

Week (10) - “Modeling the Influence of Visual Density on Cluster Perception in Scatterplots Using Topology ”

Khulood Alkhudaiddi

Scatterplots are among the popular methods used to explore the data characteristics, such as relationships between quantitative variables. When it comes to clusters, the data distribution type/size, data point number, mark type, mark size, and opacity are examples of factors that might affect the user perception.

This paper aims to explore the effectiveness of these factors for the cognition of clusters in scatter plots. Using a data structure called a *merge tree*, the researchers conducted a study and came up with two models for the clusters' perception, namely distance-based model and density-based model. Then, they verified the accuracy and the impacts of their models with these variables shown in Figure (1).

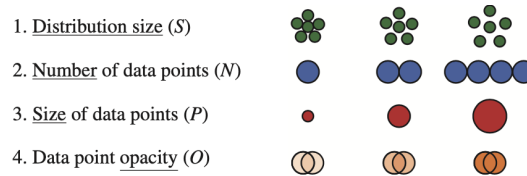
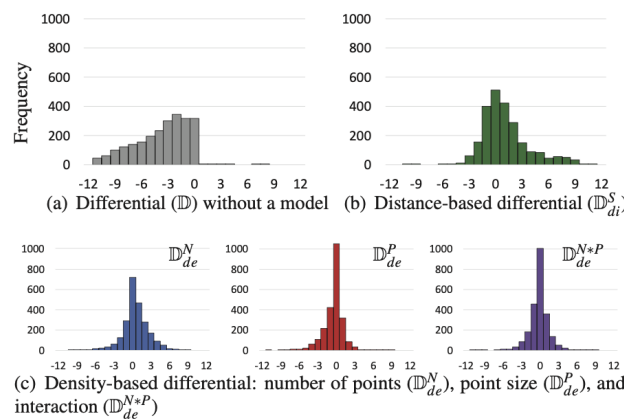


Figure (1)


























The researchers tested their two models considering S, N, and P factors. They found out that the density-based model achieves higher accuracy than the distance-based model. Besides, they analyzed the impact of these factors without considering the models. They noticed that both the point number and the distribution size factors significantly impact the counting clusters process. They also explored the interaction between both factors and found a highly significant effect between them. However, the point size factor does not have that effect; results are shown in Figure (2).



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

Figure (2)

Testing the opacity with the density-based model showed that opacity has a very high impact with dense distribution than with sparse. Also, more transparent points are good for small distribution size, while opaque points are more effective for the large distribution size scatter plot. The following table is a summary of suggestions that can help designers when designing clusters in scatter plots.

Factor	Without Model	Distance-based (T_{di})	Density-based			
			(T_{de}^N)	(T_{de}^P)	(T_{de}^{N*P})	(T_{de}^O)
Distribution size (S)						
Number of points (N)						—
Size of points (P)						—
Opacity (O)	—	—	—	—	—	
(N*P)						—
Interaction (O*S)	—	—	—	—	—	
(S*N)			—	—	—	—





 large effect;
  medium effect;
  small effect;
  negligible effect;
 — not tested

Table (1)

I like how this paper validates the importance of these factors then gives suggestions that help designers make their decisions when designing such a commonly used plot. It is fascinating to know about the importance of these choices and how they affect the users' perception.

Sources:

1. <https://ieeexplore.ieee.org/abstract/document/9222295>