Year: 2023

Visual-assisted Outlier Preservation for Scatterplot Sampling

Yang, Haiyan ; Pajarola, Renato

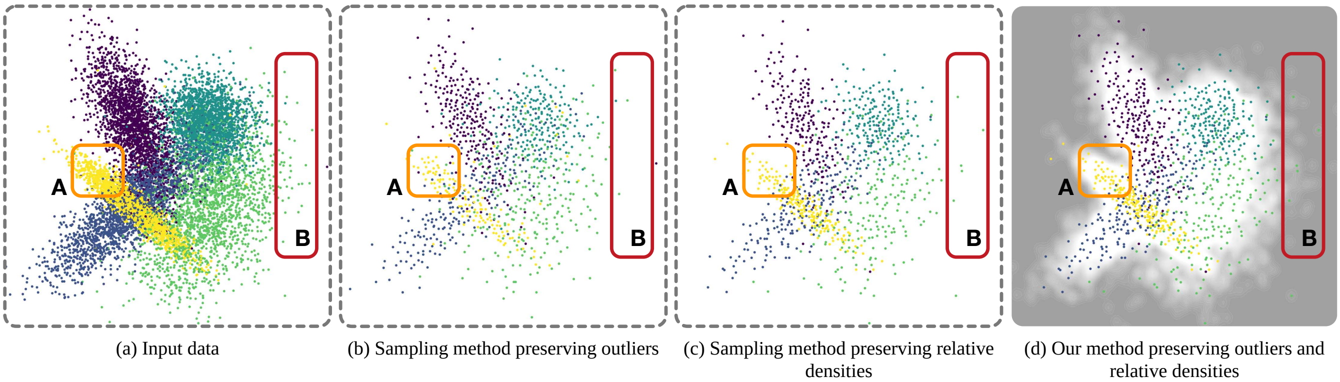
DOI: https://doi.org/10.2312/vmv.20231233

Vision, Modeling, and Visualization (2023)
T. Grosch and M. Guthe (Eds.)

Visual-assisted Outlier Preservation for Scatterplot Sampling

Haiyan Yang and Renato Pajarola[†]

University of Zürich, Switzerland



(a) Input data (b) Sampling method preserving outliers (c) Sampling method preserving relative densities (d) Our method preserving outliers and relative densities

Figure 1: *Our proposed outlier preservation method is intended to solve the existing inherent dilemma in scatterplot sampling methods. (a) is an input overplotted scatterplot. (b) shows one type of sampling approach that chooses to keep more information from the less represented data points at the expense of losing relative data densities contrast as shown in the highlighted region* **A***. whereas (c) presents another type of sampling result that maintains the relative data densities contrast by sampling all the data points with the same probability. (d) Our method adds an additional density map which captures the density distribution of low density points on top of (c).*

Abstract

Scatterplot sampling has long been an efficient and effective way to resolve the overplotting issues commonly occurring in large-scale scatterplot visualization applications. However, it is challenging to preserve the existence of low-density points or outliers after sampling for a sub-sampling algorithm if, at the same time, faithfully representing the relative data densities is of importance. In this work, we propose to address this issue in a visual-assisted manner. While the whole dataset is sub-sampled, the density of the outliers is modeled and visually integrated into the final scatterplot together with the sub-sampled point data. We showcase the effectiveness of our proposed method in various cases and user studies.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization; Visualization techniques;***

1. Introduction

Scatterplots are a widely used visualization technique in various data analytics and data visualization applications [Mun14, MPOW17, SG18, BBK*18]. By plotting the data points as dots in a 2D space with their attributes' values representing the corresponding 2D locations, scatterplots are well-known to be effective in uncovering the prevalent data distributions or clusters, thus further helping users to discover the correlations between any two attributes, or any two dimensions derived from dimension reduction techniques. However, when the number of data points increases, the possibility that many data points have close or virtually equivalent coordinates and may overlap with each other will also increase, which is referred to as *overplotting* issue [MG13].

Various methods have been proposed in the past to deal with overplotting in scatterplots [ED07], and they can roughly be categorized into two types: one is to alter the appearance of the data marks, such as mark size [WLS98, LvWM09, LMvW10], mark transparency [LvWM10, FKLT10], or mark shape [KW13]. An-

[†] Authors emails: {haiyan,pajarola}@ifi.uzh.ch

DOI: 10.2312/vmv.20231233

other one is sub-sampling. That is, reducing the number of data points plotted while still maintaining the density structure of the overall data distribution [PCM16, RRBKRD17].

To achieve better data density preservation for scatterplot sub-sampling, two challenges arise: The first is how to ensure that the sampling result can faithfully represent the original data distribution. And second, how to deal with any outliers (i.e., data points located in very low-density regions [BKNS00]) such that both the context of the outliers and the relative densities contrast [BS05, BS06, BBK*18, CGZ*20, CZF*22] between the inliers and the outliers are preserved. Most of the state-of-the-art scatterplot sampling methods focus more on tackling the first challenge, and treat outliers either the same as inliers [CCM*14] or emphasize their presence [LXL*18, CGZ*20, HSVK*20] in the sampling process. However, there is still a lack of systematical discussions on the second challenge. Specifically, there is a commonly ignored issue for any existing outlier preservation methods: If the outliers are kept with certain priorities, the final scatterplot tends to lose the relative data densities contrast between inliers and outliers (see highlighted region **A** in Fig. 1(b)). On the other hand, if outliers are sampled, such as to keep the data densities balanced among all data points, users lose the context of outliers in the low-density areas after sub-sampling (see highlighted region **B** in Fig. 1(c)). How to effectively tackle this dilemma becomes more urgent in time-series data sampling applications since the representation of outliers could be helpful for data (and outliers) structure estimation and prediction.

To address the second challenge, in this work, we will first briefly summarize and discuss the existing outlier preservation methods, and then present our proposed visual-assisted outlier preservation strategy based on a modified density estimation visualization. In summary, our work has two main contributions:

- We propose and demonstrate a new way to preserve the occurrence, likelihood and density of outliers, which will not affect at the same time the relative data densities contrast between inliers, and outliers for scatterplot sampling.
- We provide a preliminary guided user study to collect feedback on the effectiveness of our proposed visual encoding of outliers for scatterplot sampling.

2. Related Work

2.1. Outlier Preservation for Scatterplot Sampling

For large-scale scatterplots, sampling is one of the most effective way to alleviate overplotting issues [ED07, PCM16, YXX*21], and different approaches have also proposed their own outlier preservation methods. It is usually achieved by sampling the outliers in three variants: 1) Using a pre-defined sampling strategy. For example, Liu et al. [LXL*18] proposed an outlier-biased random sampling method by adopting an outlier emphasis strategy to include more outliers in the final result and then extended the same idea by applying it to blue noise sampling and density-based sampling methods [XYX*19]. 2) Providing the user the flexibility to control the sample size of the preserved outliers. Recursive subdivision based sampling (RSBS) [CGZ*20], Pyramid based sampling (PBS) [CZF*22], or Z-order sampling (MVZS) [HSVK*20] all adopt this strategy while still giving higher priority to keep outliers during the sampling process. 3) Defining the sampling ratio for the outliers to match the overall data densities. Blue noise sampling (MBNS) [CCM*14] belongs to this category.

Methods 1) and 2) are the same, except that method 1) fixes the outlier sampling ratio while method 2) leaves it as a control parameter for the users to change it in different applications flexibly. Thus, in the end, there are only two options for sampling the outliers, either sampling them the same as all the other data points or sampling them with higher priority, which means more outliers will be retained relative to inliers. However, both options have their limitations. First, treating outliers specially with a priority may heavily influence the visual impression of the overall data density after sampling, and it may cause a misinterpretation of the data distribution [CGZ*20]. Second, treating the outliers just as all the other data may also cause problems, that is, people lose track of the true low-density areas after sampling since no more samples can represent certain sparse areas that contain some outliers.

We argue that one can improve these existing methods by adding a visual channel to emphasize the distribution of the sparse regions while sampling the outliers with equal probabilities to all the other data points. This way, the context of the low-density regions is retained, and the relative data densities in dense areas compared to the sparse areas are also maintained.

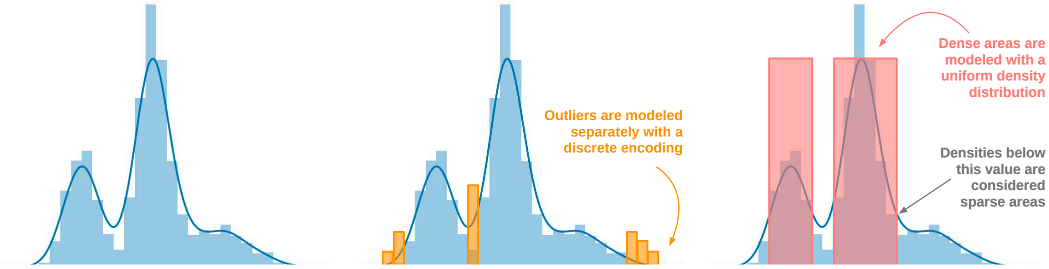
2.2. Density Visualization for Scatterplots

Applying density estimation for scatterplots has previously been explored to make scatterplots easier to read and understand. For example, Feng et al. [FKLT10] proposed to use kernel density estimation (KDE) to model the distribution of the data and integrate it with outliers in parallel coordinate plots and scatterplots for certain applications. Mayorga et al. [MG13] proposed *Splatterplot*, which turns the scatterplot into colored regions and uses a KDE based method to visualize the contour of the dense region for each class. This method can clearly show the overlap of different classes and outliers. However, it is difficult to visually distinguish the local density of each dense region since only the boundary contours of the dense regions are highlighted. Sarikaya et al. [SG18] argue that the choice of different visual designs for scatterplots can vary based on particular analysis goals, such as a hexagonal binned visual design [CLNL87] that can better reveal the spatial density. *Splatterplot* is limited in its capability of revealing the density distribution, especially in the dense areas by using contours only.

In contrast, our method first estimates the density of the scatterplot with a specific focus on the low-density areas. Then it integrates the density plot with the sampling results in the final scatterplot, which not only maintains the relative data densities through the scatterplot itself, but also highlights the low-density regions.

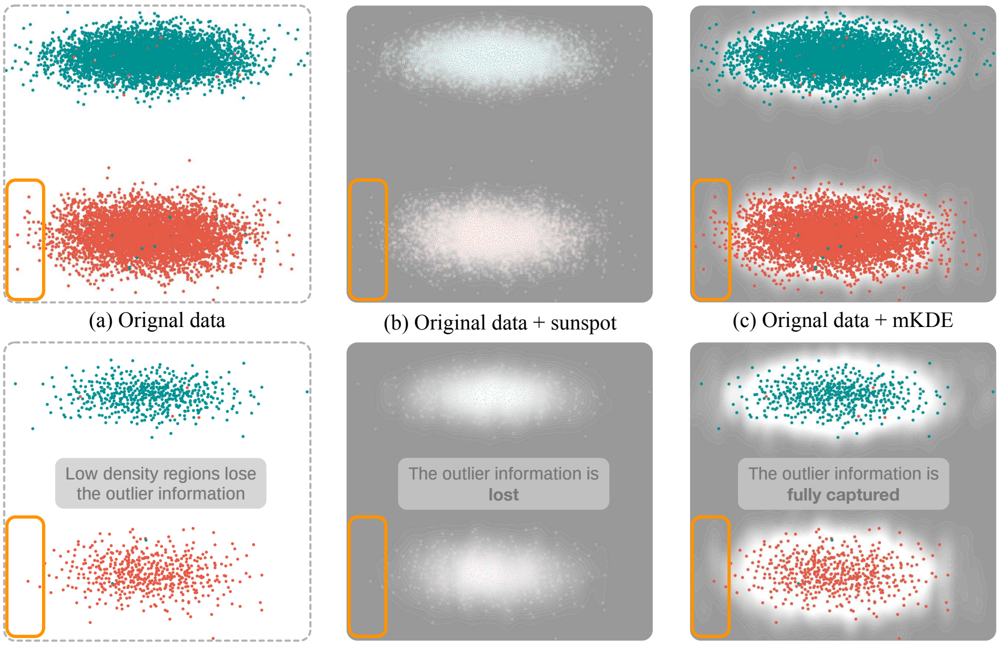
3. Density-based Visual Encoding

Our proposed visual-assisted outlier preservation strategy is based on a modified density estimation visualization. To resolve the dilemma between preserving the data densities contrast and preserving the context of outliers at the same time for scatterplot sampling as discussed above in Sec. 2.1, we summarize the following



(a) *KDE plot* **(b)** *sunspot plot* **(c)** *our mKDE plot*

Figure 2: *(a), (b), and (c) illustrate the basic ideas of the normal KDE plot, the modified kernel density estimation in sunspot plots [TBSB20] and in our work, respectively.*



(a) Orignal data (b) Original data + sunspot (c) Orignal data + mKDE

(d) Sample (e) Sample + sunspot (f) Sample + mKDE

Figure 3: *An outlier preservation demonstration. (a) is the scatterplot of the original data from two classes. (b) and (c) represent the original plot blended with sunspot plot and our modified KDE (mKDE) plot, respectively. Whereas (d), (e), and (f) show the sampled data ($\alpha = 0.1$) with or without the density plots.*

design rationales after discussing them with the visualization expert:

R1: The relative data densities should faithfully be maintained after sampling, including the outliers. And the density estimation needs to provide a guidance for the potential areas where outliers may exist.

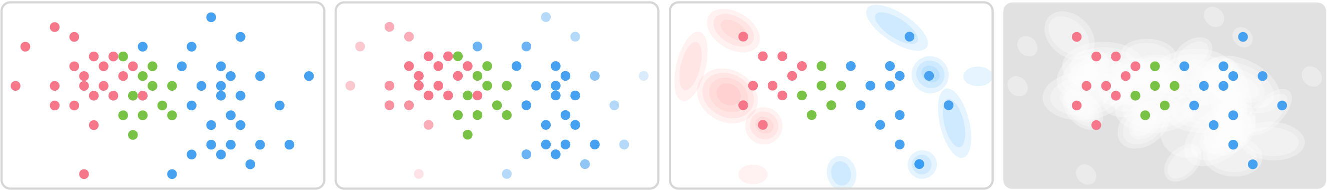
R2: The density estimation of the scatterplot needs to be able to emphasize the existence of the outliers, and the dense inlier areas.

R3: The density estimation of the scatterplot only plays a complementary role, meaning that the scatterplot itself should always capture the main focus of the users.

To support **R1**, the proportion of data points to be kept after sampling for any local regions of a scatterplot is expected to be the same. Thus the methods suggested in RSBS, PBS, and MVZS, where a priority is given to the outliers during the sampling process, are not applicable since all data points should be sampled with an equal probability. Consequently, only a few or even no data points may remain in the low-density regions after sampling. This will cause the users perceptually missing the overall data distribution for low-density areas. As a result, we employ the sampling method from MVZS to demonstrate our visual design results.

3.1. Density Encoding

In multi-class scatterplots, outliers are the points that are usually either far away from the majority of the data (low-



(a) *Input data* **(b)** *Use transparency* **(c)** *Color code the outliers* **(d)** *Our solution: modified KDE*

Figure 4: *The alternative visual designs of outlier preservation methods.*

density points) or are isolated within different classes of data points [SG18, CGZ*20]. In the latter situation, the disappearance of the outliers after sampling will not strongly affect the visual perception of the local data density for the majority class. Thus our method mainly focuses on the former case, i.e., low-density points. We borrow the initial idea from sunspot plots [TBSB20] where they combine the scatterplot with an additional density layer to strengthen the visual perception of the density distribution of scatterplots. However, as shown in Figs. 3(b) and (e), sunspot plots will lose the outlier information after sampling since the visual enhancement of the outliers is achieved by discrete encoding of each low-density data point. Additionally, this type of discrete encoding of the outliers loses the potential for predicting outliers in the sparse areas. In contrast, to preserve the outlier information in low-density areas after sampling, we combine a modified kernel density estimation (mKDE) with the scatterplot such that the low-density regions (instead of the low-density data points) are highlighted compared to the pure scatterplot sampling result without a density background (Figs. 3(a) and (d)) (**R2**). The idea is outlined below and also demonstrated in Figs. 3(c) and (f).

1. Divide the whole plotting space into M by N meshgrids and model the data density map using KDE at each meshgrid. Fig. 2(a) illustrates a 1D case.
2. Define the sparsity threshold t_s as the lowest γ percent of all the data densities. γ can vary case by case, and in our experiments, we set $\gamma = 10$.
3. Modify the data density map by compositing uniform distributions in the denser areas (areas that have data densities higher than t_s) as illustrated in Fig. 2(c).
4. Combine the modified KDE (mKDE) plot with the discrete sample data in one plot as shown in Figs. 3(c) and (f).

The key difference between our mKDE and the sunspot plots [TBSB20] is that sunspot plots combine the continuous density distribution of the dense data with discrete data points in the sparsely populated areas, as illustrated in Fig. 2(b), whereas our method suggests smoothly compositing the KDE plot with uniform density distributions in the dense areas, as shown in Fig. 2(c). One advantage of our method is that we show the probability of sparsity of the low-density regions, rather than low-density points in sunspot plots. And such difference makes our approach especially suitable for sampling results. The motivation of our design choice is twofold. First, since the sampling method could already preserve the relative data densities after sampling, the density map itself only needs to serve as a visual cue for low-density regions. Second, using a uniformly distributed density plot for the dense areas can help the users see the sampling results better without being visually distracted by the background (**R3**).

3.2. Discussions and Results

In different applications, one may favor other types of visual designs. For example, Fig. 4(b) illustrates the case when sampling is non-essential, whereas the level of the sparsity of the outliers is more important; one can use the transparency of the data points to encode the level of sparsity of the low-density points. On the other hand, when sampling is preferred, which means some (or all) of the outliers disappear after sampling, one can still obtain information on the low-density points, such as which areas possibly contain outliers from which classes before sampling? This can be answered by the visual encoding method shown in Fig. 4(c) where colored density estimations of the low-density points are applied to all the pre-detected outliers. However, this method has three potential limitations: First, the reliability of the result heavily depends on the accuracy of the outlier detection result. Second, the color-coded densities may attract users' attention more than the scatterplot and cause confusion of focus. And lastly, it is difficult to distinguish the correct order of densities between different colors.

Similarly, even though the key step of our proposed visual encoding is the modified KDE, the choice of color to plot the resulting density map is also crucial since introducing new colors from the density map into the scatterplots may cause perceptual ambiguity. And this is one key difference of our method compared to sunspot plots, which uses a high contrast colormap (yellow – purple) to encode the density map, resulting in the fact that sunspot plots are applied to single class data only, whereas our method can still visually preserve the class information of the original dataset. Thus in our experiments, we choose to use a sequential colormap of *Greys* from the matplotlib colormaps library [Hun07]. Because greyscale only contains luminance (brightness) information, and the scatterplot itself contributes the main color information for the integrated final plot. As shown in Fig. 4(d), the choice of the density map color is not very intrusive when we want the human to perceptually focus on the data points. Fig. 5 shows the effectiveness of our method from more examples. And Fig. 6 presents the runtime results of our proposed mKDE method. It can be seen that our approach scales well with the data size increases.

4. User Feedbacks

In order to evaluate our proposed outlier visual encoding method objectively, as well as to collect preliminary feedback from poten-

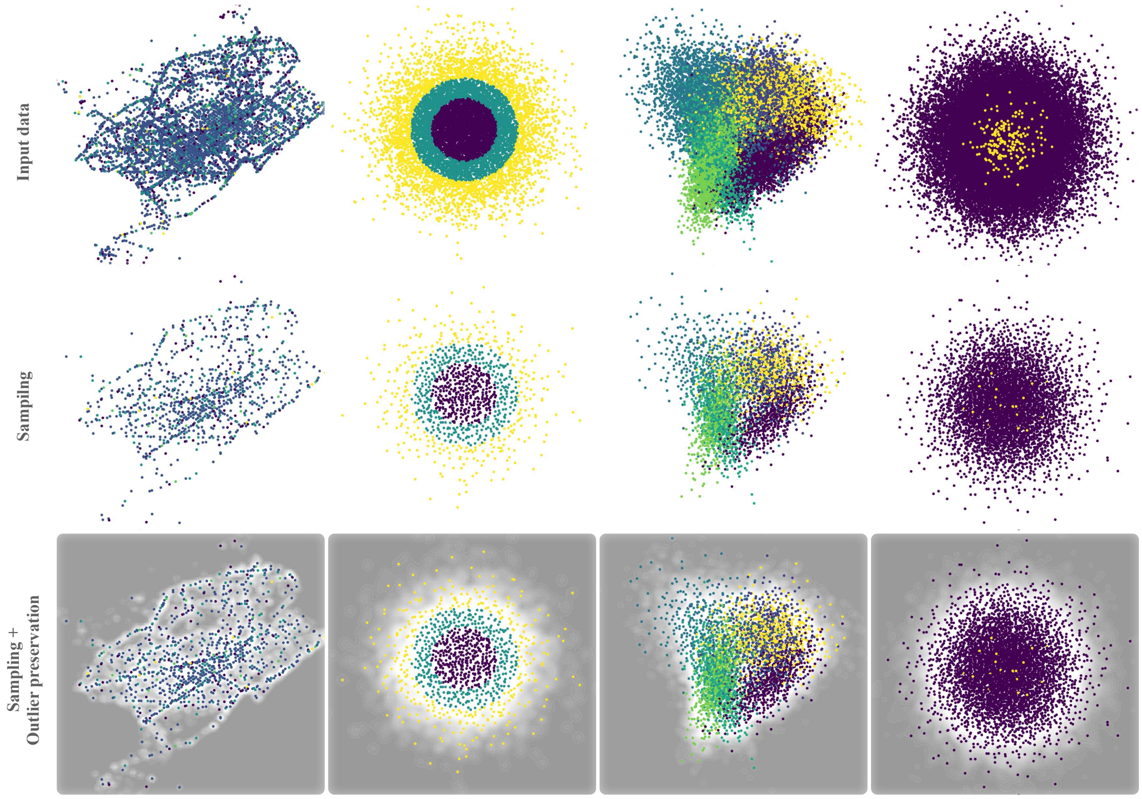


Figure 5: *The results of our proposed outlier preservation method on four different datasets. Compared to the input over-plotted scatterplots on the top row, the densities contrast preservation sampling results will lose the context of the low-density points as shown in the middle row, while in the bottom row, after combing our proposed outlier preservation method, the context of the low-density points are maintained.*

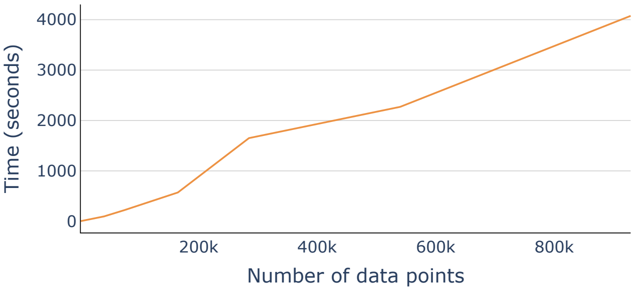


Figure 6: *The runtime (in seconds) for varying data sizes of our proposed mKDE method. All the data used in running the experiment can be found from [YXX*21].*

tial users and domain experts, we invited 7 researchers (1 visualization expert, 3 PhDs working on other data visualization topics and 3 PhDs in graphics and visualization related research areas) for a guided user study. The main purpose of this user study is to check the effectiveness of the visual-assisted outlier preservation scheme (Sec. 3). In this study, the participants are first informed about the tasks of the study, followed by an example-based explanation of how to read the examples. Then they go through each example and answer the questions. The study contains 10 examples.

In this study, for each example, we prepare six plots in the same way as in Fig. 7 and ask the participants to answer the following questions:

Q1: For KDE and mKDE, which part are you looking at first? The scatterplot, or the density plot?

Q2: Can you read the KDE and mKDE plots correctly as which parts represent dense areas and which parts sparse areas? Choose from “very clear”, “clear”, “neutral”, “unclear”, and “very unclear”.

Q3: Which density encoding do you like most to preserve the density information of outliers?

Q4: Do you think it is necessary (better) to have the density plot when preserving outlier information. Yes or No?

In all the questions, KDE refers to the normal density plot shown in Figs. 7(b1) and (b2), and mKDE refers to the modified density plot shown in Figs. 7(c1) and (c2). The purpose of this study is to evaluate the necessity of our proposed visual encoding on the low-density regions, as well as the effectiveness of such visual encoding. Given the fact that our participants have different research backgrounds, and their answers also vary from one to another. For example, for Q1, as summarized in Fig. 8(a), about 2/3 of them think that using KDE the scatterplot is more prominent, and in

Proceedings published by Eurographics - The European Association for Computer Graphics.

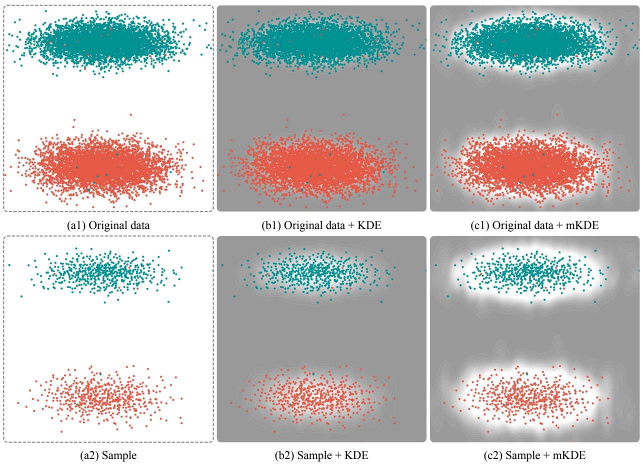


Figure 7: *One example placement of the plots used in the user study.*

mKDE, more than half think that the density plots draw their attention first. Interestingly, the visualization expert gives an opposite opinion. The main reason is that the expert is more used to and well-trained to look at certain visualizations like scatterplots and density plots. This result indicates that our visual encoding is not yet perfect and has to be improved such that the scatterplot (as the main content) can capture the users' main attention. One of the participants suggested that using a proper color encoding of the scatterplot may further help. While for Q2, all the participants agree that mKDE is slightly better than KDE regarding representing and differentiating dense areas and sparse areas in the plots as shown in Fig. 8(b). And the main reason is that with mKDE, the dense areas stand out on a grey background due to the applied uniform white color encoding. Whereas KDE may fail for certain cases, resulting in more "unclear" results.

On the other hand, all the participants come up with a uniform answer to both Q3 and Q4. For Q3, all of them think that our proposed mKDE method can better preserve the density information of outliers. And for Q4, they all agree that it is necessary to have the density plot when preserving outlier information because "*the density plot adds additional information on the data distribution, which one may not guess from the scatterplot alone*". Moreover, the visualization expert also suggested that "*current density plot does not differentiate among different classes, it is challenging, but can be a future work to do*".

In summary, the participants are generally positive about our proposed outlier preservation method by adding a modified KDE plot to the scatterplot. However, certain limitations still exist and shall be improved in the future.

5. Conclusion

Sampling is one of the most effective ways to solve the overplotting issue for large-scale scatterplots. However, the inherent dilemma exhibited in existing scatterplot sampling techniques brings new challenges; that is, to maintain the relative data densities contrast before and after sampling, the outliers (low-density points) and inliers (high-density points) need to be sampled with the same probability. And in this case, the low-density areas lose the context of the outliers, causing misinterpretation of the original data distribution. To tackle this issue, in this work, we propose a novel visual-assisted outlier preservation method for scatterplot sampling by adding a modified kernel density estimation. Our method first estimates the density of a scatterplot from all its data points, then modifies this density plot by compositing uniform distributions in the dense areas, and finally combines the modified density plot with the scatterplot sampling results. Our visual design emphasizes the low-density areas by encoding the probability of sparsity using color and contour lines from KDE (**R2**), which also ensures the perceptual focus to be on the majority of the data points (**R3**) when choosing appropriate color scheme as discussed in Sec. 3.2.

We also want to point out a few limitations of our work that could further be improved. First, our method can be applied to multi-class scatterplots without encoding the density estimation of the data points from each class with different colors since this may lead to severe clutter issues perceptually, especially when the number of classes is large. Thus, our current visual design can preserve the density distribution of all the low-density areas regardless of their classes. For class-sensitive applications, our method may not be suitable. Second, from the user feedback, we learn that choosing the right color scheme for both the scatterplot and the modified density plot is critical, and a wrong color choice can sometimes bring unsatisfactory or even negative results. *Splatterplot* [MG13] is one of the successful cases that use color blending techniques to solve the clutter issue of multi-class scatterplots, and it is worth investigating applying similar approaches to scatterplot sampling applications.

References

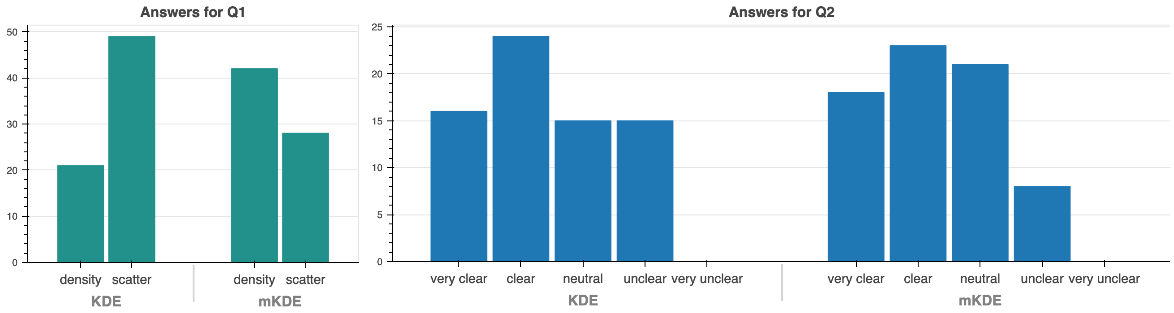
[BBK*18] Behrisch M., Blumenschein M., Kim N. W., Shao L., El-Assady M., Fuchs J., Seebacher D., Diehl A., Brandes U., Pfister H., et al.: Quality metrics for information visualization. *Computer Graphics Forum 37*, 3 (2018), 625–662. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13446, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13446, doi:https://doi.org/10.1111/cgf.13446. 1, 2

[BKNS00] Breunig M. M., Kriegel H.-P., Ng R. T., Sander J.: Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2000), SIGMOD '00, Association for Computing Machinery, pp. 93–104. URL: https://doi.org/10.1145/342009.335388, doi:10.1145/342009.335388. 2

[BS05] Bertini E., Santucci G.: Improving 2d scatterplots effectiveness through sampling, displacement, and user perception. In *Ninth International Conference on Information Visualisation (IV'05)* (2005), pp. 826–834. doi:10.1109/IV.2005.62. 2

[BS06] Bertini E., Santucci G.: Give chance a chance: Modeling density to enhance scatter plot quality through random data sampling. *Information Visualization 5*, 2 (2006), 95–110. URL: https://doi.org/10.1057/palgrave.ivs.9500122, doi:10.1057/palgrave.ivs.9500122. 2

[CCM*14] Chen H., Chen W., Mei H., Liu Z., Zhou K., Chen W., Gu W., Ma K.-L.: Visual abstraction and exploration of multiclass scatterplots. *IEEE Transactions on Visualization and Computer*



(a) *User study Q1*

(b) *User study Q2*

Figure 8: *Results for the user study on questions 1 and 2.*

Graphics 20, 12 (2014), 1683–1692. doi:10.1109/TVCG.2014.2346594. 2

[CGZ*20] Chen X., Ge T., Zhang J., Chen B., Fu C.-W., Deussen O., Wang Y.: A recursive subdivision technique for sampling multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 729–738. doi:10.1109/TVCG.2019.2934541. 2, 4

[CLNL87] Carr D. B., Littlefield R. J., Nicholson W., Littlefield J.: Scatterplot matrix techniques for large n. *Journal of the American Statistical Association 82*, 398 (1987), 424–436. URL: https://doi.org/10.1080/01621459.1987.10478445, doi:10.1080/01621459.1987.10478445. 2

[CZF*22] Chen X., Zhang J., Fu C.-W., Fekete J.-D., Wang Y.: Pyramid-based scatterplots sampling for progressive and streaming data visualization. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2022), 593–603. doi:10.1109/TVCG.2021.3114880. 2

[ED07] Ellis G., Dix A.: A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1216–1223. doi:10.1109/TVCG.2007.70535. 1, 2

[FKLT10] Feng D., Kwock L., Lee Y., Taylor R.: Matching visual saliency to confidence in plots of uncertain data. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 980–989. doi:10.1109/TVCG.2010.176. 1, 2

[HSVK*20] Hu R., Sha T., Van Kaick O., Deussen O., Huang H.: Data sampling in multi-view and multi-class scatterplots via set cover optimization. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 739–748. doi:10.1109/TVCG.2019.2934799. 2

[Hun07] Hunter J. D.: Matplotlib: A 2d graphics environment. *Computing in Science & Engineering 9*, 3 (2007), 90–95. doi:10.1109/MCSE.2007.55. 4

[KW13] Krzywinski M., Wong B.: Points of view: Plotting symbols. *Nature methods 10* (06 2013), 451. doi:10.1038/nmeth.2490. 1

[LMvW10] Li J., Martens J.-B., van Wijk J. J.: A model of symbol size discrimination in scatterplots. CHI '10, Association for Computing Machinery. URL: https://doi.org/10.1145/1753326.1753714, doi:10.1145/1753326.1753714. 1

[LvWM09] Li J., van Wijk J. J., Martens J.-B.: Evaluation of symbol contrast in scatterplots. In *Proceedings IEEE Pacific Visualization Symposium* (2009), pp. 97–104. doi:10.1109/PACIFICVIS.2009.4906843. 1

[LvWM10] Li J., van Wijk J. J., Martens J.-B.: A model of symbol lightness discrimination in sparse scatterplots. In *Proceedings IEEE Pacific Visualization Symposium* (2010), pp. 105–112. doi:10.1109/PACIFICVIS.2010.5429604. 1

[LXL*18] Liu S., Xiao J., Liu J., Wang X., Wu J., Zhu J.: Visual diagnosis of tree boosting methods. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 163–173. doi:10.1109/TVCG.2017.2744378. 2

[MG13] Mayorga A., Gleicher M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics 19*, 9 (2013), 1526–1538. doi:10.1109/TVCG.2013.65. 1, 2, 6

[MPOW17] Micallef L., Palmas G., Oulasvirta A., Weinkauf T.: Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (2017), 1588–1599. doi:10.1109/TVCG.2017.2674978. 1

[Mun14] Munzner T.: *Visualization analysis and design*. CRC press, 2014. 1

[PCM16] Park Y., Cafarella M., Mozafari B.: Visualization-aware sampling for very large databases. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)* (2016), pp. 755–766. doi:10.1109/ICDE.2016.7498287. 2

[RRBKRD17] Ramos Rojas J. A., Beth Kery M., Rosenthal S., Dey A.: Sampling techniques to improve big data exploration. In *Proceedings IEEE Symposium on Large Data Analysis and Visualization* (2017), pp. 26–35. doi:10.1109/LDAV.2017.8231848. 2

[SG18] Sarikaya A., Gleicher M.: Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 402–412. doi:10.1109/TVCG.2017.2744184. 1, 2, 4

[TBSB20] Trautner T., Bolte F., Stoppel S., Bruckner S.: Sunspot plots: Model-based structure enhancement for dense scatter plots. *Computer Graphics Forum 39*, 3 (2020), 551–563. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14001, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14001, doi:https://doi.org/10.1111/cgf.14001. 3, 4

[WLS98] Woodruff A., Landay J., Stonebraker M.: Constant density visualizations of non-uniform distributions of data. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 1998), UIST '98, Association for Computing Machinery, pp. 19–28. URL: https://doi.org/10.1145/288392.288397, doi:10.1145/288392.288397. 1

[XYX*19] Xiang S., Ye X., Xia J., Wu J., Chen Y., Liu S.: Interactive correction of mislabeled training data. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2019), pp. 57–68. doi:10.1109/VAST47406.2019.8986943. 2

[YXX*21] Yuan J., Xiang S., Xia J., Yu L., Liu S.: Evaluation of sampling methods for scatterplots. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (February 2021), 1720–1730. doi:https://doi.org/10.1109/TVCG.2020.3030432. 2, 5