Evaluation of Sampling Methods for Scatterplots

Category: Research

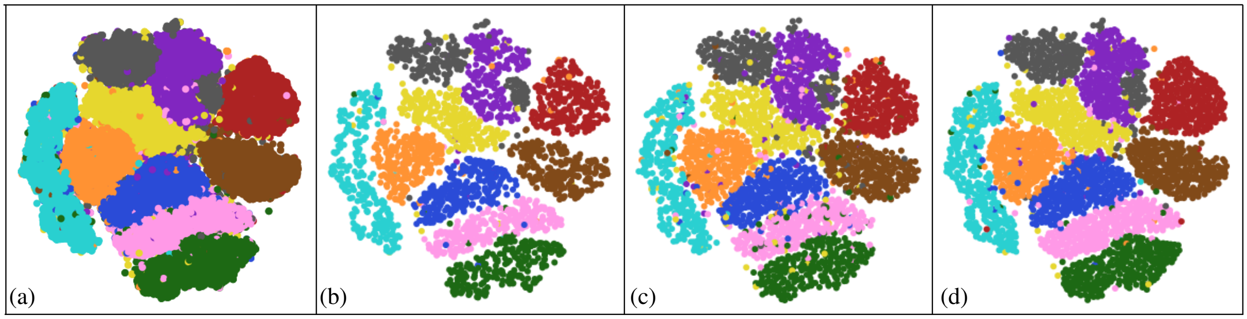


Fig. 1: Different sampling results of the MNIST dataset: (a) the original scatterplot; (b) the results of random sampling; (c) the result of outlier biased density based sampling; (d) the result of bluue noise sampling. The three sampling methods can preserve relative density, outliers and the overall shapes in terms of human perception, respectively.

Abstract— Given a scatterplot with tens of thousands of points or even more, a natural question is which sampling method should be used to create a small but "good" scatterplot for a better approximation. We present the results of a user study that investigates the influence of different sampling strategies on multi-class scatterplots. The main goal of this study is to understand the capability of these sampling methods in preserving the density, outliers, and overall shape of a scatterplot. To this end, we comprehensive review the literature and select 8 representative datasets as well as 7 typical sampling strategies. We then design four controlled experiments to understand the performance of different strategies in maintaining the 1) overall density; 2) relative density among different classes; 3) outliers; and 4) overall shape in the sampling results. The results show that 1) random sampling is preferred for preserving the overall and relative density; 2) outlier biased density based sampling and recursive subdivision based sampling perform the best in keeping more outliers; and 3) blue noise sampling outperforms the others in maintaining the overall shape of a scatterplot.

Index Terms—Scatterplot, data sampling, empirical evaluation.

1 Introduction

Scatterplots are one of the most widely used visual representations in exploratory data analysis [21, 32]. Their flexibility enables discovering free-form patterns in two dimensional data, such as trends, clusters, and outliers [12]. With the companion of dimensionality reduction approaches, scatterplots are also the dominant visualization tool to explore high-dimensional data [17, 31, 33]. However, scatterplots become less effective when data grows in size. First, the overdraw issue will affect the understanding of scatterplots [20]. Second, the speed of producing visualization, i.e., loading and rendering source data, will be a considerable issue [26].

To overcome the scalability issue in scatterplots, sampling has been well studied in data mining [26] and visualization [3, 5]. Generally, sampling aims to select a statistically unbiased representation of the full dataset. In different scenarios, many sampling strategies have been developed to enhance specific aspects of the full dataset, e.g., density [3], outlier [40], shape [16], and class ratio [5, 6, 10]. Rojas et al. [30] interviewed 22 data scientists and concluded that random sampling, which is statistically unbiased, is the only choice of these scientists for data exploration. Although other sampling strategies can provide different insights for data exploration, data scientists are not familiar with them and do not know which strategy to use in a specific scenario. For instance, although blue-noise sampling has been widely used in computer graphics and visualization, we have not found its application in data mining in our literature review.

Nevertheless, the researches on sampling strategy design have performed many quality comparisons. Performing a perception-based evaluation study is still essential to provide guidelines of choosing sampling strategies. On the one hand, most of existing comparisons are based on objective quality measures, e.g., density, class ratio, and number of outliers. Their conclusions may not be suitable for visualization tasks due to perceptual biases [19, 36]. On the other hand, these strategy-oriented evaluations are limited to a subset of tasks and approaches. Thus, a comprehensive evaluation of representative approaches is missing.

In this paper, we conduct four experiments to study the effects of typical sampling strategies on 2D scatterplots. First, we select seven widely used or task-specific sampling strategies by a comprehensive literature review. The strategies include random sampling [23], blue noise sampling [8], density biased sampling [25], multi-class blue noise sampling [35, 5], outlier biased density based sampling [40], multi-view Z-order sampling [10], and recursive subdivision based sampling [6]. Second, we identify four typical analytical tasks in multi-class scatterplots analysis, including identifying relative region density, relative class density, outliers, and shapes. Third, we formulate four hypotheses based on our experience and literature review. We hypothesize that (1) for a scatterplot without class information, all other sampling strategies perform better than random sampling in relative region density identification tasks in terms of accuracy and efficiency; (2) for a scatterplot with class information, multi-class sampling strategies perform better than other sampling strategies in relative class density identification tasks in terms of accuracy and efficiency (3) outlier biased density based sampling is the best in the outlier identification task; (4) blue noise sampling and multi-class blue noise sampling perform better than other strategies in preserving the overall shape.

We select eight datasets that present different patterns and various degrees of visual clutter. 100 participants are recruited for the formal experiments. Before the formal study, we perform a pre-study on 160 participants to determine the sampling ratio and color stimuli. In the formal study, we conduct a series of experiments on different sampling strategies and datasets and record the results in completing these experiments. We also design subjective questionnaires to obtain subjective experience of the participants.

Based on the experiment results, we perform a comprehensive statistical analysis. The analysis results of the two objective metrics suggest that (1) H1 is rejected; with random sampling, participants use shorter time to complete the region density identification tasks with higher accuracy; (2) H2 is partially confirmed; multi-class sampling strategies achieve higher accuracy than other strategies except for blue noise sampling; with random sampling, participants use shorter time to complete the class density identification tasks. (3) H3 is partially confirmed; outlier biased density based sampling, blue noise sampling and recursive subdivision based sampling perform better than other strategies in identifying outliers. (4) H4 is partially confirmed; blue noise sampling performs the best in shape preserving while multi-class blue noise sampling performs at a middle level. The analysis results of the subjective questionnaires provide useful insights into the sampling strategies. They disclose subjective reasons of the objective metric results. After the analysis, we summarize the ability of the seven sampling strategies to support our identified tasks.

In summary, we present a comprehensive perception-based evaluation of sampling strategies for scatterplots. We contribute a carefully designed evaluation and a series of instructive findings, which provide guidelines for choosing sampling strategies in task-specific scenarios. In addition, we also contribute a Python library for scatterplot sampling, which contains XXX commonly used sampling algorithms and is available at `http://shixialiu.com/libsampling/`.

2 Related Work

2.1 Sampling Strategies for Scatterplots

The scatterplot sampling methods can be categorized into two classes, single-class sampling and multi-class sampling.

Single-class sampling. This category of sampling strategies aims to preserve the properties of interest (e.g., density) of the original dataset without considering class information. Random sampling, the most widely used sampling method, is a classical single-class sampling method. It employs a uniform sampling strategy that treats all samples equally and selects each sample with the same probability.

On the contrary, non-uniform sampling strategies assign varying sampling probability to data so that some specific properties of the original datasets can be better preserved. For example, in some cases, samples are required to be better spatially separated [16, 42]. Blue noise sampling [42, 41] achieves this by selecting samples with blue noise properties so that the selected samples will distribute evenly in the sample space. Farthest point sampling [1] can also select samples with better space separation. It randomly picks the first sample, and then iteratively selects samples of maximal minimum distances to the previously selected ones. Liu *et al*. [16] developed a dual space sampling strategy. It computes a density field of the original sample space and maps the samples from the original density space to a uniform density space through a warping function. Then it selects the samples via orthogonal least squares or weight sample elimination in the mapped space in order to maintain good spatial separation among selected samples. Lastly, the selected samples are mapped back into the original density space.

There are also sampling strategies developed to preserve density-related properties. Density-biased sampling [25] tends to over-sample sparse regions and under-sample dense regions in the sample space. It can counterbalance samples from both regions, thus preserving small clusters and more solitary samples. Bertini *et al*. [2] proposed a non-uniform sampling strategy aiming at preserving the relative region density difference. It divides the sample space into uniform grids, then determines the represented density of each grid and finally selects samples from each grid according to the density. Joia *et al*. [11] formulated the sampling problem as a matrix decomposition problem and solved it with singular value decomposition (SVD). This method performs SVD on the original dataset and selects the samples with the biggest correlation with top-k basis vectors in the SVD result, where k is a rank parameter indicating the number of principal components of interest. The SVD based sampling strategy can counterbalance the number of points from regions with different densities.

Outlier preservation is another common goal in sampling strategies. A typical method for achieving this goal is to alter exisiting sampling strategies, making them probabilistically accept more outliers according to specified outlier scores [18, 40]. For instance, Liu *et al*. [18] proposed outlier biased random sampling that assigns higher sampling probabilities to outliers in random sampling. Similarly, Xiang *et al*. [40] increased the accepting probability of outliers in the sampling process of blue noise sampling and density biased sampling, thus developing outlier biased blue noise sampling and outlier biased density based sampling, respectively. Moreover, Cheng *et al*. [7] sampled the point clouds on their color mapping display using a hashmap based stratified sampling technique to preserve outliers while keeping the main distribution. This strategy maintains a 2D hashmap during the sampling process to reduce the samples selected in dense regions and include more samples in sparse regions.

Multi-class sampling. Unlike single-class sampling strategies, multi-class sampling strategies need to take the class labels of a dataset as input. They aim to preserve the properties of interest (e.g., density) of each individual class as well as their union. Wei [35] extended blue noise sampling to multi-class scenarios to maintain the blue noise properties of each class of samples and of the whole dataset. Based on the multi-class blue noise sampling, Chen *et al*. [5] employed a hierarchical sampling strategy that selects samples round by round. It first selects samples from the coarsest level using multi-class blue noise sampling, and when the selected samples are not enough, it reduces the restricted distance of the selected samples by half and adds more samples in the final result. Recently, a recursive subdivision based sampling strategy proposed by Chen *et al*. [6] meets several requirements in multi-class scatterplots exploration, including preserving relative densities, maintaining outliers, and minimizing visual artifacts. It splits the visual space into a binary KD-tree and determines which class of instances should be selected at each leaf node based on relative class density by a backtracking procedure. Additionally, Hu *et al*. [10] developed multi-view Z-order sampling based on Z-order curve methods [45] and formulated it as a set cover problem. The sets were constructed by segmenting the Z-order curves of the samples of each class and the whole dataset, respectively. This strategy selects samples by greedily solving such set cover problem, and gets satisfying results in terms of minimizing kernel density estimation error.

Sampling strategy	MC	NU	S	D	O
Random sampling [44, 9, 28, 39]					
Blue noise sampling [42, 40]		√	√		
Farthest point sampling [1]		√	√		
Dual space sampling [16]		√	√		
Density biased sampling [40, 25]		√		√	
Non-uniform sampling [2]		√		√	
SVD based sampling [11]		√		√	
Outlier biased random sampling [18, 43]		√			√
Outlier biased density based sampling [40]		√		√	√
Outlier biased blue noise sampling [40]		√	√		√
Hashmap based noise sampling [7]		√			√
Multi-class blue noise sampling [35, 5]	√	√	√		
Multi-view Z-order sampling [10]	√	√		√	
Recursive subdivision based sampling [6]	√	√		√	√

Table 1: Characteristics of our collected sampling methods. MC refers to multi-class sampling strategies; NU refers to non-uniform sampling strategies; S refers to considering spatial separation; D refers to considering density; O refers to considering outlier preservation.

2.2 Evaluation Studies of Sampling Methods

A few studies have paid attention to the evaluation of sampling methods before. However, they either evaluate the sampling methods in certain situations (e.g., graph sampling) or evaluate a specific sampling method to show its capability.

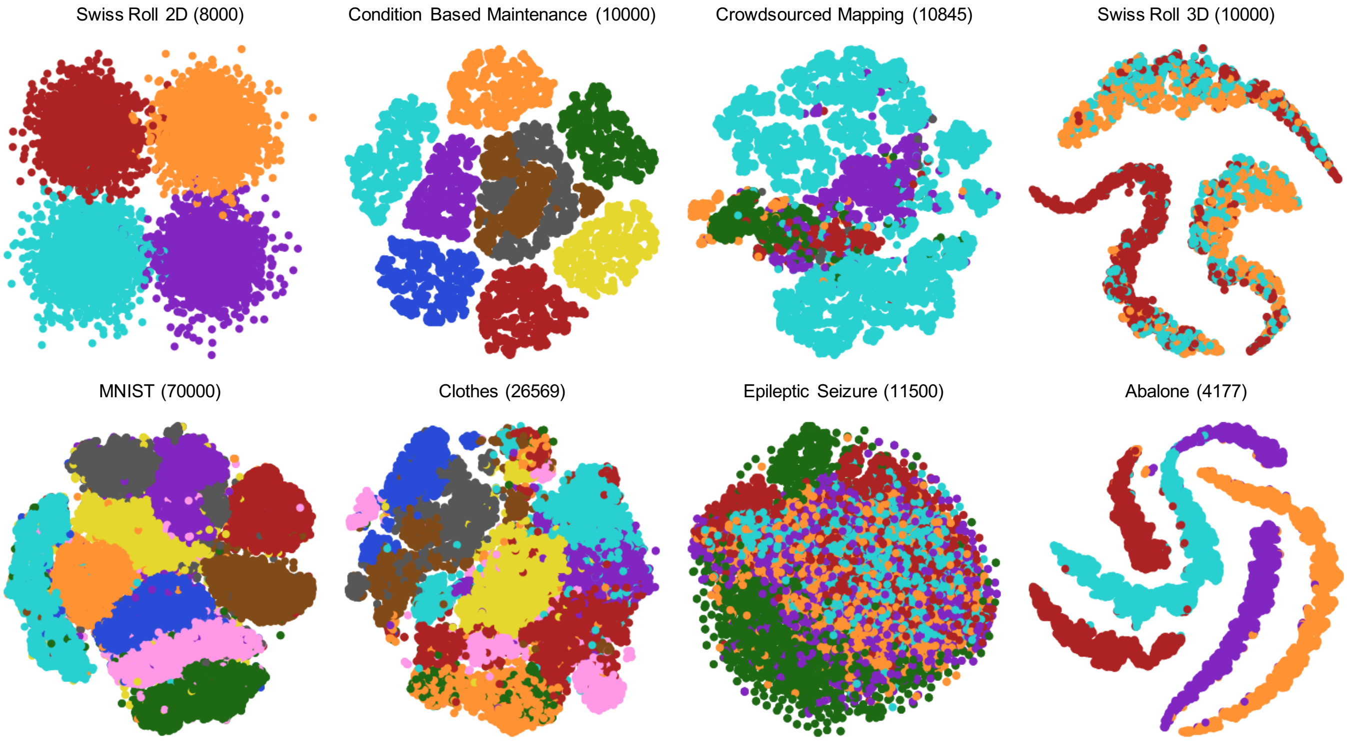


Fig. 2: Datasets selected for our evaluation. The numbers in the brackets indicate the sizes of the datasets.

Generic Evaluation. Previous studies concentrated on evaluating sampling methods for graph data [38, 24]. Wu *et al*. [38] conducted a survey on graph sampling methods and performed an empirical evaluation about the preservation of the three most important visual factors on five selected methods. Later, Nguyen *et al*. [24] proposed a family of quality metrics to evaluate the stochastic graph sampling methods in an objective manner.

Instance-oriented Evaluation. When proposing new sampling methods, researchers would also conduct evaluations to demonstrate their effectiveness. Some of them used quality metrics to make a quantitative evaluation in terms of data features. Chen *et al*. [6] designed four metrics based on their design requirements and compared their results with three baseline methods. They also presented three case studies to show the usefulness of their method in multi-dimension data analysis. However, as numerical measures do not always agree with human perception [34], other efforts focused on empirically evaluating perceptual subjects through user studies. For example, both hierarchical multi-class blue noise sampling [5] and multi-view Z-order sampling [10] showed their superiority on the recognition of data classes and densities.

To the best of our knowledge, there has never been a systematic evaluation of sampling on scatterplots from the perspective of visualization. Inspired by [38], in this paper, we collected the representative sampling strategies on scatterplots from the visualization society and then conducted four experiments to evaluate their ability to retain data features on perception.

3 Evaluation Landscape

Based on a comprehensive literature review, we select a set of representative sampling strategies, datasets, and visual factors to be evaluated.

3.1 Selection of Sampling Strategies

To comprehensively summarize the sampling methods used in visualization society, we first surveyed papers from the the journal of IEEE TVCG and three mainstream visualization conferences (IEEE VIS, PacificVis, and EuroVis) published in 2010 – 2019. We used Google Scholar to search for the papers with the keyword ”sampling” from the sources above. There were 1562 papers in the initial result. Next, we filtered out papers which are not relevant to sampling in visualization. We kept papers either that applied sampling for visualization purposes or that proposed new sampling strategies in visual analytics or in information visualization. Finally, 26 papers remained in our survey. We further summarized the sampling strategies discussed in these papers.

The collection of the sampling strategies are listed in Table 1. We decided to focus on the widely used and recent advanced task-specific sampling strategies since there are diverse sampling strategies used for different visualization purposes, and obviously, it is impractical to evaluate all in our work. As a result, the selected sampling strategies for our evaluation have covered all the categories listed in Table 1. We selected random sampling (RS) [23], because it is the most widely used sampling strategy. We also selected other representative strategies, including blue noise sampling (BNS) [8], density biased sampling (DBS) [25], and multi-class blue noise sampling (MCBNS) [35, 5]. Besides, outlier biased density based sampling (OBDBS) [40], multi-view Z-order sampling (MVZS) [10], and recursive subdivision based sampling (RSBS) [6] are those strategies that perform the best in terms of their design requirements, respectively. These seven strategies are all selected in our study.

3.2 Selection of Datasets

To ensure the reliability of the evaluation results, we selected datasets from the previous studies in visualization as our experiment data. More specifically, we collected the datasets that were used in the works in our survey. Since most of them are high-dimensional data, we first transform them into 2D space using t-SNE and normalized to $[0,1] \times [0,1]$. In the results, points are located as clusters in different shapes in the obtained multi-class scatterplots. In addition, the number of points and the clutter degrees of these scatterplots vary within a wide range. According to the observations above, we finally selected eight representative datasets shown in Fig. 2 with different characteristics: six datasets where points are located as clusters (*Swiss Roll 2D*, *Condition Based Maintenance*, *Crowdsourced Mapping*, *MNIST* [14], *Clothes* [40], and *Epileptic Seizure*); and two where points are located as curved stripes (*Swiss Roll 3D* and *Abalone*). The clutter degrees of them vary from slight to severe as the order listed in each bracket. The number of points in the selected datasets ranges from thousands to tens of thousands as listed in Fig. 2.

3.3 Selection of Visual Factors

We identified the most critical visual factors for the sampling methods by comprehensively reviewing existing works on sampling and scatterplots. Previous studies have shown that there are basically five goals related to the scatterplot exploration, outlier identification, shape examination, trend analysis, density detection, and coherence analysis [21, 37]. In order to determine which of these factors mentioned in the aforementioned goals are concerned in the existing sampling strategies, we carefully examine the 26 selected papers and extracted the visual factors that are considered in these works. Specifically, outlier maintenance is the most mentioned, which appears in seven papers out of the 26 papers. Three papers concern about preserving the relative density and two concern about the overall shape of a scatterplot, respectively. As a result, in the study, we decide to investigate the capabilities of different sampling strategies in preserving **outliers**, **density**, and the **overall shape** of a scatterplot.

4 Pre-Study

The purpose of the pre-study was to specify two experiment choices for the formal study: (1) how many points should be sampled from each dataset, and (2) whether color encoding should be used in the experiment of comparing region densities of multi-class scatterplots in the formal study. *Region density* refers to the density of data points regardless of class information. We conducted two experiments to answer the questions above, respectively.

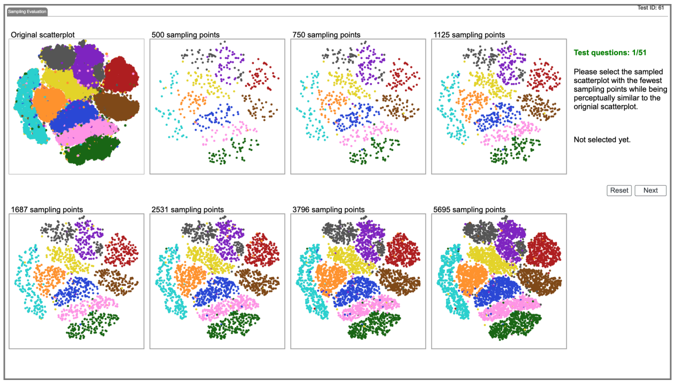
4.1 Experiment 1: Sampling Number Identification

We conducted a subjective experiment to specify the proper number of sampling points for each dataset. On the one hand, we would like to set a small sampling number to clearly show the motivation of sampling, i.e., addressing the issue of visual clutter. On the other hand, the patterns of the original scatterplots would not be preserved if the sampling number is too small. Because the patterns are different in different datasets, it is essential to choose a proper sampling number for each dataset.

Task and Procedure. We used the same eight datasets as those in the formal study. For each dataset, we showed the participants the original scatterplot as well as a series of sampled scatterplots with different sampling numbers. Participants were asked to select the sampled scatterplot that has the smallest number of points while being perceptually similar to the original scatterplot. Weber-Fechner Law states that the perceived intensity is proportional to the logarithm of the stimulus [29]. Therefore, we adopted a seven-level sampling with 500, 500×1.5, 500×1.5^2, ..., 500×1.5^6 points (a geometric sequence) by random sampling.We choose random sampling because it has no preference on certain property of dataset. Leskovec *et al.* [15] show that different sampling strategies produce very similar results when the sampling rate is greater than 50%. Therefore, we cut off the sampled scatterplots in a sequence if their sampling rate is greater than 50% to avoid meaningless comparison. As a result, there are seven sampled scatterplots (at most), along with the original scatterplot with all points, to be displayed. The eight scatterplots are arranged in a 2×4 matrix layout, as shown in Fig. 3. It took about 5 minutes for each participant to finish the experiment. Participants were asked which visual factors they are concerned in their judgments in this experiment in the post-experiment questionnaire.

Participants and Apparatus. We recruited 160 participants (130 males, 30 females, aged 18 - 60 years). All participants were either students or researchers with computer science background. 69 participants reported to be familiar or very familiar with visualization, 32 moderately familiar, and 61 unfamiliar or very unfamiliar. 65 participants reported previous experience with sampling.

The experiment was conducted through a web prototype. Participants were asked to perform the experiment on a screen with a resolution higher than 1920 × 1080. The points in the scatterplots were rendered with a radius of 3 pixels in a random order without transparency.



Test questions: 1/51

Please select the sampled scatterplot with the fewest sampling points while being perceptually similar to the orignial scatterplot.

Not selected yet.

Fig. 3: Interface of Experiment 1 in the pre-study: the original scatterplot is always located at the top-left with the sampling results of increasing sampling number.

Results. Given the fact that when there are more points in the sampled scatterplot, it will look more like the original scatterplot, we assume that participants will consider the sampling results still similar to the original scatterplot when the sampling number is more than the selected one. In addition, considering that other sampling strategies may perform better than random sampling, we need to leave the space to show their superiority in our experiments. Based on these considerations, we have chosen the optimal sampling number by requiring that the sampling numbers of the scatterplots selected by 80%, rather than 100%, of the participants was smaller than the optimal one. Fig. 4 presents the result, which shows that the optimal sampling number for most datasets (*MNIST*, *Swiss Roll 2D*, *Crowdsourced Mapping* and *Condition Based Maintenance*) is 2531, while the sampling rates of them are 3.6%, 31.6%, 23.3%, 25.3%, respectively. Four exceptions are the datasets *Clothes*, *Epileptic Seizure*, *Swiss Roll 3D* and *Abalone*, whose optimal sampling numbers with the corresponding sampling rates are 3796 (14.3%), 3796 (33.0%), 1687 (16.9%), and 1125 (26.9%), respectively. According to the results of subjective questionnaire in the experiment, when judging the similarity between scatterplots, over 75% of the participants took the overall shape of each class of points into consideration, followed by density (55%) and outliers (35%).

4.2 Experiment 2: Understanding Color Effect on Region Density Identification

This controlled experiment aims at understanding the effect of color when comparing region density in multi-class scatterplots. Though color is not related to the definition of region density, encoding class labels with color may affect the human perception. Therefore, although we prefer to perform the formal study in color-encoded scatterplots rather than in single-color scatterplots, we should figure out whether color affects the perception of region density.

Comments: merge the Task and Experiment Design as one paragraph. Split the part of Procedure as a separate paragraph.

Task and Experiment design. We generated ten synthetic datasets using mixed Gaussian distributions with 3 to 10 classes. These datasets were different from the ones in the formal study. In each question, two rectangular regions were marked on the same scatterplot and participants were asked to compare their region density and select the region which had higher density. This experiment adopted a within-subject design, and the only variable was whether the scatterplot was colored or not. We asked the same questions with multiple color or with only dark grey color. We provided two questions on each synthetic dataset, so there were 20 different questions. Thus, in total, we had

$$160\ (participants) \times 20\ (questions) \times 2\ (colors) = 6400$$

results. In order to eliminate the learning effect, the multi-colored and

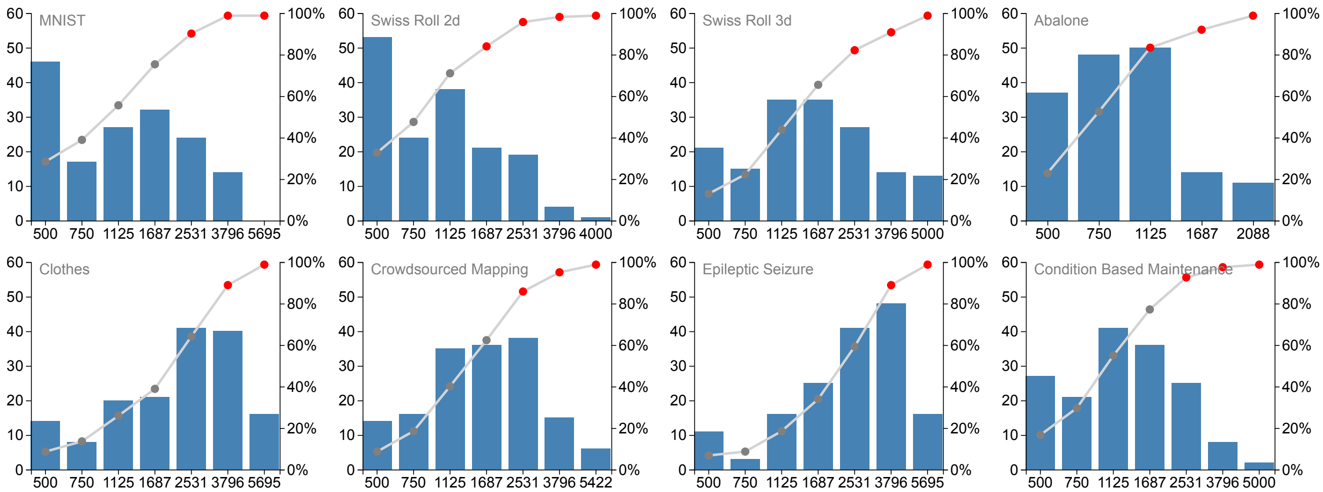


Fig. 4: Results of Experiment 1 in the pre-study: Sampling number identification. The bar indicates the number of participants selected the corresponding option, while the line indicate the cumulate proportion of participants selecting the option with a sampling number not more than the current one.

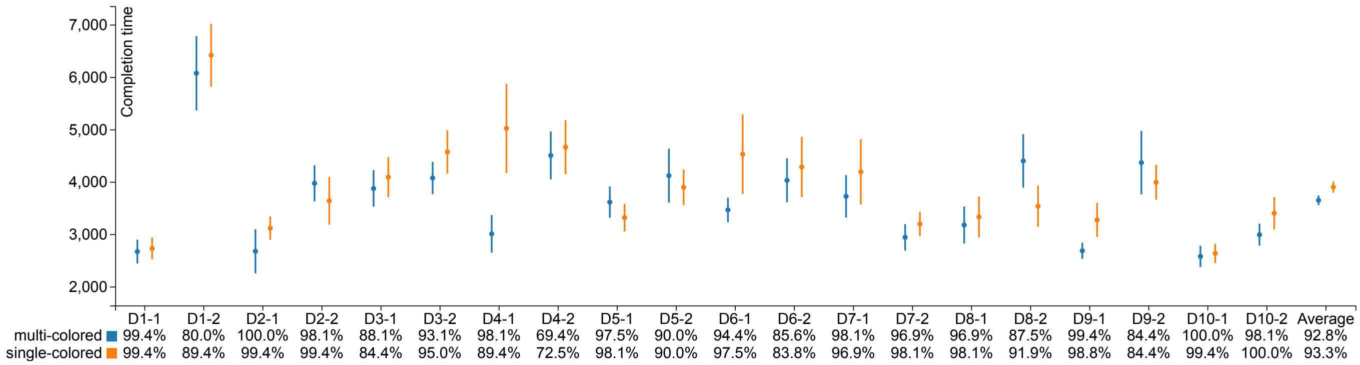


Fig. 5: Results of Experiment 2 in the pre-study: Understanding color effect on all-class density comparison. Error bars represent 95% confidence intervals. Below the x-axis listed the accuracy of the colored and the uncolored version of each question and all questions.

single-colored versions of the same question were arranged to appear in a random order and not consecutively. It took about 5 minutes for each participant to finish the experiment.

Procedure. In order to help the participants to get familiar with the task, the experiment started with a training session of three questions. In the training session, the correct answers were shown to the participants after they submitted their answers. Participants could ask questions during the training session. Time and accuracy were not recorded. As long as the participants reported that they had fully understood the experiment, we started the real study, where completion time and accuracy for each question were recorded. Thus, in the real study, we reminded participants that they needed to finish the experiment as fast and precisely as possible. After the experiment, participants were asked to finish a questionnaire and rate the color effect on a five-point Likert scale.

Participants and Apparatus. This part was the same as Experiment 1.

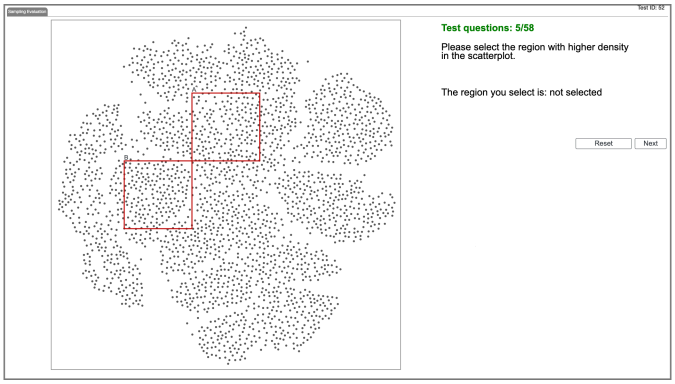
Results. The results are shown in Fig. 5. The average accuracy of the multi-colored and single-colored questions are 92.9% and 93.3%, respectively, while the average completion time of the multi-colored questions is 3581ms and that of the single-colored ones is 3616ms. Since the data are not subject to the normal distribution according to the Shapiro-Wilk test, we conduct a Wilcoxon test to check for the significance of their difference with the significance level $\alpha = 0.05$. No statistical significance in the difference of their accuracy is reported through hypothesis tests ($p = 0.2648 > 0.05$). But significant difference exists in terms of completion time ($p = 1.551 \times 10^{-10} < 0.001$), which means that participants spent significantly shorter time in completing single-colored questions than multi-colored ones. In the subjective questionnaires, the average score of the color deficiency is 2.03, indicating that the participants felt that color affected the region density comparison slightly. This is also coherent with our numerical results. A participant commented that *”the salient color may have an influence on my judgment of density”*. Consequently, we decided to use single-colored scatterplots in the experiment of all-class density comparison in the formal study to eliminate the color effect.

5 Formal Study

5.1 Hypotheses

The formal study aims to evaluate the performance of selected seven sampling strategies in the aspects of preserving three identified visual factors, including relative density, outlier, and overall shape. According to the three visual factors, we formulate four hypotheses to guide the experiments design. Specifically, we formulate two hypotheses on relative density in terms of *region density* and *class density*, respectively. *Region density* refers to the density of data points regardless of class information. *Class density* is the density of data points belonging to a certain class.

H1: All other sampling strategies perform better than random sampling in preserving relative region density.
Maintaining relative density is a common goal for many sampling



Test questions: 5/58

Please select the region with higher density in the scatterplot.

The region you select is: not selected

Fig. 6: Example interface of Experiment 1 in the formal study.

strategies [5, 6, 10]. Compared to the random sampling, these strategies are designed with delicate algorithms. They often report positive results when comparing with random sampling in different scenarios [6]. Therefore, we assume all other sampling strategies should perform better than random sampling in preserving relative region density.

H2: Multi-class adapted sampling strategies perform better than other sampling strategies in preserving relative class density. Many sampling strategies are customized for the scenario of multi-class scatterplots. In these strategies, preserving the individual class properties, e.g., density, is an important goal. Therefore, we assume that multi-class adapted sampling strategies, including multi-view Z-order sampling [10], recursive subdivision based sampling [6], and multi-class blue noise sampling [5], perform better than the other four sampling strategies in preserving relative class density.

H3: Outlier biased density based sampling is the best in preserving outliers. Many sampling strategies have shown their ability in preserving outliers in their reports. Among them, outlier biased density based sampling is designed specially for preserving outliers. It is the only strategy that integrates outlier measure into sampling process. Therefore, we assume that outlier biased density based sampling is the best strategy in preserving outliers.

H4: Blue noise sampling and multi-class blue noise sampling perform better than other strategies in preserving the overall shape. In our observation, uniform distribution facilitates the description of shapes by minimizing the effects of other visual factors, such as outliers and inhomogeneous density. Based on the observation, we assume that sampling strategies that aim to generate uniform samples with blue noise property, i.e., blue noise sampling and multi-class blue noise sampling, should perform the best in preserving the overall shape.

5.2 Experiments

Guided by the four hypotheses (**H1** – **H4**), we designed four experiments: Experiment 1 (**E1**) was designed for the perception of relative region density preservation (**H1**) and Experiment 2 (**E2**) was designed for the perception of relative class density preservation (**H2**); Experiment 3 (**E3**) was designed for the perception of outlier maintenance (**H3**); and Experiment 4 (**E4**) was for the perception of overall shape preservation (**H4**). Note that, **E1** – **E3** are controlled experiments and **E4** is a subjective experiment.

E1: Perception of relative region density preservation. This experiment was used to evaluate the ability of different sampling strategies in preserving relative region density in the aspect of visual perception. Specifically, in the experiment, we aimed to test if the region with higher region density can be still recognized as the higher one after sampling. Thus, in each question, we randomly marked out two rectangle regions with the size of $\frac{w}{5} \times \frac{w}{5}$, where w is the width of scatterplots. Participants were asked to select the region with higher density without considering class labels.

Based on the result of the pre-study, color would interfere and slow down the judgments of the participants. Thus, we rendered all data points in dark grey regardless of their labels. We had eight datasets and we generated two questions for each dataset. For each question, we generated seven trails corresponding to seven sampling strategies, respectively. In the seven trails of the same question, the locations of the rectangle regions were the same. In total, we had

$$7\ (sampling\ strategies) \times 8\ (datasets) \times 2\ (questions) = 112$$

trials for each participant.

E2: Perception of relative class density preservation. This experiment was used to test whether the class have higher density can be still recognized as the higher one after sampling. In contrast to **E1**, **E2** focuses on preserving of relative density of specific classes in the same region, instead of relative region density. Thus, the scatterplots are rendered using color to encode class labels. Specifically, in each question, we marked out a rectangle region with the size of $\frac{w}{5} \times \frac{w}{5}$ in a scatterplot. We specified two classes in the question and the participants were asked to choose the class with higher average density in the marked region. Similar to **E1**, we generated two questions for each dataset and seven trails for each question. In total, we had

$$7\ (sampling\ strategies) \times 8\ (datasets) \times 2\ (questions) = 112$$

trials for each participant.

E3: Perception of outlier maintenance. This experiment was used to evaluate the ability of sampling strategies in preserving outliers in the aspect of perception. We tested: (1) whether an outlier in the original dataset can be still preserved and perceived as an outlier after sampling; and (2) in case a point is perceived as an outlier after sampling, whether it is indeed an outlier in the original dataset.

There are diverse definitions of outlier, however, here we focused on the outlier in two scenarios: first, when considering class labels, the point that is of different class with its neighboring points; second, when not considering class labels, the points which are located at abnormal distances from its class. We followed the definition in the class purity algorithm [22] in the first scenario and followed the local outlier factor algorithm [27] in the second scenario.

In each question, we marked out a region in a scatterplot. Participants were asked to mark the outliers in case they judge the outliers existed in the marked region. Note that, the outliers were referring to all points in the entire scatterplot, instead of the marked region. We had considered three ways for participants identifying outliers: first, marking out all the outliers in the entire scatterplot; second, marking out a specified number of outliers (e.g., 10) in the entire scatterplot; and third, marking out all outliers in a given rectangle region. Considering the huge size of outliers in the entire dataset, it is not feasible to mark all out in the entire range in the limited experiment time. Besides, the accuracy would be very low since a lot of outliers would be missed. If we limit the number of target outliers as noted in the second option, due to the big number of outliers, participants may easily mark out all requested number of outliers. So that the accuracy would be high for all strategies and it would be hard to distinguish the performances of different sampling strategies. Thus, we finally chose the third option and asked the participants to mark all the outliers out in a fixed range. The only disadvantage of this option was that participants might misselect outliers referring to the local distribution. To avoid this, we reminded the participants to refer to global distribution and we also corrected the observed errors in the training session.

In total, we had

$$7\ (sampling\ strategies) \times 8\ (datasets) = 56$$

trials for each participant in this experiment.

E4: Perception of overall shape preservation. This experiment was used to compare the abilities of sampling strategies in preserving the overall shape of scatterplots in terms of visual perception. In contrast to **E1**–**E3**, **E4** was a subjective experiment. In each trial, we sampled a dataset using seven strategies and displayed these seven sampling results together with the original scatterplot (see Fig. 3). Participants were asked to rank the seven sampling results based on the shape similarities between the sampling results and the original scatterplot. Participants were reminded that class labels should be taken into account in comparing the shape similarities. Parallel ranking was allowed when participants could not distinguish the difference among the sampling results. In total, each participant had eight trials for the eight datasets in this experiment.

5.3 Participants, Apparatus and Testing Data

Participants. We recruited 100 participants (78 males, 22 females, aged 18–50 years, average: 24) for the formal study. 16 of them are researchers in visualization and computer graphics. The others are undergraduate or graduated students majoring in computer science. 34 participants reported previous experience with sampling. None of them reported color blindness or color weakness. Each participant was rewarded $20 per hour for completing the experiments.

Apparatus. The experiments were conducted online and a web prototype was implemented for the formal user study (see Figure 6). Participants were required to visit it remotely on the Chrome browser and finish the experiment on a screen with a resolution of 1920 × 1080. They were asked to share their screen with the instructor during the experiments to enable remote monitoring.

Testing data. We generated scatterplots based on the selected eight datasets for the experiments. For each dataset, we created one scatterplot of the original dataset and seven scatterplots of sampling results by the seven sampling strategies, respectively. The sampling rates of each dataset were determined by the results of the pre-study. Since multi-view Z-order sampling and recursive subdivision based sampling cannot set the exact sampling rate, we controlled the error in 1%. The points were rendered with a radius of 3 pixels without transparency. The size of the scatterplots was 1000 × 1000 pixels in **E1**–**E3**, and 300 × 300 pixels in **E4**. To avoid imbalanced occlusion between classes, the points in the scatterplots were rendered in a random order. Except for **E1**, we selected Boynton's color palette [4] to encode classes in the scatterplots. All scatterplots were generated in advance. For the training session, we generated synthetic datasets following the Gaussian mixed distribution. In the real testing session, the order of all questions were counterbalanced by following a Latin square to avoid the learning effect.

5.4 Procedure

Each experiment included a training session and a real test session. In the beginning of the training session, the instructor explained the experiments as well as the related concepts (e.g., outliers). After the explanation, several practice trails (three for **E1**–**E3**, and one for **E4**) were presented to help participants get familiar with the experiments. For controlled experiments **E1**–**E3**, the correct answers were shown to the participants after they submitted their answers. The participants were encouraged to ask questions in the training sessions to facilitate their understanding of the experiments. After they reported that they have fully understood the tasks, we started the real test session. Participants were allowed to have a break of five minutes before each experiment. After participants completed all the experiments, they were asked to answer a questionnaire for their backgrounds and subjective feedback on the experiments. The entire process lasted approximately one hour and 20 minutes for each participant.

The questionnaire included three parts. First, we asked participants' backgrounds and basic information, familiarity with visualization, and experience with sampling strategies. Second, participants were asked to rate the importance of preserving relative density, outliers, and overall shape for a sampling strategy using a five-point Likert scale. They were also encouraged to add extra abilities that a sampling method should provide. At last, we asked their focus in each experiment in order to learn the important visual factors for human perception.

6 Experimental Results

6.1 Analysis Approach

We recorded the objective measurements from **E1**–**E3** and the subjective measurement from **E4**. For **E1** and **E2**, we recorded the correctness and completion time of each trail. For **E3**, we calculated the precision and recall of each trial. In each trail, we denote the set of outliers marked out by a participant as M and the ground truth as N. The precision is the ratio of $|N \cap M|$ to $|M|$, and the recall is the ratio of $|N \cap M|$ to $|N|$, where $|\cdot|$ denotes to the cardinality of a finite set. Note that the recall refers to the ratio of outliers that are preserved by sampling and then perceived by participants. To avoid small values, we normalized the recall by the maximal outlier preserving ratio among seven sampling strategies on each dataset. Without loss of clarity, we use the term *recall* to refer to normalized recall in the rest part of paper. For **E4**, we record the ranking of each sampling strategy in each trail and transferred the rankings into scores. Specifically, the 1st–7th sampling strategies get 7–1 points, respectively.

Following the common means in evaluating user performance [13], we reported the mean value and confidence interval of the objective measurements and performed significance analysis to test our hypotheses. For the results of **E1**–**E3**, we performed Shapiro-Wilk test and found that they do not follow the normal distribution. Therefore, we employed a non-parametric method to examine whether significant differences exist among the sampling strategies. Specifically, we chose the Friedman test with a standard significance level $\alpha = 0.05$ in our analysis. If there was significant differences, we conducted Conover test as the post-hoc test to examine the pairwise significance. In **E4**, we also reported the mean values and the confidence interval of the rating scores for sampling strategy.

6.2 Objective Results Analysis

H1: We assume that all other sampling strategies perform better than random sampling in preserving relative region density. This hypothesis is rejected.

Fig. 7 shows the results of **E1**. Overall, random sampling have the highest accuracy (98.63%) and the shortest completion time (2904*ms*) in **E1**. The Friedman tests show that statistical significance among different sampling strategies exists in accuracy ($\chi^2(6) = 13.56, p = 0.0349$) and average completion time ($\chi^2(6) = 20.28, p = 0.0025$) in **E1**. Fig. 8 depicts the pairwise significance relationships between each pair of sampling strategies in terms of accuracy. Random sampling performs significantly better than multi-class blue noise sampling ($p = 0.0051$) and outlier biased density based sampling ($p = 0.0232$) in accuracy. Fig. 9 depicts the pairwise significance relationships in terms of average completion time. Random sampling performs significantly better than multi-view Z-order sampling ($p = 0.0328$), multi-class blue noise sampling ($p = 0.0011$), outlier biased density based sampling ($p = 0.0221$), and recursive subdivision based sampling ($p = 0.0048$) in average completion time. No sampling strategy performs significantly better than random sampling either in either accuracy or in average completion time.

H2: We assume that multi-class adapted sampling strategies, including multi-class blue noise sampling, multi-view Z-order sampling, and recursive subdivision based sampling, perform better than random

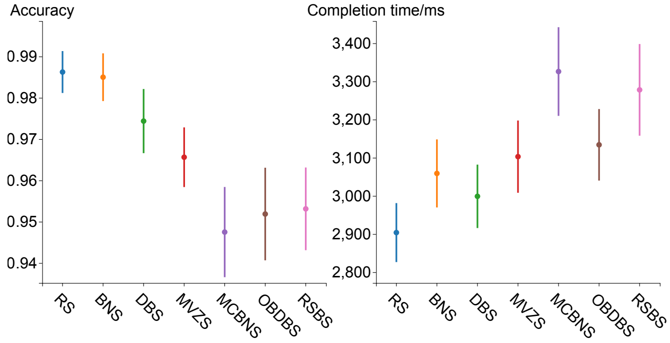


Fig. 7: Average accuracy and completion time of Experiment 1 in the formal study: Perception of region density. Error bars represent 95% confidence intervals.

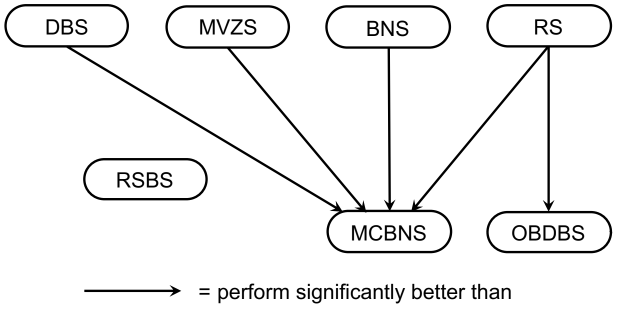


Fig. 8: Directed acyclic graph depiction on the pairwise significance relationships of the accuracy differences of the sampling strategies in **E1**. An edge indicates that the origin sampling strategy performs significantly better than the destination one. Same as below.

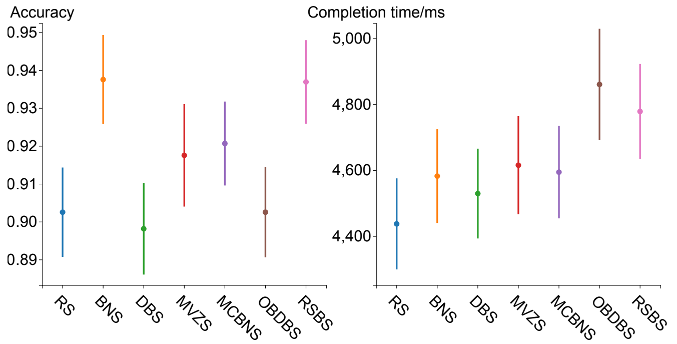


Fig. 10: Average accuracy and completion time of Experiment 2 in the formal study: Perception of class density. Error bars represent 95% confidence intervals.

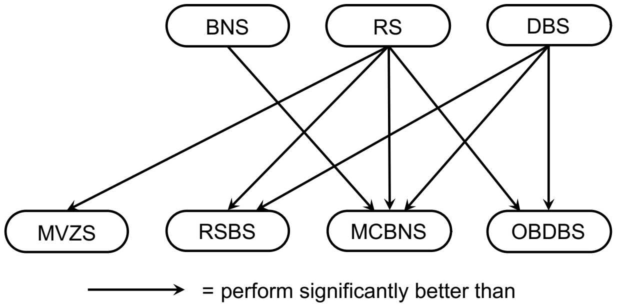


Fig. 9: Directed acyclic graph depiction on the pairwise significance relationships of the completion time differences of the sampling strategies in **E1**.

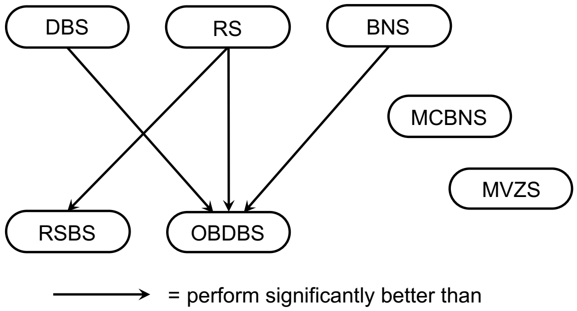


Fig. 11: Directed acyclic graph depiction on the pairwise significance relationships of the completion time differences of the sampling strategies in **E2**.

sampling, blue noise sampling, density biased sampling and outlier biased density based sampling in preserving relative class density. **H2** is partially confirmed.

The results of **E2** are displayed in Fig. 10. The blue noise sampling have the highest accuracy at 93.75%. The accuracies of recursive subdivision based sampling (93.69%), multi-class blue noise sampling (92.06%) and multi-view Z order sampling (91.75%) are higher than the remaining three strategies. However, difference significance is not found in accuracy according to the Friedman test ($\chi^2(6) = 4.019, p = 0.6741$).

In the aspect of average completion time, random sampling performs the best. The three multi-class adapted sampling strategies present no advantage than other strategies. The Friedman test shows that significant difference exists in completion time ($\chi^2(6) = 12.78, p = 0.0467$). Fig. 11 shows the pairwise significance relationships in terms of average completion time. Random sampling performs significantly better than recursive subdivision based sampling ($p = 0.0221$). Besides the hypothesis testing, we also find that outlier biased density based sampling performs significantly worse than density based sampling ($p = 0.0270$), random sampling ($p = 0.0095$), and blue noise sampling ($p = 0.0270$).

H3: We assume that outlier biased density based sampling is the best in preserving outliers. **H3** is partially confirmed.

As shown in Fig. 12, outlier biased density based sampling is ranked at the fourth in term of precision. Blue noise sampling, multi-class blue noise sampling, and recursive subdivision based sampling have higher accuracy than outlier biased density based sampling. However, the Friedman test presents no significant differences in precisions of the sampling methods ($\chi^2(6) = 10.53, p = 0.1040$). In the aspect of recall, outlier biased density based sampling is ranked at the second, while recursive subdivision based sampling has the highest recall. As listed in Table **??**, outlier biased density based sampling and recursive subdivision based sampling tend to sample more outliers, which contributes to their high recall. The Friedman test shows that significant differences exist ($\chi^2(6) = 18.78, p = 0.0045$). The post-hoc tests show that outlier biased density based sampling performs significantly better than random sampling and density based sampling (Fig. 13). Fig. 13 shows the discovered pairwise significance relationships. Besides the hypothesis test, we also have some interesting findings. First, the blue noise sampling has the highest precision and the third high recall. Although it has no significance relationship with other strategies in terms of precision and recall, it is worth to be recommended in the scenario of outlier preserving besides outlier biased density based sampling and recursive subdivision based sampling. In contrary, the random sampling and density based sampling have relative lower precision and recall than other strategies. The significance difference exist between them and outlier biased density based sampling, multi-class blue noise sampling, and recursive subdivision based sampling.

H4: We assume that blue noise sampling performs the best in preserving the overall shape. This hypothesis is also partially confirmed by the subjective experiment **E4**.

Fig. 14 shows the average ranking scores of sampling strategies in eight datasets and their average. Blue noise sampling has the highest ranking score in all eight datasets with an average score of 6.37, while the performance of multi-class blue noise sampling is in middle level. Moreover, the recursive subdivision based sampling, outlier biased density based samplings, and multi-class blue noise sampling have similar ranking scores and ranked at the 2nd, 3rd, and 4th with average scores of 4.82, 4.74, 4.27, respectively. The ranking is stable across eight datasets.

6.3 Subjective Results Analysis

Here we discuss the visual factors that most affect the participants' perception on relative density, outliers and overall shapes. More than half of them mentioned that the distributed area and the distances

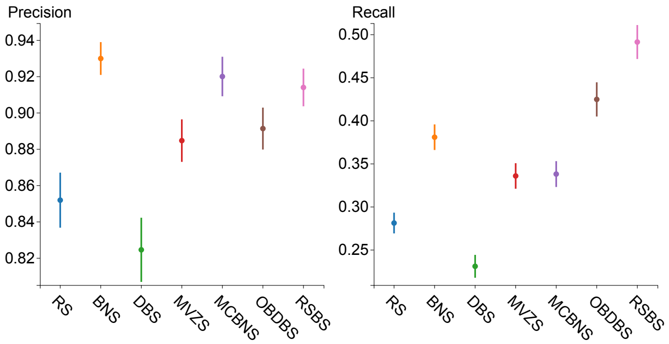


Fig. 12: Precision and recall of Experiment 3 in the formal study: Perception of outlier maintenance.

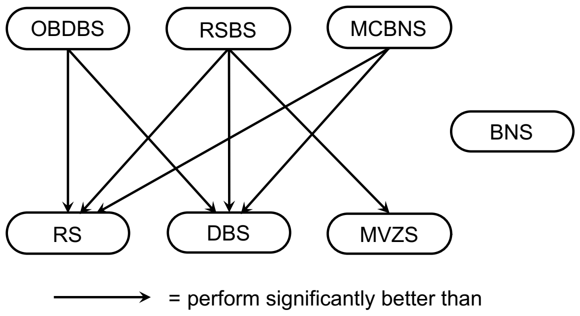


Fig. 13: Directed acyclic graph depiction on the pairwise significance relationships of the recall differences of the sampling strategies in **E3**.

between samples affect their judgments on the region density in **E1**. As for class density recognition, in addition to the previous mentioned factors, participants said that the color and the distribution of the two classes to be compared would also affect their judgments. In **E3**, some researchers suggested that the continuity of the points of a class would influence their judgments of whether a point was located in the main distribution of its class, thus affecting their judgments of outliers. Lastly, people commented that when comparing the preservation of overall shapes, the outline, the hollows, and the size of the shapes are the most important visual factors in their criterion.

We also collected the visual factors that participants would be most concerned in sampling in the questionnaire. The result is displayed in Fig. 15. For the three visual factors evaluated in our experiments, participants favored preserving overall shapes most in the sampling for scatterplots, then relative density, and outliers last. They also pointed out that such rating greatly depends on the specific analysis tasks in practical use. Besides, when asked for other important visual factors, fifteen of them mentioned that they were interested in the areas where points of different classes mix with each other, while another three participants would focus on the clutter degree in these areas specifically.

The feedback on the experiments about the influence factors helps us better understand the performance of the participants in the experiments, which is also beneficial to the further experiment design. Besides, the rating results on our three selected visual factors confirm that the evaluation targets of our experiments make sense, while the extra important visual factors raised by participants may shed light on new design requirements of sampling. Both of them are helpful to develop sampling methods from a perception-driven perspective.

7 Discussion

Reflection on the experiment design. In the subjective questionnaires, participants pointed out many factors that affected their perception of certain visual factors. For instance, the overlap of the rendered points may hide some outliers in the sampling results in **E3**. However, it is beyond the acceptable scale of the experiments if including so many factors. Although it is a limitation that the experiments do not consider the effect of these factors, we have plentiful trials with different datasets in the experiments so that we can reduce the bias caused by these confounding factors to some extent.

TODO: Discuss participants' background

Reflection on the experiment results. When we examined the sampling results of different sampling strategies, we found that outlier biased density based sampling and recursive subdivision based sampling tend to sample more outliers than other sampling strategies. In fact, they have the highest outlier retention ratio, and this contributes to their high recall. The results in **E2** show that participants spent most time in relative class density identification with these two sampling strategies. It can also be accounted by the observation that the scatterplots of their sampling results look more chaotic with so many outliers.

In Experiment 4, an interesting finding is that the better-performing sampling method, i.e. blue noise sampling, outlier biased density based sampling, multi-class blue noise sampling, and recursive subdivision based sampling, tend to sample the boundary points of a class or between different classes. It may contribute to the presevation of overall shapes since such boundary points help to outline the shape of the distribution of the points of each class.

Future work. In recent years, the newly developed sampling techniques mostly consider 2D data. However, as dimension reduction can be a time-consuming process sensitive to the size of data, it will be more efficient to first sample the data and then reduce their dimensionality when sampling and visualizing high-dimensional data. Apart from the sampling methods related to the visual space like multiview Z order sampling and recursive subdivision based sampling, other sampling methods evaluated can be also applied on high-dimensional data. A promising future work is to explore the perception effect when high-dimensional data are first sampled and then projected to the 2D scatterplots.

8 Conclusion

In this paper, we presented an empirical evaluation of sampling methods for scatterplots from the perspective of perception. We identified seven representative sampling strategies and three critical visual factors for scatterplots following a comprehensive survey on the existing literature. Based on the results, we formulated four hypotheses and designed four experiments to evaluate the ability of our selected sampling strategies to preserve the identified visual factors. We first conducted a pre-study to determine the proper sampling number of each dataset and confirm the negative effect on region density identification caused by color.

References

[1] M. Berger, K. McDonough, and L. M. Seversky. cite2vec: Citation-driven document exploration via word embeddings. *IEEE transactions on visualization and computer graphics*, 23(1):691–700, 2016.

[2] E. Bertini and G. Santucci. By chance is not enough: preserving relative density through nonuniform sampling. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 622–629. IEEE, 2004.

[3] E. Bertini and G. Santucci. Give chance a chance: Modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5:110 – 95, 2006.

[4] R. M. Boynton. Eleven colors that are almost never confused. In B. E. Rogowitz, editor, *Human Vision, Visual Processing, and Digital Display*. SPIE, Aug. 1989.

[5] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, and K. Ma. Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1683–1692, 2014.

[6] X. Chen, T. Ge, J. Zhang, B. Chen, C. Fu, O. Deussen, and Y. Wang. A recursive subdivision technique for sampling multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):729–738, 2020.

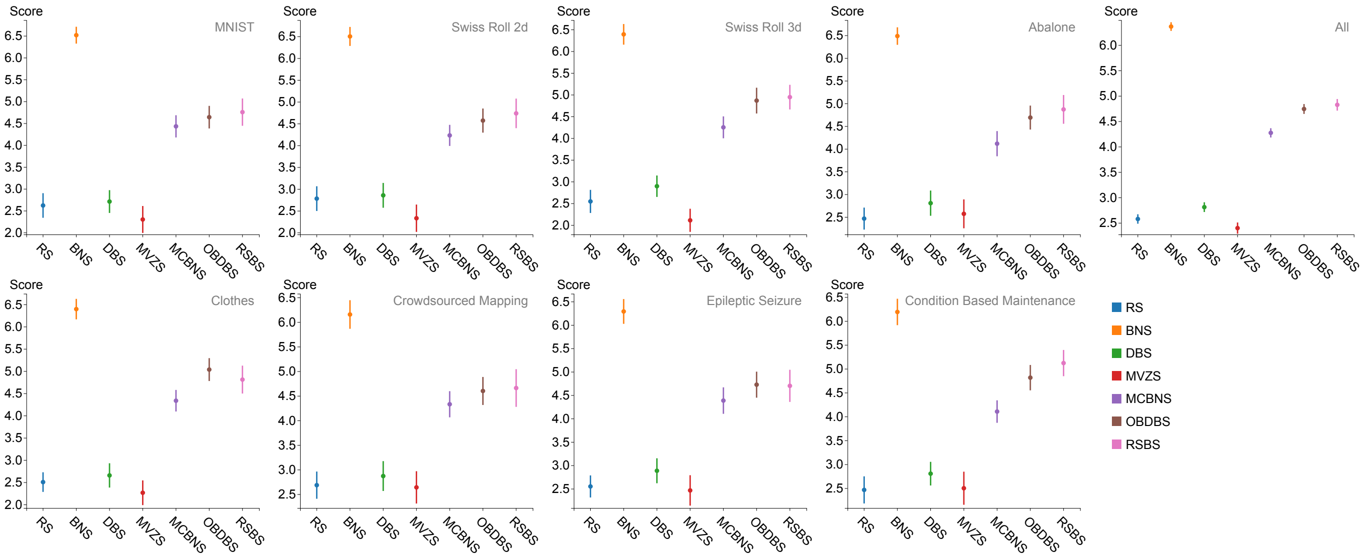


Fig. 14: Average rating scores of Experiment 4 in the formal study: Perception of overall shapes.

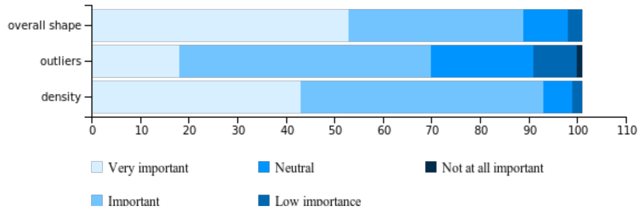


Fig. 15: The importance rating of the three visual factors evaluated in our experiments.

[7] S. Cheng, W. Xu, and K. Mueller. ColorMapND: A data-driven approach and tool for mapping multivariate data to color. *IEEE Transactions on Visualization and Computer Graphics*, 25(2):1361–1377, Feb. 2019.

[8] R. L. Cook. Stochastic sampling in computer graphics. *ACM Transactions on Graphics (TOG)*, 5(1):51–72, Jan. 1986.

[9] E. P. dos Santos Amorim, E. V. Brazil, J. Daniels, P. Joia, L. G. Nonato, and M. C. Sousa. iLAMP: Exploring high-dimensional spacing through backward multidimensional projection. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Oct. 2012.

[10] 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*, 26(1):739–748, 2020.

[11] P. Joia, F. Petronetto, and L. Nonato. Uncovering representative groups in multidimensional projections. *Computer Graphics Forum*, 34(3):281–290, June 2015.

[12] E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 73–82, 2012.

[13] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Sept. 2012.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 06*. ACM Press, 2006.

[16] L. Liu, A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2165–2178, Sept. 2017.

[17] S. Liu, C. Chen, Y. Lu, F. Ouyang, and B. Wang. An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):235–245, 2019.

[18] S. Liu, J. Xiao, J. Liu, X. Wang, J. Wu, and J. Zhu. Visual diagnosis of tree boosting methods. *IEEE transactions on visualization and computer graphics*, 24(1):163–173, 2017.

[19] Y. Ma, A. K. H. 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*, pages 1–1, 2018.

[20] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, 2013.

[21] 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.

[22] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira. Graph-based measures to assist user assessment of multidimensional projections. *Neurocomputing*, 150:583–598, Feb. 2015.

[23] P. Mukhopadhyay. *Theory and methods of survey sampling*. PHI Learning Pvt. Ltd., 2008.

[24] Q. H. Nguyen, S.-H. Hong, P. Eades, and A. Meidiana. Proxy graph: Visual quality metrics of big graph sampling. *IEEE transactions on visualization and computer graphics*, 23(6):1600–1611, 2017.

[25] C. R. Palmer and C. Faloutsos. Density biased sampling: An improved method for data mining and clustering. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 82–92, 2000.

[26] Y. Park, M. J. Cafarella, and B. Mozafari. Visualization-aware sampling for very large databases. *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 755–766, 2015.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[28] J. Poco, R. Etemadpour, F. Paulovich, T. Long, P. Rosenthal, M. Oliveira, L. Linsen, and R. Minghim. A framework for exploring multidimensional data with 3d projections. *Computer Graphics Forum*, 30(3):1111–1120, June 2011.

[29] R. Portugal and B. F. Svaiter. Weber-fechner law and the optimality of the logarithmic scale. *Minds and Machines*, 21(1):73–81, 2011.

[30] J. A. R. Rojas, M. Beth Kery, S. Rosenthal, and A. Dey. Sampling techniques to improve big data exploration. In *IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)*, pages 26–35, 2017.

[31] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, 2017.

[32] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018.

[33] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE transactions on visualization and computer graphics*, 19(12):2634–2643, 2013.

[34] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE transactions on visualization and computer graphics*, 26(1):759–769, 2019.

[35] L.-Y. Wei. Multi-class blue noise sampling. In *ACM Transactions on Graphics (TOG)*, volume 29, page 79. ACM, 2010.

[36] Y. Wei, H. Mei, Y. Zhao, S. Zhou, B. Lin, H. Jiang, and W. Chen. Evaluating perceptual bias during geometric scaling of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):321–331, 2020.

[37] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE.

[38] Y. Wu, N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui. Evaluation of graph sampling: A visualization perspective. *IEEE transactions on visualization and computer graphics*, 23(1):401–410, 2016.

[39] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A. K. Tung. LDSScanner: Exploratory analysis of low-dimensional structures in high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):236–245, Jan. 2018.

[40] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu. Interactive correction of mislabeled training data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2019.

[41] D.-M. Yan, J.-W. Guo, B. Wang, X.-P. Zhang, and P. Wonka. A survey of blue-noise sampling and its applications. *Journal of Computer Science and Technology*, 30(3):439–452, 2015.

[42] J. Zhang, E. Yanli, J. Ma, Y. Zhao, B. Xu, L. Sun, J. Chen, and X. Yuan. Visual analysis of public utility service problems in a metropolis. *IEEE transactions on visualization and computer graphics*, 20(12):1843–1852, 2014.

[43] X. Zhao, W. Cui, Y. Wu, H. Zhang, H. Qu, and D. Zhang. Oui! outlier interpretation on multi-dimensional data via visual analytics. *Computer Graphics Forum*, 38(3):213–224, June 2019.

[44] Y. Zhao, X. Luo, X. Lin, H. Wang, X. Kui, F. Zhou, J. Wang, Y. Chen, and W. Chen. Visual analytics for electromagnetic situation awareness in radio monitoring and management. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):590–600, Jan. 2020.

[45] Y. Zheng, J. Jestes, J. M. Phillips, and F. Li. Quality and efficiency for kernel density estimates in large data. In *Proceedings of the 2013 international conference on Management of data - SIGMOD 13*. ACM Press, 2013.