 1 of our models.

3.3 Data Generation

Datasets are synthesized using 5 input parameters (see Fig. 2): stimuli dimensions ($[X \times Y]$ pixels); number of clusters (C); distribution size, i.e., standard deviation (S pixels); number of points (N); and signal-to-noise ratio (SNR). First, C cluster centers are randomly placed within a "safe zone" defined as 1 standard deviation from the stimuli (image) border, in other words, $x \in [S, X-S]$ and $y \in [S, Y-S]$. Each cluster is assigned an equal share of the available points (N/C). Points are randomly placed around their cluster center using a normal distribution with a standard deviation of S pixels. Points outside of the image dimensions are discarded without replacement. Next, an additional N/SNR points representing noise are placed randomly using a uniform distribution across the image dimensions. Finally, to generate images, 2 more input parameters are used: point size (P pixels) and point opacity (O). The points are drawn as filled circles of P area with O opacity. Example stimuli are shown in Fig. 3.

In all experiments some inputs were kept constant:

- Stimuli dimensions ($[X \times Y]$): $[550_{px} \times 550_{px}]$ — The vertical size was selected such that the image would fit on the majority of desktop monitors without scrolling [75]. The horizontal resolution was selected to match, avoiding any directional bias.
- Signal-to-noise ratio (SNR): $10:1$ — We manually optimized the SNR by looking for a high level of noise that would not overwhelm the clusters. We ended at $10:1$, making the maximum total number of data points in any given dataset $N + 0.1 \cdot N$.

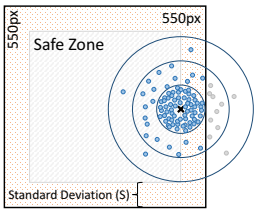


Fig. 2. Illustration of the data generation. All plots are 550×550. Cluster centers are placed within a "safe zone", 1 standard deviation away from the boundary. Points are sampled from a normal distribution (blue), and points outside of the image are discarded without replacement (gray).

4 Preliminary Experiment

We performed a preliminary user study to test initial hypotheses and calibrate parameters for the larger AMT experiment. Based on our observation and study of prior work, we drafted the following hypotheses:

[H1] The distribution size of clusters affects the accuracy in cluster count identification in scatterplots.

[H2] The number of data points affects the accuracy in cluster count identification in scatterplots.

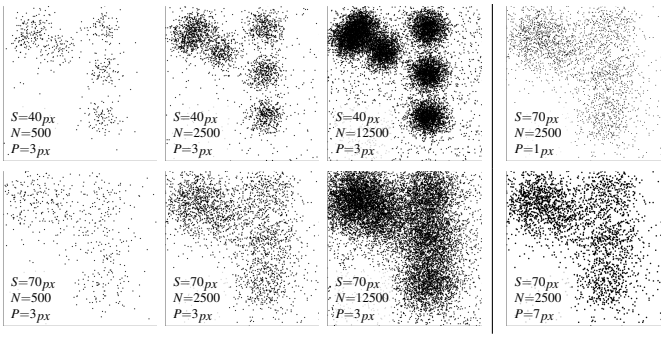


Fig. 3. Example stimuli with the same cluster centers ($C = 5$), but varying distribution size (S), number of points (N), and size of points (P).

4.1 Properties and Data Generation

As aligned with previous empirical studies, e.g., [69], we selected parameter values to maintain a reasonable level of difficulty. We designed the task such that the response time for a single stimulus would be 5 to 20 seconds. We selected the following experimental factors:

- Number of clusters (C): $\{4-12\}$ — The number of clusters was selected using trial-and-error to avoid tasks that were too easy (i.e., trivial to count) or too difficult (i.e., larger number or sparse clusters).
- Data point size/area (P): $\{20_{px}\}$ — Experimental calibration was not needed for point size, as reasonable values could be determined analytically. Therefore, the point size was fixed in order to calibrate other factors.
- Number of data points (N): $\{1000, 5000, 10000\}$
- Distribution size (S): $\{20_{px}, 35_{px}, 50_{px}, 65_{px}, 90_{px}\}$ — N and S were the main factors to test/calibrate. The value ranges were selected using our observation of sample stimuli and judgment of factors from prior work, considering sufficient range, minimum and maximum values, and the number of experimental conditions that could be reasonably tested.
- Data point opacity (O): $\{100\%\}$ — Points were fully opaque.

The dependent variable we tested was:

- User-selected number of clusters (U): $[1-15]$

Dataset generation for the preliminary experiment was done in the following manner—for every combination of S and N, 500 stimuli (i.e., images) were generated with a random number of clusters, C. Other parameters were fixed as described, leading to a pool of $|S| \times |N| \times 500 = 7500$ stimuli.

4.2 Study Procedure

We developed a webpage for the experiments, where each participant was shown 50 images from the pool of 7500, one at a time, and asked the number of clusters they could see. Answers were recorded using a drop-down box with options $1-15$. The maximum allocated time for each task was 20 seconds. At the expiration of time, the page was automatically advanced. To mitigate any effects or bias, we placed a blank screen between every 2 tasks [43]. At the beginning of the experiment, we included a brief introduction to clustering and 3 training tasks for each participant, which were similar to the study tasks that followed. The experiment was expected to last 20-30 minutes, including demographic details and training tasks.

We recruited 30 participants from the College of Engineering at the University of South Florida for the IRB approved study. Participant ages ranged from 18-28 (μ_{age}=23), with 24 males and 6 females. No compensation was provided. In total, 50 trials $\times$ 30 participants = 1500 responses were collected. While performing data quality checks on the responses, we found discrepancies—participants responding to

stimuli in less than 1 second or those with responses of 1 cluster to all stimuli—in 4 participants results and removed them from analysis, leaving 1300 responses (26×50). We further identified and removed 161 stimuli that had been reused from the pool only keeping the first occurrence[1], leaving 1139 responses for analysis.

4.3 Analysis and Result

To measure accuracy for a given scatterplot, τ, we use the *differential*: $\mathbb{D}(\tau) = U_\tau - C_\tau$, where U_τ is the user response and C_τ is the number of clusters in the data. We analyzed the *differential* against the independent factors *distribution size* (S) and the *number of data points* (N) using a 2-way ANOVA test. We also calculated partial eta-squared (η^2). We observed that S and N have a significant effect on the accuracy in identifying the number of clusters, ($F_S(4,1130) = 48.57, p < 0.01, \eta^2 = 0.12$) and ($F_N(2,1130) = 8.29, p < 0.01, \eta^2 = 0.02$), respectively. **These results confirm [H1] and [H2]**.

Although we found a significant effect, user accuracy had a low average $\mu_\mathbb{D} = -3.27$ and a high standard deviation $\sigma_\mathbb{D} = 2.66$ (accurate predictions would have an average of 0 with a small standard deviation). Fig. 4 shows the histogram of differentials, which appear as a truncated normal distribution. The negative shift in the $\mu_\mathbb{D}$ revealed that of the *number of points* or *distribution size* alone is insufficient to model the number of clusters users would perceive—an accurate model needs to consider the overlap of clusters. For example, in Fig. 3, all images have an identical number of generated clusters, but the interaction between clusters causes differing numbers of clusters to appear. Instead, the distance between clusters or the visual density of the data needs to be considered as an additional factor in cluster perception modeling. The next section introduces models that each considers one of these factors.

5 Topology-based Modeling of Clustering

We propose 2 models for capturing human perception of clusters based upon approaches from Topological Data Analysis (TDA) [84]. TDA is a set of approaches used to study the "shape" of data, including scalar fields [67, 79], vector fields [82], and high-dimensional data [13, 55].

Both models capture the clustering structure using a data structure called the *merge tree*. The merge tree encodes a series of topological events in the form of creation and merging of components (specifically, 0-dimensional homology groups), based upon properties of the space under a real-valued function. The first model, based upon the distance between cluster centers, is captured using a technique called *persistent homology* [24]. The second model, based upon the visual density of points, is captured by calculating the join tree of a scalar field [9].

5.1 Distance-based Model

The distance-based model tries to capture human perception of clusters by considering the spatial resolution at which 2 or more cluster distri-

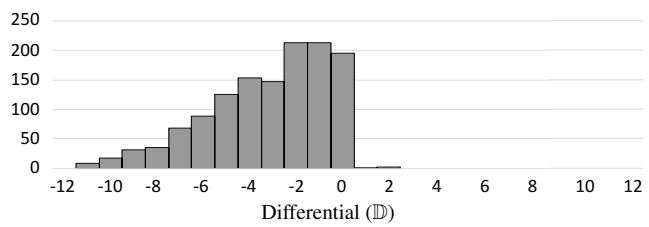


Fig. 4. The histogram of user responses *differential*, $\mathbb{D}$, against frequency for the preliminary experiment, appears as a truncated normal distribution.

Construction begins with a finite set of points V representing cluster centers embedded in Euclidean space (i.e., their positions on the scatterplot). Given a real number $D >= 0$, we consider a set of balls of diameter D centered at points in V. Continuously increasing the diameter, $0 = D_0 \leq D_1 \leq D_2 \leq \cdots \leq D_m = \infty$, forms a 1-parameter family of nested unions of balls. If at a given diameter D_i, 2 balls overlap, we consider these balls as a single component. Fig. 5(b) shows an example dataset with 4 values of D_i. As D_i increases, more balls intersect and merge into larger components. At D_∞, all balls will overlap, forming a single component.

To compute the PH, the points V form the vertices of a graph. A 1-simplex (an edge) is formed between 2 points in V if and only if their balls intersect (i.e., the distance between them is $\leq D_i$). Sweeping D_i from $0 \to \infty$, as D_i increases, new edges are added to the graph. Components are efficiently calculated at each step by finding connected components of the graph using the set union data structure. The total computation time is $O(|E|\alpha(|V|))$, where E are the edges of the graph, and α refers to the inverse Ackermann function, an extremely slow-growing function. At D_∞, PH forms the complete graph. Therefore, there are $O(|V|^2)$ edges.

Creating the merge tree from the prior construction is relatively simple. The merge tree is parameterized with respect to D. At $D_0 = 0$, all cluster center components are *born*. In other words, the balls have 0 volume. These birth events appear in the merge tree as 1 node per cluster, e.g., see the bottom of Fig. 5(c). As D_i increases, when 2 components first merge at a given D_i, a *merge node* is added to the merge tree at D_i connecting those components. For example, at D_1, the purple and pink components intersect, causing them to merge into a single component. From that point forward, 1 of the merged components *dies* (i.e., no longer exists), while the other takes on the identity of the new merged component (in this context, it does not matter which). Referring back to Fig. 5(c), when purple and pink merge at D_1, pink dies, while purple takes on the identity of the merged component. When the components finally merge into a single component, yellow in our example, this component dies at ∞. In other words, no matter how large the balls get, that 1 component will exist. It is also relevant to note that this particular construction has many parallels to single-linkage hierarchical clustering.

This model has 2 main limitations: (1) it assumes that clusters are isotropic and have similar distributions; and (2) it requires knowledge about the location of cluster centers. Our next model uses a related framework to overcome these limitations.

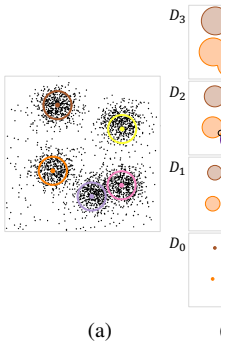
5.2 Density-based Model

The density-based model attempts to directly identify the *relative* visual density at which users will differentiate between clusters. The density-based model is found by calculating the join tree of a scalar field. We again provide a simplified treatment—for a detailed description, see [9].

First, a 2D histogram of the visual density is created for the scatterplot (i.e., a density plot). The image plane is divided into a set of grid cells of uniform width and height (selection of this resolution is discussed in our evaluation). Within each grid cell, the number of *white* pixels is counted, and this is considered the density[2], f_{xy}. For illustrative purposes, this value is mapped to the range $F \in [0, 255]$, where 0 is empty (i.e., completely black) and 255 is full (i.e., completely white), as shown in Fig. 6(a).

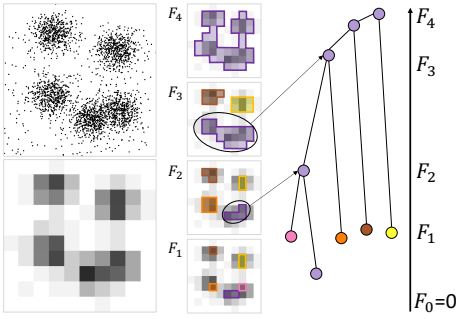
The components of the density histogram are identified by sweeping F, such that $0 = F_0 < F_1 < F_2 < \cdots < F_m = \infty$. At each F_i, histogram cells where $f_{xy} \leq F_i$ are extracted and components found by joining neighboring cells (we use the 8 surrounding neighbors). This is computed by treating histogram cells as graph nodes, V, iff $f_{xy} \leq F_i$. Graph edges, E, connect vertices that are neighbors in the density histogram, and connected components are extracted using the set union data structure with performance $O(|E|\alpha(|V|))$. Since only immediate neighbors are considered for connecting, there are $O(|V|)$ edges.

[2]We acknowledge this is not the usual calculation of density, e.g., see [20], which would count the number of black pixels. However, our configuration makes the remainder of the discussion easier.



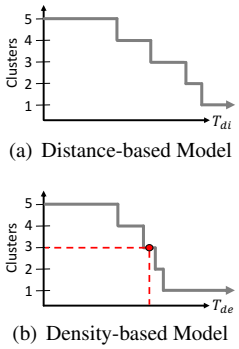
(a)

Fig. 5. With the distance-based model, (a) given the input data, (b) cluster centers are analyzed at a series of ball diameters. Connected components receive a unique color at each diameter. (c) The scan of diameters leads to the creation of the merge tree, which marks topological events (i.e., creation and merging of components) with nodes.



(a) (b) (c)

Fig. 6. With the density-based model, (a) given the input data (top), a density histogram (bottom) is calculated. (b) The space is analyzed at different density values, and components are extracted. (c) Tracking components across density values leads to the creation of the merge tree, which marks topological events with nodes.



(a) Distance-based Model

(b) Density-based Model

Fig. 7. The persistence threshold plots for the (a) distance- and (b) density-based models in Fig. 5 and Fig. 6, respectively. The horizontal axis represents the threshold, while the vertical axis shows the number of clusters. The red line shows how a threshold can be extracted from a given number of clusters and vice versa.

To construct the merge tree, sweeping F_i from $0 \rightarrow \infty$, nodes are *born* at the first F_i, where a new component appears. As F_i is increased, the components expand until they merge with another component. When components merge, the component with the more recent birth (i.e., higher f_{xy}) *dies*, while the component with the lower f_{xy} continues. For example, in Fig. 6(c), at F_1, the pink and purple components are about to merge. When they do at F_2, the pink component dies since it was born more recently (i.e., $f_{pink} > f_{purple}$), and the merged component in purple continues. Once all clusters have merged into a single component, that component dies at ∞ (i.e., it always exists, no matter how large F_i gets).

The value of this model over the distance-based model is that it only requires the input scatterplot. It needs no information about the cluster centers, and it makes no assumptions about the distribution of points within those clusters.

5.3 Persistence Threshold Plot

Thus far, the models only encode the clustering structure as a function of distance or as a function of density in the merge tree. The method to select the number of clusters that will be perceived by a user is calculated similarly, irrespective of the underlying model, though the input parameters (distance vs. density) have different meanings.

For this, we generate a persistence threshold plot. For a given merge tree, each component has its *persistence*, ρ, calculated. The persistence is the difference between birth and death values of the component (i.e., $\rho = death - birth$)[3]. The fundamental intuition behind persistence is that it measures the relative scale of a feature (e.g., the relative change in density), as opposed to the absolute scale of the feature (e.g., the absolute density value). We use persistence as a threshold to model the number of clusters a user would count in a scatterplot and vice versa.

This information is represented in a *persistence threshold plot* or *threshold plot*. To form the plot, for the threshold $T \in [0, \infty)$, at a given T_i, we count the number of clusters whose $\rho > T_i$. This information is encoded into the line chart (see Fig. 7) by plotting the threshold T horizontally and the number of clusters vertically.

Given these functions, we have the ability to determine critical thresholds (using either model) for the visual separation of clusters. For example, the red dashed lines in Fig. 7(b) show the persistence threshold (T_{de}) that corresponds to perceiving 3 clusters and vice versa. With this relationship, our models can now be used to estimate the number of clusters that a user would select in a given scatterplot.

[3]For the distance-based model, birth is always 0 making $\rho = death$. However, the full definition unifies the distance- and density-based models.

6 Main Experiment

We evaluate how well the merge tree models estimate the number of clusters perceived in a scatterplot by studying 3 factors (S, N, P). In addition to revisiting [H1] and [H2], we include 3 new hypotheses:

[H3] Data point size (P), having a direct impact on visual density, affects the accuracy in cluster count identification in scatterplots.

[H4] Using a persistence threshold correlated to the distribution size (S) of normally distributed clusters, the distance-based model will estimate the number of clusters perceived by users.

[H5] Using a persistence threshold correlated to the size of data point (P), the number of data points (N), and by their interaction effect ($N * P$), the density-based model will estimate the number of clusters perceived by users.

6.1 Properties and Data Generation

Using the information learned in preliminary experiment, following values were modified for the main experiment (i.e., all others remained the same, see Sect. 4.1):

- Data point size/area (P): $\{1_{px}, 3_{px}, 5_{px}, 7_{px}\}$ — On the low end, 1_{px} point size is the minimum possible value. On the high end, 7_{px} was chosen in combination with the number of points to limit the maximum visual density to $\sim 30\%$ of a given stimulus.
- Number of data points (N): $\{500, 2500, 12500\}$ — To decide the number of data points, we considered if data points are uniformly distributed, the *maximum visual density* is $MVD = \frac{N*a}{X*Y}$, where $[X \times Y]$ are stimuli dimensions $[550 \times 550]$. With a target of $< 30\%$, using $P = 7_{px}$ and $N = 12500$ the visual density, $MVD = 0.29$, i.e., 29% of pixels filled. We noted a logarithmic effect in the preliminary experiment. Therefore, logarithmic intervals (base 5) are used.
- Distribution size (S): $\{25_{px}, 40_{px}, 55_{px}, 70_{px}, 85_{px}\}$ — The distribution size was chosen to be similar to the preliminary experiment, slightly adjusted to have fixed intervals of 15_{px} between values.

The data generation process is kept similar to the preliminary experiment. A key difference is that task stimuli are generated for each participant covering all combinations of factors. For each subject, $|S| \times |N| = 15$ dataset are generated and rendered into $|15| \times |P| = 60$ scatterplot stimuli. Each participant received similar variability and the same combination of factors in their stimuli.

6.2 Study Procedure

This study was designed similarly to the preliminary exper
Sect. 4.2) with the following variations. Each subject wa
stimuli from their own pool of 60 stimuli in random ord
included a post-test questionnaire, asking participants to de
criteria for selecting the number of clusters.

We recruited participants from Amazon's Mechanical T
for the IRB approved study [8, 19]. Based upon a post
analysis of the preliminary experiment data, we recruited a
participants (21 male, 19 female; ages: $[18-64]$, median
$[25-34]$) limited to the US or Canada. 45% of participan
having corrected vision. All participants had a HIT appro
$\geq 95\%$, and were compensated at US Federal minimum wa

In total, 60 trials $\times$ 40 participants $=$ 2400 responses wer
We carried out some data quality checks on responses, and th
responses were eliminated—9 responses with task completi
less than 1 second and 27 responses that ran out of time —leaving a
total of 2364 responses for analysis.

Suitability of Studying Point Size Using AMT Studying visual factors, mark size in particular, on a crowdsourced environment has potential biases due to lack of control of user hardware, retinal size, viewing distance, ambient lighting, etc. For example, search task performance decreases as the viewing angle increases [30]. However, this lack of control is a commonly accepted limitation in crowdsourced studies—numerous recent AMT studies have considered mark size, among other properties, that could be impacted by this lack of environmental control, e.g., Szafir's study of perceived color differences [77], Chung et al.'s evaluation of orderability in visual channels [14], and Kim and Heer's study of the effectiveness of multiple visual encodings [46].

6.3 Analysis Methodology

We ran our data and user responses through the merge tree-based models. For the distance-based model, we first take the centers of each cluster to build the model. Then, we use the user response to the number of clusters (U) to extract a persistence threshold, T_{di}. After generating the threshold for all stimuli, a linear regression, using linear least squares, is calculated for T_{di} on the factor distribution size, $T^S_{di}(s) = c_1 \cdot s + c_2$, where the distribution size, s, is input, and c_1 and c_2 are calculated by the regression. Fig. 8(a) shows the resulting regression.

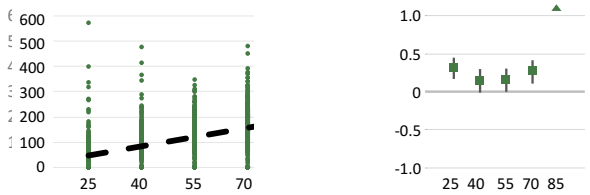
The density-based model is built by using the scatterplot to generate a visual density histogram, which is the input to the model. Then, the user response to the number of clusters (U) is used to extract a persistence threshold, T^*_{de}. For the density-based model, multiple factors are tested (N, P, and $N*P$), each requiring their own linear regression, i.e., $T^N_{de}(n) = c_1 \cdot n + c_2$; $T^P_{de}(p) = c_1 \cdot p + c_2$; and $T^{N*P}_{de}(n,p) = c_1 \cdot n + c_2 \cdot p + c_3$.

Threshold functions (T^S_{di} and T^*_{de}) from the merge tree are used to calculate the *model-predicted number of clusters*. To measure the accuracy of the user response on a given scatterplot, τ, we add new *differentials*, $\mathbb{D}^S_{di}$ and $\mathbb{D}^*_{de}$, for the distance- and density-based models, respectively:

$$\mathbb{D}^S_{di}(\tau) = U_\tau - C_{di}(T^S_{di}(\tau)) \qquad \mathbb{D}^*_{de}(\tau) = U_\tau - C_{de}(T^*_{de}(\tau))$$

where U_τ is the user response, and C_{di} and C_{de} are the number of clusters produced by the models using a given threshold. We used the value of *differentials* ($\mathbb{D}$, $\mathbb{D}^S_{di}$, and $\mathbb{D}^*_{de}$) as the primary measure to analyze the effects of the factors in the cluster counting. The histograms of the differentials for both models can be found in Fig. 9.

The study followed a within-subjects design, where all 40 subjects were exposed to all the same treatment. Hence, we use repeated measures (RM) ANOVA to analyze the effects of the factors on $\mathbb{D}$. For some results, due to violations of sphericity, according to Mauchly's test, reported degrees of freedom and p-values are Greenhouse-Geisser corrected (highlighted in green) [40, 58]. Along with RM ANOVA, we calculated partial eta-squared (η^2). As per Cohen's guidelines for measures of η^2: 0.01 denotes small effect, 0.06 denotes medium effect, and 0.14 denotes large effect [17].



(a) Threshold plot (b) Differential plot

Fig. 8. Plots of the threshold (T^S_{di}) and differential ($\mathbb{D}^S_{di}$) against distribution size (S) for the distance-based model. (a) The regression (in black) $T^S_{di}(x)$ is extracted for the distribution size (S), in green. (b) The regression is, in turn, used to calculate the model-predicted number of clusters, which are plotted by the mean and 95% confidence interval of the differential ($\mathbb{D}^S_{di}$) against different values of distribution size (S).

6.4 Results

6.4.1 Model Accuracy

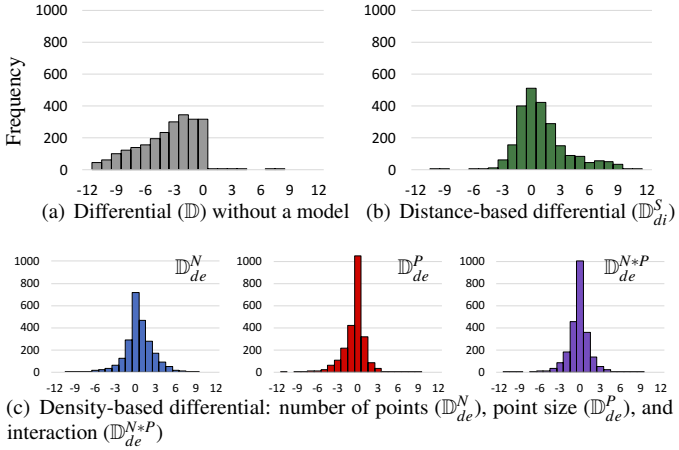
The distance- and density-based models both successfully estimated user perception for counting clusters. Fig. 9 shows the performance of all models in terms of differential. From our analysis, we observed the highest estimation accuracy was achieved using the density-model, from best to worst, $\mathbb{D}^{N*P}_{de}$: ($\mu = 0.18$, $\sigma = 1.58$); $\mathbb{D}^P_{de}$: ($\mu = 0.50$, $\sigma = 1.67$); and $\mathbb{D}^N_{de}$: ($\mu = -0.53$, $\sigma = 2.14$). The distance-based model performs next best, $\mathbb{D}^S_{di}$: ($\mu = 1.12$, $\sigma = 2.64$). Whereas, without a model performed the worst, $\mathbb{D}$: ($\mu = -3.74$, $\sigma = 3.00$).

6.4.2 Factor Effect Analysis Without a Model

We performed 3-factor RM ANOVA testing to analyze the factors, distribution size (S), number of points (N), and point size (P) in terms of the effect on the differential without a model, $\mathbb{D}$.

We observed that the *distribution size* (S) and the *number of points* (N) had a significant effect for counting clusters with respect to the differential, $\mathbb{D}$, with ($F_S(4, 2304) = 286.11, p < 0.001, \eta^2 = 0.32$) and ($F_N(1.98, 1576.43) = 33.98, p < 0.001$, $\eta^2 = 0.029$), respectively. On the other hand, *data point size* (P) failed to reach significance, with ($F_P(3, 2304) = 0.21, p = 0.889$, $\eta^2 = 0.0002$). We also tested for interaction effects and only observed a significant effect between S and N, ($F_{S*N}(8, 2304) = 8.18, p < 0.001, \eta^2 = 0.028$).

The η^2 analysis showed a large effect size on distribution size (S) and a small on the number of points (N) and interaction effect $S*N$. This is likely because smaller distributions create denser clusters with



(a) Differential ($\mathbb{D}$) without a model (b) Distance-based differential ($\mathbb{D}^S_{di}$)

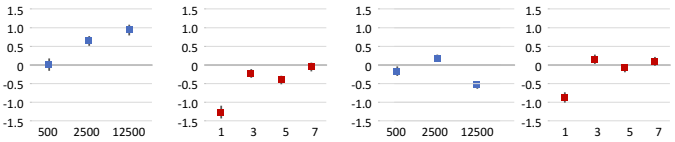
(c) Density-based differential: number of points ($\mathbb{D}^N_{de}$), point size ($\mathbb{D}^P_{de}$), and interaction ($\mathbb{D}^{N*P}_{de}$)

Fig. 9. Histograms for user response differential (horizontally) against frequency (vertically) for (a) no model (skew due to users' underestimation), (b) the distance-based model, and (c) density-based models. Responses that are closer to 0 imply a good fit for the data.

Table 1. Distance-based model Repeated-Measures ANOVA results on the effect of visual factors on the *differential*, $\mathbb{D}^S_{di}$.

Factors	Df	F-value	P-value	η^2	σ
Distribution size (S)	4	552.9	≤ 0.001	0.4900	1.91
Number of points (N)	1.98	051.39	≤ 0.001	0.0420	2.61
Size of points (P)	3	000.41	0.746	0.0005	2.64
S*N	8	002.72	0.006	0.0100	-
S*P	12	000.13	1.000	0.0006	-
N*P	6	000.08	0.998	0.0002	-
S*P*N	24	000.21	1.000	0.0020	-

-: not calculated; Greenhouse-Geisser corrected.



(a) $\mathbb{D}^N_{de}$ for the number of points (b) $\mathbb{D}^P_{de}$ for the size of points (c) $\mathbb{D}^{N*P}_{de}$ for the number (left) and size (right) of points.

Fig. 11. Density-based model mean and 95% confidence intervals of user response differential, $\mathbb{D}^*_{de}$, on factors with density histogram $[20_{px} \times 20_{px}]$.

better separation, while larger distributions blend to create ambiguous boundaries. From these results, **both [H1] and [H2] are reconfirmed**. The lack of significance on point size (P) indicates that [H3] should be rejected. However, we will revisit this hypothesis later.

6.4.3 Distance-based Model Factor Analysis

Using persistence threshold on distribution size, T^S_{di}, we calculated the *differential* ($\mathbb{D}^S_{di}$) and performed 3-factor RM ANOVA to observe the main effects of the individual factors distribution size (S), number of points (N), and point size (P), as well as interaction effects (see Table 1).

The analysis identified a significant effect of distribution size (S) and the number of points (N) on the *differential* ($\mathbb{D}^S_{di}$), but the point size (P) failed to reach significance. In particular, we found a large effect for distribution size (S) on $\mathbb{D}^S_{di}$. We also observed a small-medium effect in the number of points (N) and a negligible effect on the point size (P). We did not anticipate any interaction effects, and only $S*N$ showed a small effect. In terms of accuracy, as pointed out in Sect. 6.4.1, the distance-based model improved overall accuracy over using no model (see Fig. 9(b)). Investigating further, Fig. 8(b) shows the accuracy per distribution size. Note that the accuracy was sound for all distribution sizes, except at $S = 85$, which negatively impacted overall performance. We speculate that this is due to the significant blending of distributions at this extreme. Given the large effect in S and overall improvement in accuracy, **we consider [H4] confirmed**.

6.4.4 Density-based Model

For the density-based model, we calculate 3 variations of the threshold and differential that use the factors that most directly influence visual density. Those are the number of data points ($T^N_{de}/\mathbb{D}^N_{de}$), the size of data points ($T^P_{de}/\mathbb{D}^P_{de}$), and their interaction ($T^{N*P}_{de}/\mathbb{D}^{N*P}_{de}$). For each, we performed 3-factor RM ANOVA testing on the individual factors the distribution size (S), the number of points (N), and the point size (P), as well as interaction effects (see Table 2).

Histogram Resolution The density-based model uses the *visual density* of a given scatterplot to model cluster perception. To calculate visual density, a 2D histogram is calculated on the image with bins of uniform width and height, $[B_{px} \times B_{px}]$. The choice of bin size for the density histogram is potentially influential in our analysis, as bins that are too small may cause instability, and bins that are too large may miss clusters. To determine the appropriate bin size, we performed an analysis on the data from Fig. 3. A set of stimuli images are generated with fixed values for factors ($C = 6$, $N = 2500$, and $P = 7_{px}$) and different values for $S = \{25_{px}, 40_{px}, 55_{px}, 70_{px}, 85_{px}\}$. We plotted the normalized density threshold (i.e., density threshold divided by area of a bin, i.e., T_{de}/B^2) generated by $U = 6$ clusters for 5 different bin sizes (see Fig. 10). The results showed instability in the density threshold for smaller values and a stable result starting at $[20 \times 20]$. For this reason, 3 resolutions of histogram cell sizes are reported: $[10_{px} \times 10_{px}]$, $[20_{px} \times 20_{px}]$, and $[40_{px} \times 40_{px}]$, but our main discussion focuses on $[20_{px} \times 20_{px}]$.

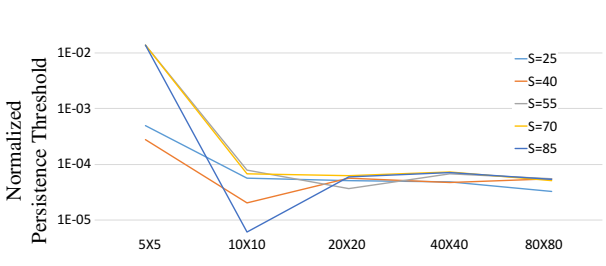


Fig. 10. Normalized persistence threshold for the density-based model for the data from Fig. 3 shows stability at histogram bins size $[20_{px} \times 20_{px}]$ and larger.

Number of Points Model ($T^N_{de}/\mathbb{D}^N_{de}$) RM ANOVA results demonstrate significant and consistent main effects of S, N, P, and interaction effect of $N*P$, which can be seen in Table 2. Point size has a large effect on $\mathbb{D}_{de}$, confirming our hypotheses and previous work (e.g., [68]) of density's influence on cluster perception. The number of points showed a small-medium effect size on $\mathbb{D}_{de}$, also align with our hypotheses. The accuracy of the number of points model was the worst of the 3 density models, though still significantly better than no model (see Fig. 9(c)). The accuracy of the model, plotted by the number of points in Fig. 11(a), shows lower accuracy as the number of points increases.

Point Size Model ($T^P_{de}/\mathbb{D}^P_{de}$) In this model, the number of points showed a medium-large effect size, while point size demonstrated a medium effect size for the differential. On the other hand, interaction of $N*P$ results small values of η^2 (see Table 2). The overall accuracy of this model was better than the number of points model (see Fig. 9(c)). Fig. 11(b) shows the accuracy per point size. The model was largely accurate, except for the smallest size, $P = 1_{px}$.

Interaction Model ($T^{N*P}_{de}/\mathbb{D}^{N*P}_{de}$) Similar to the previous 2 models, significant effects were observed for all factors. However, only point size demonstrated medium effect size (see Table 2). This model showed the best overall accuracy of any model tested (see Fig. 9(c)). This makes logical sense, as the density is the combination of the number of points and their size. Fig. 11(c) shows the accuracy per number of points and per point size. For both cases, the accuracy was improved. However, $P = 1_{px}$ was still the worst performing category.

Our analysis showed the number of points, point size, and their interaction all had significant effects and improved accuracy over no model. **Therefore, we consider [H5] confirmed**. Furthermore, we identified some large effects with point size for the density-based model, and **this indirect relationship confirms [H3]**.

6.4.5 Post-Test Questionnaire

To further support our hypotheses, we asked the participants to state the criteria that influenced their counting of clusters in a free-response format at the end of the experiment. Their responses largely mirrored our findings—10% cited the size of symbol; 25% responses cited something amounting to distribution size; 25% cited distance between clusters; and 65% of responses included density as a factor[4].

6.5 Follow-up Study on Opacity

We now evaluate opacity modification, which studies have shown to be more effective in overdraw reduction than other techniques, such as

[4]Some subjects listed multiple criteria.

Table 2. Density-based model Repeated-Measures ANOVA results on the effect of visual factors for user response accuracy (i.e., *differential*), $\mathbb{D}^*_{de}$.

Model	Grid	Distribution Size (S) Df	F-val	p-val	η^2	Number of Points (N) Df	F-val	p-val	η^2	Point Size (P) Df	F-val	p-val	η^2	Interaction (N*P) Df	F-val	p-val	η^2	σ
$\mathbb{D}^N_{de}$	$10_{px} \times 10_{px}$	4	061.45	≤ 0.001	0.090	2	122.60	≤ 0.001	0.090	3	086.28	≤ 0.001	0.099	6	11.28	≤ 0.001	0.028	1.640
	$20_{px} \times 20_{px}$	4	014.99	≤ 0.001	0.020	2	058.80	≤ 0.001	0.045	3	301.80	≤ 0.001	0.280	6	10.20	≤ 0.001	0.025	1.760
	$40_{px} \times 40_{px}$	4	129.22	≤ 0.001	0.180	2	216.00	≤ 0.001	0.150	3	613.20	≤ 0.001	0.430	6	56.90	≤ 0.001	0.120	2.440
$\mathbb{D}^P_{de}$	$10_{px} \times 10_{px}$	4	088.58	≤ 0.001	0.130	2	096.50	≤ 0.001	0.075	3	006.00	≤ 0.001	0.086	6	1.62	0.103	0.004	1.680
	$20_{px} \times 20_{px}$	4	004.98	≤ 0.001	0.008	2	143.80	≤ 0.001	0.100	3	074.30	≤ 0.001	0.080	6	11.40	≤ 0.001	0.028	1.520
	$40_{px} \times 40_{px}$	4	094.45	≤ 0.001	0.130	2	439.70	≤ 0.001	0.270	3	135.40	≤ 0.001	0.140	6	010.80	≤ 0.001	0.250	2.070
$\mathbb{D}^{N*P}_{de}$	$10_{px} \times 10_{px}$	4	057.95	≤ 0.001	0.090	2	383.90	≤ 0.001	0.240	3	246.30	≤ 0.001	0.240	6	130.30	≤ 0.001	0.250	1.710
	$20_{px} \times 20_{px}$	4	007.56	≤ 0.001	0.012	2	047.10	≤ 0.001	0.038	3	061.20	≤ 0.001	0.070	6	11.40	≤ 0.001	0.028	1.470
	$40_{px} \times 40_{px}$	4	062.81	≤ 0.001	0.096	2	273.30	≤ 0.001	0.180	3	195.50	≤ 0.001	0.190	6	19.10	≤ 0.001	0.046	1.910

The $[20_{px} \times 20_{px}]$ grid is highlighted to indicate it is the primary focus of our analysis.

reducing point size or changing the shape of the data point [36]. Given the results for the density-based model, we hypothesize that it will be able to model the perception of clusters when opacity is applied to the data points.

[H6] Using a density-threshold correlated to the opacity of data points (O), the density-based model will have a significant effect for the number of clusters perceived by the viewer.

Properties and Data Generation Our data synthesis and rendering of scatterplots is similar to main experiments (see Sect. 6.1) in most aspects. We fix the number of points $N = 200,000$ and point size $P = 7_{px}$ to overdraw the data on the scatterplot (see Fig. 12 for an example). The distribution size is the same as in the main experiment $S = \{25_{px}, 40_{px}, 55_{px}, 70_{px}, 85_{px}\}$. The data point opacity was selected on a logarithmic interval, $O = \{1\%, 10\%, 100\%\}$ over a white background. Each subject sees each condition 2 times. Thus, for each subject, $2 \times |S| \times |N| = 10$ datasets are generated and rendered into $|10| \times |P| \times |O| = 30$ scatterplot stimuli.

Study Procedure The task for the study was identical to the main experiment in Sect. 6.2. We recruited 40 participants (21 male, 19 female; median age group: $[35-44]$) from AMT, limited to subjects located in the US or Canada. 45% of participants reported having corrected vision. Subjects were compensated at US federal minimum wage. In total 30 tasks $\times$ 40 participants resulted in 1200 responses. We carried out data quality checks on responses—2 participants (60 total responses) were discarded because the majority of responses were the default value of 1, and 23 responses that ran out of time were rejected. A total of 1117 responses were analyzed.

6.5.1 Analysis and Results

The analysis was performed similarly to Sect. 6.4.4. We performed 2-factor RM ANOVA testing and evaluated the effect of factor opacity of data point (O) with varying the distribution size (S) on measure $\mathbb{D}^O_{de}$.

The calculation of the visual density histogram was modified such that it summed the pixel intensities, instead of counting the number of filled pixels. For building the histograms, we only considered the bin of size $[20_{px} \times 20_{px}]$. We calculated the density threshold on the opacity factor using linear least squares regression on $T^O_{de}(o)$ Using T^O_{de}, we calculated the *differential* $\mathbb{D}^O_{de}$.

Opacity showed a medium-large effect ($F_O(2, 1102$ $0.001, \eta^2 = 0.09$), followed by a medium effect for tion of data points ($F_S(4, 1102) = 17.44, p < 0.001$, The interaction effect of O and S also showed mediur ($F_{O*S}(8, 1102) = 12.33, p < 0.001, \eta^2 = 0.12$). The eff observed **confirm [H6]**.

Perhaps unsurprisingly, further investigation suggeste has a more substantial effect when the distribution of the and a smaller effect when the data distribution is spars confirm previous findings on scatterplot overplotting [56], a larger distribution size in a scatterplot requires more c whereas a narrower distribution size requires more trans

7 Model Usage

Identifying clusters is an important low-level visual analytics task [3], as well as in data analysis in general [64]. Still, as mentioned in Sect. 1, clustering is an ill-posed problem, with the "correct result" being subject to the constraints of the algorithm or individual performing the analysis. Through our evaluation, we have shown that our models, the density-based model, in particular, performed well in estimating the number of clusters an average human would perceive. However, this in and of itself is not the real application value of the models. Instead, the models can be used to optimize the visual encodings to maximize the saliency of the visualization. Furthermore, the threshold plots provide an evidence-based rationale for design decisions.

7.1 Controlling Design Factors

The models we have introduced construct a bridge for visualization designers between their choice of visual encodings and how users perceive clusters. Table 3 summarizes our findings and suggests how designers should focus their design decisions on selecting visual properties that robustly support cluster identification in scatterplots. For the distance-based model, *distribution size* (S) was the only factor with a large effect size. On the other hand, with the density-based model, the *number of points* (N), *point size* (P), and *opacity* (O) all showed large effects on cluster count perception.

Distribution Size — Visualization designers generally *do not* have control over distribution size (S) in the data. Although distributions are rarely known a priori, they can be extracted from scatterplots, e.g., using Gaussian mixture models, which, combined with the distance-based model, could be used to help the designer to understand the number of clusters a user is likely to see. Nevertheless, this approach is unlikely to be helpful to the majority of designers.

Number of Points — Visualization designers have *limited control* of the number of points (N), mostly in terms of data subsampling, e.g., [12,45,47], which influences visual density. Using uniform random sampling, e.g., [21,28], or targeted nonuniform sampling, e.g., by using density [6], can reduce the number of points and, thus, the visual density. The density-based model can be used to evaluate what level of sampling provides optimal saliency of clusters.

Factor	Without Model	Distance-based (T_{di})	Density-based (T^N_{de})	(T^P_{de})	(T^{N*P}_{de})	(T^O_{de})
Distribution size (S)	●●●	●●●	○	○	○	●●
Number of points (N)	○	○	○	●●	○	—
Size of points (P)	∅	∅	●●	●●	○	—
Opacity (O)	—	—	—	—	—	●●
Interaction (N*P)	∅	∅	○	○	○	—
Interaction (O*S)	—	—	—	—	—	●●
Interaction (S*N)	○	○	—	—	—	—

●●● large effect; ●● medium effect; ○ small effect; ∅ negligible effect; — not tested

Point Size — The point size (P) is the first design factor with *complete control* in scatterplots. As pointed out earlier, increasing the area of pixels also increases the visual density. Once again, the density-based model can be used to help select the point size that provides the optimal saliency. There is an important interplay between the number of points and the point size, as adjusting either can influence the visual density.

Opacity — Opacity (O) is another factor for which the designer has *complete control*, from fully transparent to fully opaque, once again impacting the visual density of the scatterplot. As suggested by Urribarri and Castro, when selecting opacity, there is a trade-off with picking a point size [80] and, given our analysis, also with the number of data points shown. Nevertheless, using the density-based model, various opacity levels, along with the number of points and their size, can be evaluated and the optimal configuration selected.

7.2 Threshold Plots for Optimizing Cluster Saliency

As our goal is to improve the effectiveness of the visualization design, it is important to understand how designers can use our models to reduce ambiguity in the data, and thereby reduce the chance of misinterpretation, e.g., by having a visualization that is too sparse or over-saturated. Using pre-studies of the effects of different visual encoding configurations in scatterplots, visualization practitioners can pick the configuration that maximizes the visibility of clusters. See <https://usfdatavisualization.github.io/TopoClusterPerceptionDemo> for a demo.

Consider Fig. 12, for example. The 3 scatterplots are each plotted in the persistence threshold plot. The horizontal axis reveals for each plot how salient each of the clusters are. The 10% opacity plot shows between 1 and 8 clusters are visible, but either 2 or 6 clusters are most visible. That is not to say other numbers of clusters are not visible, but they are simply not as distinctive. Of the 3 scatterplots in Fig. 12, 10% provides the best saliency, followed by 1%, then 100% opacity.

7.3 Case Study

To demonstrate the utility of the models on real data, we showed how the choice of visual encoding impacts the cluster perception. We performed a case study using dimensionality reduction on the MNIST dataset [48], which is an extensive database of handwritten letters commonly used to test machine learning techniques. Here we explore the dataset, which consists of 70K samples with 10 labels of handwritten digits (the labels are not used in sampling or rendering). We applied both t-SNE (see Fig. 1(a)) and PCA (see Fig. 1(b)) to plot the features on a 2D scatterplot and demonstrate the influence of 2 factors, the number of points and opacity.

In Fig. 1(a), we show the results of varying the number of points (after dimension reduction), where $N = \{500, 2500, 12500\}$, by using random sampling. The resulting persistence threshold plot shows that for $N = 500$, in red, clusters are difficult to differentiate. For $N = 2500$, in blue, and $N = 12500$, in purple, both have similar levels of effectiveness, with 12500 having a slight advantage, making it the better choice for representing this example.

In Fig. 1(b), we show the results of varying the opacity of the data points, where $O = \{1\%, 5\%, 10\%, 50\%, 100\%\}$. The results in the persistence threshold plot fall into 3 groups. On the first extreme, $O = 50\%$, in purple, and $O = 100\%$, in orange, provide no differentiation of any clusters. On the other extreme, $O = 1\%$, in blue, shows that a relatively low level of saliency for 1, 2, or 3 clusters. The final group, $O = 5\%$, in pink, and $O = 10\%$, in green, both show identically high levels of saliency for 1, 2, or 3 clusters in the data, making either of these the better choice for representing this example.

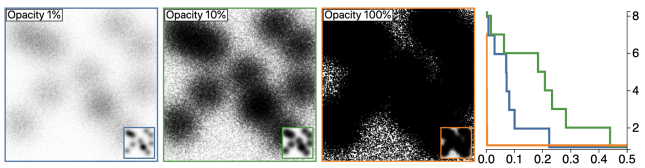


Fig. 12. Example of overplotted stimuli with $C = 11$, $N = 200{,}000$, $S = 55$, $P = 7_{px}$, but varying opacity, 1% (blue), 10% (green), and 100% (orange).

8 Discussion & Conclusions

Scatterplots are a common type of visualization, used to identify clusters in datasets. In this work, we tested and validated the importance of 4 visual factors—distribution size, the number of data points, the size of data points, and the opacity of data points—in cluster perception and built 2 models: a distance- and density-based model for the task. Our results confirm the theoretical models of Sadahiro, which states data points distribution (proximity), and number and size of data points (concentration and density change) affect cluster perception [68]. Finally, our findings confirm the important role that the choice of visual factors can have on cluster identification—visualization practitioners may apply these models for optimizing properties of their visualizations.

Model Limitations Both of our models have some limitations. In the distance-based model, we required knowledge of the centers of the clusters with a fixed-size isotropic normal distribution for the model—considering other distributions would likely require modifications to the model. This requirement is particularly restrictive with respect to non-synthetic data. We showed that user response accuracy in the density-based model was significantly better than the distance-based model. However, choosing the correct density histogram resolution is a critical task that may also be dependant on the data. A choice of an extremely high or low resolution could reduce the accuracy of the threshold value. Additionally, although the density model does not directly consider a normal distribution, we have only tested it against fixed-size normal distributions. Using the model with other types of distribution should be treated with caution.

Study Limitations The study itself has some additional limitations. First, we have not considered some other factors that could influence performance in either model, e.g., chart size, screen resolutions, etc. We have also not extensively analyzed variance between individuals, although we did note some small variation during our analysis (i.e., some individuals had over- or under-estimation tendencies). Another limitation that we have only provided a limited analysis of mixing effects, e.g., changing the size of points, while also changing the opacity. A final limitation is that we have not considered the correlation between confidence, which is highly related to the nature of data [33], and the correctness on each model.

Alternative Models Alternative models could potentially be developed to similarly explain the variance. With respect to distance, hierarchical clustering could be used, which is functionally equivalent to our distance model. For density, since stimuli are built on Gaussian distributions, a Gaussian Mixture Model (GMM) could be used. GMMs, being numerically extracted, cannot provide the same theoretical guarantees as our models, which are technically combinatorial. The theoretical guarantees, coming from persistent homology, also include *stability* guarantees. With stability, small changes in the input are guaranteed to produce only small changes to the output. A consequence of stability is robustness to noise. The noise has low persistence, not influencing the selection of the number of clusters.

Automatic Parameter Optimization One natural extension of this work is to develop a (semi-)automatic model for selecting design factors for a dataset. Unfortunately, using threshold plots as-is represents an under-constrained optimization, and it requires, *at the very least*, a user specification of the number of clusters in the data.

Code: <https://github.com/USFDataVisualization/TopoClusterPerception>
Demo: <https://usfdatavisualization.github.io/TopoClusterPerceptionDemo>
Data: <DOI: 10.17605/OSF.IO/49CD2>

Acknowledgments

We thank Bei Wang, Les Piegl, Jorge Adorno Nieves, Zach Beasley, Junyi Tu, and our reviewers for their input on this project. This project was supported in part by National Science Foundation (IIS-1845204).

REFERENCES

[1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: 10.1111/cgf.13684

[2] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2014. doi: 10.1109/VAST.2014.7042493

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization (InfoVis)*, 2005. doi: 10.1109/INFVIS.2005.1532136

[4] G. Anobile, G. M. Cicchini, and D. C. Burr. Number as a primary perceptual attribute: A review. *Perception*, 45(1-2):5–31, 2016. doi: 10.1177/0301006615602599

[5] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 1–8, 2016. doi: 10.1109/PACIFICVIS.2016.7465244

[6] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006. doi: 10.1057/palgrave.ivs.9500122

[7] L. A. Best, A. C. Hunter, and B. M. Stewart. Perceiving relationships: A physiological examination of the perception of scatterplots. In *International Conference on Theory and Application of Diagrams*, 2006. doi: 10.1007/11783183_33

[8] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee. Information visualization evaluation using crowdsourcing. *Computer Graphics Forum*, 37(3):573–595, 2018. doi: 10.1111/cgf.13444

[9] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Computational Geometry: Theory and Applications*, 24, 2003. doi: 10.1016/S0925-7721(02)00093-7

[10] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, and K.-L. Ma. Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 20, 2014. doi: 10.1109/TVCG.2014.2346594

[11] H. Chen, S. Engle, A. Joshi, E. D. Ragan, B. F. Yuksel, and L. Harrison. Using animation to alleviate overdraw in multiclass scatterplot matrices. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018. doi: 10.1145/3173574.3173991

[12] X. Chen, T. Ge, J. Zhang, B. Chen, C.-W. Fu, O. Deussen, and Y. Wang. A recursive subdivision technique for sampling multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 2019. doi: 10.1109/TVCG.2019.2934541

[13] A. I. Choudhury, B. Wang, P. Rosen, and V. Pascucci. Topological analysis and visualization of cyclical behavior in memory reference traces. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 9–16. IEEE, 2012. doi: 10.1109/pacificvis.2012.6183557

[14] D. H. Chung, D. Archambault, R. Borgo, D. J. Edwards, R. S. Laramee, and M. Chen. How ordered is it? on the perceptual orderability of visual channels. *Computer Graphics Forum*, 35(3):131–140, 2016. doi: 10.1111/cgf.12889

[15] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080

[16] E. H. Cohen, M. Singh, and L. T. Maloney. Perceptual segmentation and the perceived orientation of dot clusters: The role of robust statistics. *Journal of Vision*, 8(7):6–6, 2008. doi: 10.1167/8.7.6

[17] J. Cohen. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, 33(1):107–112, 1973. doi: 10.1177/001316447303300111

[18] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2018.2864907

[19] K. Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pp. 210–221. Springer, 2012. doi: 10.1007/978-3-642-35142-6_14

[20] T. N. Dang and L. Wilkinson. Transforming scagnostics to reveal hidden features. *IEEE Transactions on Visualization and Computer Graphics*, 2014. doi: 10.1109/TVCG.2014.2346572

[21] A. Dix and G. Ellis. By chance enhancing interaction with large data sets through statistical sampling. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 167–176, 2002. doi: 10.1145/1556262.1556289

[22] M. E. Doherty, R. B. Anderson, A. M. Angott, and D. S. Klopfer. The perception of scatterplots. *Perception & Psychophysics*, 69(7), 2007. doi: 10.3758/BF03193961

[23] E. Domany. Superparamagnetic clustering of data the definitive solution of an ill-posed problem. *Physica A: Statistical Mechanics and its Applications*, 263(1-4):158–169, 1999. doi: 10.1016/S0378-4371(98)00494-4

[24] H. Edelsbrunner and J. Harer. Persistent homology-a survey. *Contemporary Mathematics*, 453:257–282, 2008. doi: 10.1090/conm/453/08802

[25] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010. doi: 10.1090/mbk/069

[26] W. C. Eells. The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, 21, 1926. doi: 10.1080/01621459.1926.10502165

[27] G. Ellis, E. Bertini, and A. Dix. The sampling lens: making sense of saturated visualisations. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1351–1354, 2005. doi: 10.1145/1056808.1056914

[28] G. Ellis and A. Dix. Density control through random sampling: an architectural perspective. In *International Conference on Information Visualisation*, pp. 82–90. IEEE, 2002. doi: 10.1109/IV.2002.1028760

[29] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 2007. doi: 10.1109/TVCG.2007.70535

[30] J. M. Enoch. Effect of the size of a complex display upon visual search. *Journal of the Optical Society of America*, 49(3):280–286, 1959. doi: 10.1364/JOSA.49.000280

[31] R. Etemadpour and A. G. Forbes. Density-based motion. *Information Visualization*, 16(1):3–20, 2017. doi: 10.1177/1473871615606187

[32] R. Etemadpour, L. Linsen, C. Crick, and A. G. Forbes. A user-centric taxonomy for multidimensional data projection tasks. In *Information Visualization Theory and Applications (IVAPP)*, pp. 51–62, 2015. doi: 10.5220/0005313400510062

[33] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. De Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94, 2014. doi: 10.1109/TVCG.2014.2330617

[34] R. Etemadpour, B. Olk, and L. Linsen. Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. In *Information Visualization Theory and Applications (IVAPP)*, pp. 233–246. IEEE, 2014. doi: 10.5220/0004675802330246

[35] X. Z. Fern, I. Davidson, and J. G. Dy. Multiclust 2010: discovering, summarizing and using multiple clusterings. *ACM SIGKDD Explorations Newsletter*, 12(2):47–49, 2011. doi: 10.1145/1964897.1964910

[36] S. Few and P. Edge. Solutions to the problem of over-plotting in graphs. *Visual Business Intelligence Newsletter*, 2008.

[37] M. Friendly and D. Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 2005. doi: 10.1002/jhbs.20078

[38] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2316–2325, 2013. doi: 10.1109/TVCG.2013.183

[39] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1953–1962, 2014. doi: 10.1109/TVCG.2014.2346983

[40] S. W. Greenhouse and S. Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959. doi: 10.1007/BF02289823

[41] C. Gutwin, A. Cockburn, and A. Coveney. Peripheral popout: The influence of visual angle and stimulus intensity on popout effects. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2017. doi: 10.1145/3025453.3025984

[42] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979

[43] C. Healey and J. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012. doi: 10.1109/TVCG.2011.127

[44] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1303–1312, 2009. doi: 10.1145/1518701.1518897

[45] R. Hu, T. Sha, O. Van Kaick, O. Deussen, and H. Huang. Data sampling in multi-view and multi-class scatterplots via set cover optimization. *IEEE Transactions on Visualization and Computer Graphics*, 2019. doi: 10.1109/TVCG.2019.2934799

[46] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. *Computer Graphics Forum*, 2018. doi: 10.1111/cgf.13409

[47] B. C. Kwon, J. Verma, P. J. Haas, and C. Demiralp. Sampling for scalable visual analytics. *IEEE Computer Graphics and Applications*, 2017. doi: 10.1109/MCG.2017.6

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. doi: 10.1109/5.726791

[49] J. Lewis, L. Van der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, 2012.

[50] J. Li, J.-B. Martens, and J. J. van Wijk. A model of symbol size discrimination in scatterplots. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 2553–2562. ACM, 2010. doi: 10.1145/1753326.1753714

[51] J. Li, J. J. van Wijk, and J.-B. Martens. A model of symbol lightness discrimination in sparse scatterplots. In *IEEE Pacific Visualization Symposium (PacificVis)*, 2010. doi: 10.1109/PACIFICVIS.2010.5429604

[52] M. Li, F. Choudhury, Z. Bao, H. Samet, and T. Sellis. Concavecubes: Supporting cluster-based geographical visualization in large data scale. *Computer Graphics Forum*, 37(3):217–228, 2018. doi: 10.1111/cgf.13414

[53] H. Liao, Y. Wu, L. Chen, and W. Chen. Cluster-based visual abstraction for multivariate scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, pp. 2531–2545, 2018. doi: 10.1109/TVCG.2017.2754480

[54] Y. Ma, A. K. Tung, W. Wang, X. Gao, Z. Pan, and W. Chen. Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, pp:0, 2018. doi: 10.1109/TVCG.2018.2875702

[55] D. Maljovec, B. Wang, P. Rosen, A. Alfonsi, G. Pastore, C. Rabiti, and V. Pascucci. Rethinking sensitivity analysis of nuclear simulations with topology. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 64–71, 2016. doi: 10.1109/pacificvis.2016.7465252

[56] J. Matejka, F. Anderson, and G. Fitzmaurice. Dynamic opacity optimization for scatter plots. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2015. doi: 10.1145/2702123.2702585

[57] J. Matute, A. C. Telea, and L. Linsen. Skeleton-based scagnostics. *IEEE Transactions on Visualization and Computer Graphics*, 2017. doi: 10.1109/TVCG.2017.2744339

[58] J. W. Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940. doi: 10.1214/aoms/1177731915

[59] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 2013. doi: 10.1109/TVCG.2013.65

[60] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf. Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1588–1599, 2017. doi: 10.1109/TVCG.2017.2674978

[61] T. Munzner. *Visualization Analysis and Design*. CRC press, 2014. doi: 10.1201/b17511

[62] H. Nguyen, P. Rosen, and B. Wang. Visual exploration of multiway dependencies in multivariate data. In *SIGGRAPH ASIA Symposium on Visualization*, p. 2. ACM, 2016. doi: 10.1145/3002151.3002162

[63] J. O'Callaghan. Human perception of homogeneous dot patterns. *Perception*, 3(1):33–45, 1974. doi: 10.1068/p030033

[64] N. Otter, M. Porter, U. Tillmann, P. Grindrod, and H. Harrington. A roadmap for the computation of persistent homology, 2017. doi: 10.1140/epjds/s13688-017-0109-5

[65] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 3659–3669. ACM, 2016. doi: 10.1145/2858036.2858155

[66] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[67] P. Rosen, A. Seth, B. Mills, A. Ginsburg, J. Kamenetzky, J. Kern, C. R. Johnson, and B. Wang. Using contour trees in the analysis and visualization of radio astronomy data cubes. In *Topological Methods in Data Analysis and Visualization (TopoInVis)*, 2019.

[68] Y. Sadahiro. Cluster perception in the distribution of point objects. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 34(1):49–62, 1997. doi: 10.3138/Y308-2422-8615-1233

[69] B. Saket, A. Endert, and Ç. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2018.2829750

[70] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2017.2744184

[71] A. Sarikaya, M. Gleicher, and D. Szafir. Design factors for summary visualization in visual analytics. *Computer Graphics Forum*, 2018. doi: 10.1111/cgf.13408

[72] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632

[73] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 2012. doi: 10.1111/j.1467-8659.2012.03125.x

[74] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991. doi: 10.1002/acp.2350050106

[75] StatCounter. Desktop screen resolution stats worldwide. `http://gs.statcounter.com/screen-resolution-stats/desktop/worldwide`, 2019.

[76] M. Stone. Visualization viewpoints: In color perception, size matters. *IEEE Computer Graphics and Applications*, 32(2):8, 2012. doi: 10.1109/mcg.2012.37

[77] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2017.2744359

[78] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 2016. doi: 10.1167/16.5.11

[79] J. Tierny. *Topological data analysis for scientific visualization*. Springer, 2017. doi: 10.1007/978-3-319-71507-0

[80] D. K. Urribarri and S. M. Castro. Prediction of data visibility in two-dimensional scatterplots. *Information Visualization*, 16(2):113–125, 2017. doi: 10.1177/1473871616638892

[81] R. Veras and C. Collins. Saliency deficit and motion outlier detection in animated scatterplots. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019. doi: 10.1145/3290605.3300771

[82] B. Wang, P. Rosen, P. Skraba, H. Bhatia, and V. Pascucci. Visualizing robustness of critical points for 2d time-varying vector fields. *Computer Graphics Forum*, 32(3pt2):221–230, 2013. doi: 10.1111/cgf.12109

[83] C. Ware and M. D. Plumlee. Designing a better weather display. *Information Visualization*, 12(3-4):221–239, 2013. doi: 10.1177/1473871612465214

[84] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018. doi: 10.1146/annurev-statistics-031017-100045

[85] L.-Y. Wei. Multi-class blue noise sampling. *ACM Transactions On Graphics*, 29(4):79, 2010. doi: 10.1145/1778765.1778816

[86] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization (InfoVis)*, 2005. doi: 10.1109/INFVIS.2005.1532142

[87] A. K. Wong and E.-S. A. Lee. Aligning and clustering patterns to reveal the protein functionality of sequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(3):548–560, 2014. doi: 10.1109/TCBB.2014.2306840

[88] B. Wong. Points of view: gestalt principles. *Nature Methods*, 2010. doi: 10.1038/nmeth1110-863