Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

ClustMe: A Visual Quality Measure for Ranking Monochrome Scatterplots based on Cluster Patterns – Supplemental Material –

Mostafa M. Abbas[1], Michaël Aupetit[1], Michael Sedlmair[2], and Halima Bensmail[1]

[1] QCRI, HBKU, Doha, Qatar
[2] VISUS, University of Stuttgart, Germany

1 Parameters values for merging techniques

The Table 1 gives the default values of the parameters proposed by Hennig in the `fpc` R-package and used in our experiments.

Table 1: *Default parameter values of the merging methods used in our experiments.*

Method	Cutoff	Interval width
bhat	0.1	NA
ridge.uni	1	0.005
ridge.ratio	0.2	0.005
demp	0.025	NA
dipuni	0.05	0.005
diptantrum	0.05	0.005
predictive	0.75	NA, m=50

2 Stratified sampling

2.1 Generating various clustering patterns

In order to get various cluster patterns that cover well the full space of possible shapes, we could not simply uniformly sample the GMM parametric space that generated the shape as parameters are not straightforwardly related to shape perceived complexity. Instead we used a stratified sampling approach similar to the one used by Pandey *et al.* [PKF*16]. We refer to the section 2 of the supplementary material regarding this procedure.

that uses visual quality measures as indicators of perceived shape complexity and primary sampling factors. Therefore, we computed two computational measures of *class separation* S_{NN} and S_{Class}, that is, measures that are designed to quantify how different (color-coded) classes are separated in scatterplots. We can use them here, as we know *the component labels* u or v from our generating process. S_{NN} is among the best class-separation VQMs proposed [AS16]: for each point it counts how many among its two nearest neighbors have its own label (0, 1 or 2), and returns half the average value of this count over all the points. S_{Class} is proposed here and based on the misclassification probability: it computes the probability that points x actually generated by component u, are classified as being generated by component v and vice-versa. The maximum of these two probabilities is used as an indicator of the overlap of the components. The S_{Class} is defined as:

$$S_{Class}(X,\Theta) = 1 - \max\Big(\frac{1}{|U|}\sum_{x\in U} P(v|x,\Theta), \frac{1}{|V|}\sum_{x\in V} P(u|x,\Theta)\Big) \quad (1)$$

where $P(i|x,\Theta) = \pi_i g_i(x)/p(x|\Theta)$ is the posterior probability that component i generated point x in the scatterplot (X,Θ).

We applied K-means clustering [MJ66] in the 2 dimensional Euclidean space defined by $Score_{NN}$ and $Score_{Class}$ to partition the 2073600 scatterplots into $K = 20$ groups. From each group, we selected randomly 51 scatterplots to collect 1020 in total (See K-means sampling in the figure 1).

Each selected scatterplot has been rotated randomly by an angle $\alpha \in \{0, \pi/2, 5\pi/4\}$ before rendering. The parameter values of the 1020 selected scatterplots are distributed uniformly except for μ (See histograms in the figure 2). The sample contains more cases with small values of μ, that is, those where components are more likely to overlap forming a single cluster. This imbalance distribution over μ is favorable to better cover the cases where the grouping patterns are actually more difficult to decide and depend on all the other parameters.

2.2 Graphical Analysis of the Stratified Sampling Process

The figure 1 shows the 1000 scatterplots stimuli in the space of $Score_{NN}$ and $Score_{Class}$, color-coded with the 20 group labels assigned by K-means.

2.3 Quality of the sampling of the parameter space

The figure 2 shows the histograms of all the parameters of the 1000 scatterplots stimuli selected for the human subjects study, sampled from the parameter space of a 2-component bivariate Gaussian Mixture Model. The parameter values of the 1000 selected scatterplots are distributed uniformly except for μ. The sample contains more cases with small values of μ, that is, those where the component are more likely to overlap forming a single cluster.

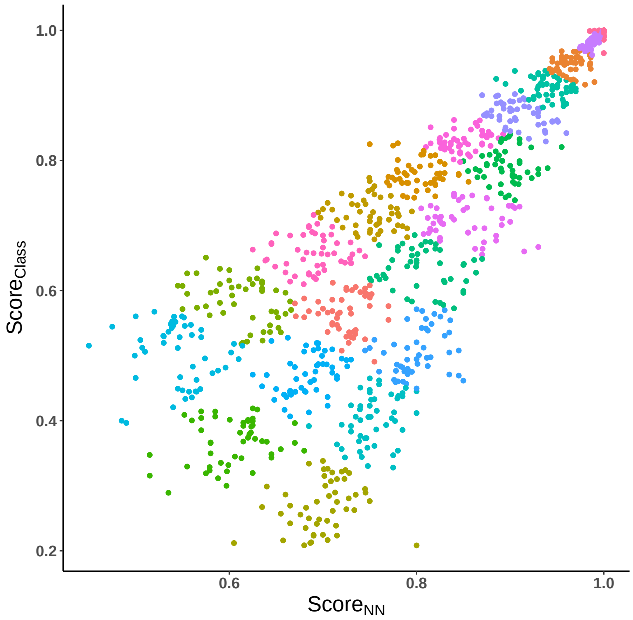

Figure 1: ***Stratified Sampling of the GMM parameter space*** *The values of Score$_{NN}$ and Score$_{Class}$ for the* 1000 *selected scatterplots color-coded with the K-means groups*

2.4 Clumpiness versus Human Judgment in Experiment 1

Figure 3 (Left) shows the distribution of the 1000 plots comparing the human judgment to the Clumpiness scagnostic for each plot. It appears that low values of Clumpiness are independent of the human separation score, while high Clumpiness values always match with high human separation score. Clumpiness is not sensitive to certain grouping patterns, as exemplified in Figure 3 (Right). We would expect a monotonic positive relation between both measures, but this result confirms our empirical observations that led us to engage in seeking a new VQM for grouping patterns.

3 Instructions given for Experiments 1 and 2

The figure 4 shows the instructions given to the subjects for each experiment.

4 257 benchmark data ranked with ClustMe and Clumpiness

In figures 5 to 14 we show the top and bottom 100 scatterplots of the 257 benchmark datasets ranked with ClustMe and Clumpiness.

References

[AS16] Aupetit M., Sedlmair M.: Sepme: 2002 new visual separation measures. *Proc. IEEE Pacific Visualization Symp.* (2016), 1–8. doi:10.1109/PACIFICVIS.2016.7465244. 1

[MJ66] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability* (1966). 1

[PKF*16] Pandey A. V., Krause J., Felix C., Boy J., Bertini E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)* (2016), pp. 3659–3669. doi:10.1145/2858036.2858155. 1

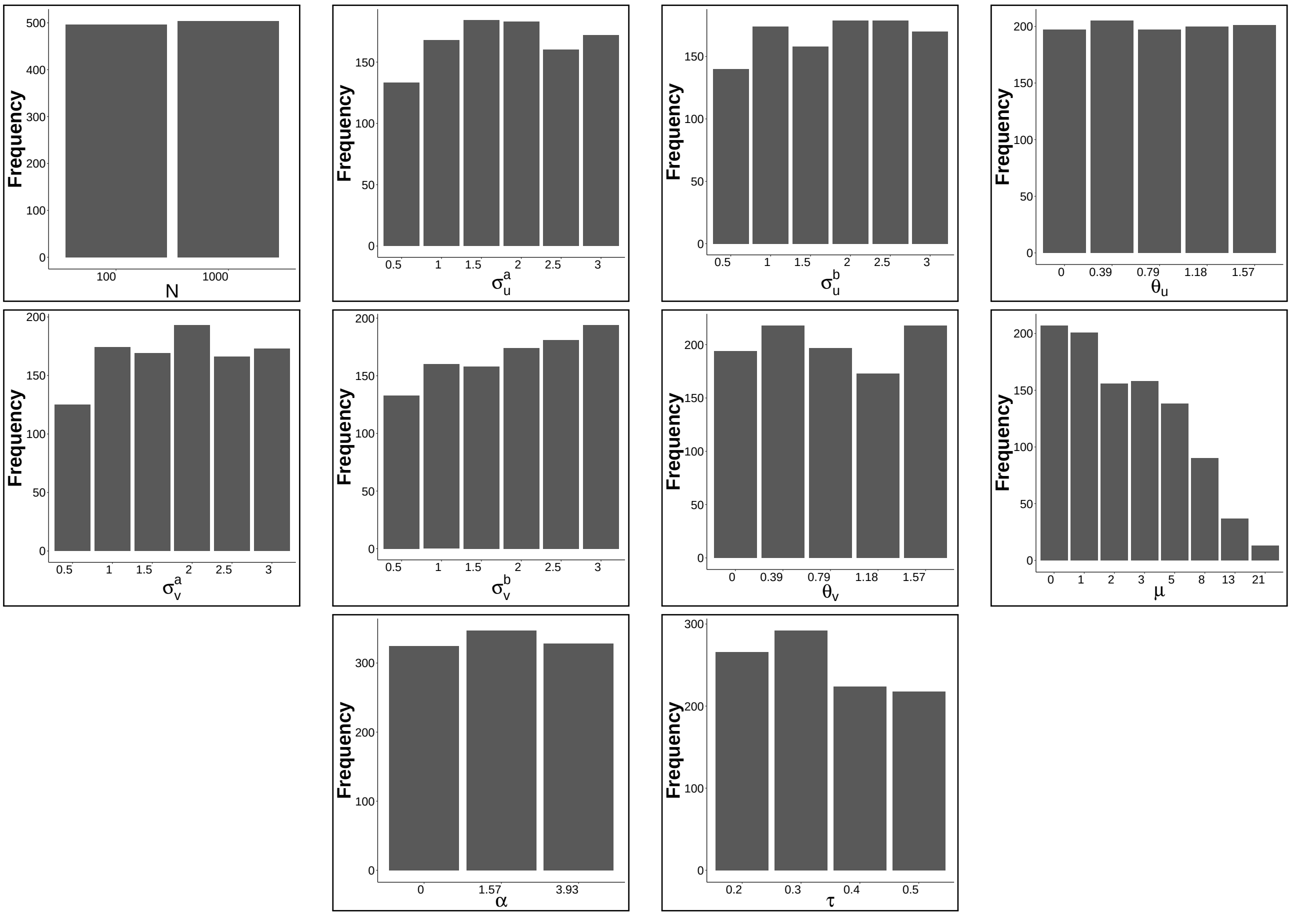

Figure 2: ***Quality of the sampling of the parameter space:*** *Histograms of all the parameters of the* 1000 *scatterplots selected for the human subjects study, sampled from the parameter space of a 2-component bivariate Gaussian Mixture Model.*

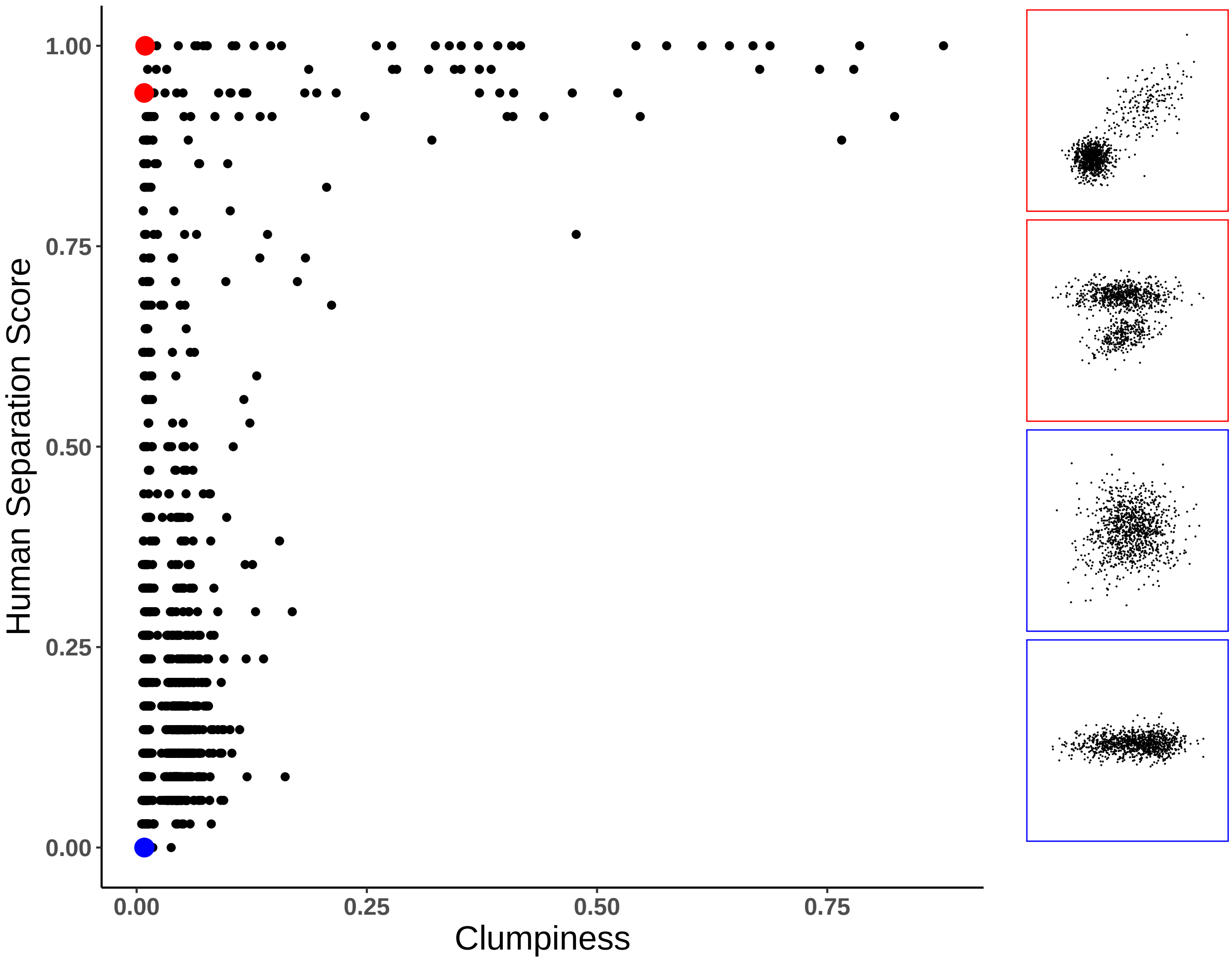

Figure 3: ***Left:*** *Overview over the* 1000 *scatterplots stimuli of Experiment 1 based on their human separation and Clumpiness scores. The human separation score is the proportion of humans subjects perceiving* 2 *or more clusters in a scatterplot. Clumpiness score is high for more structured patterns and should be low when no strong structure appears. A good perception-based VQM would show no points in the top left and bottom right corners.* ***Right:*** 4 *low Clumpiness stimuli with same frame color as the points selected in the left view.*

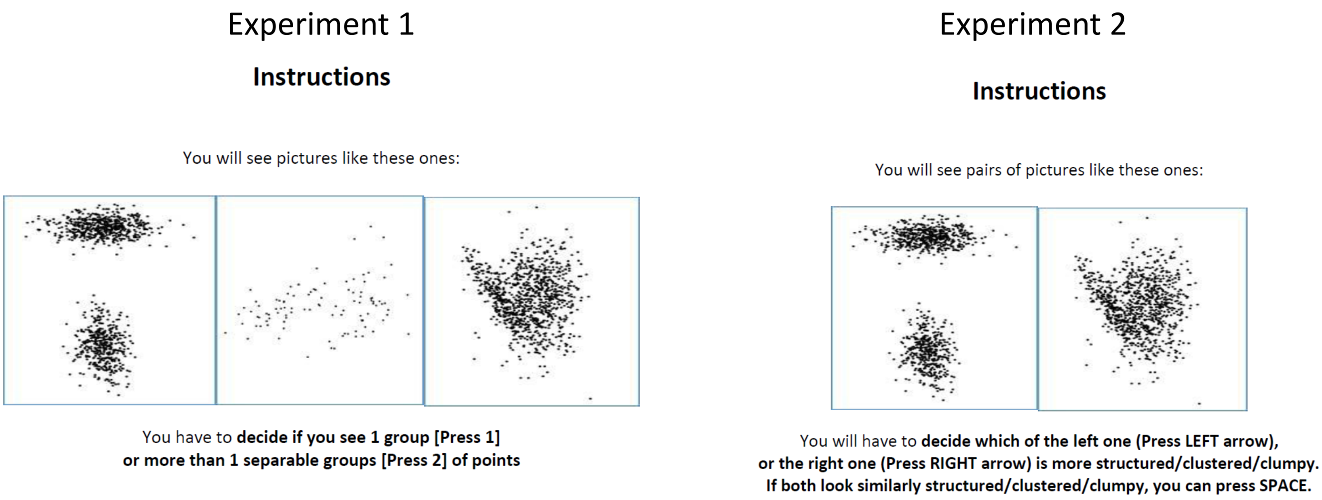

Figure 4: ***Instructions*** *given to participants for Experiment 1 (left) and Experiment 2 (right).*

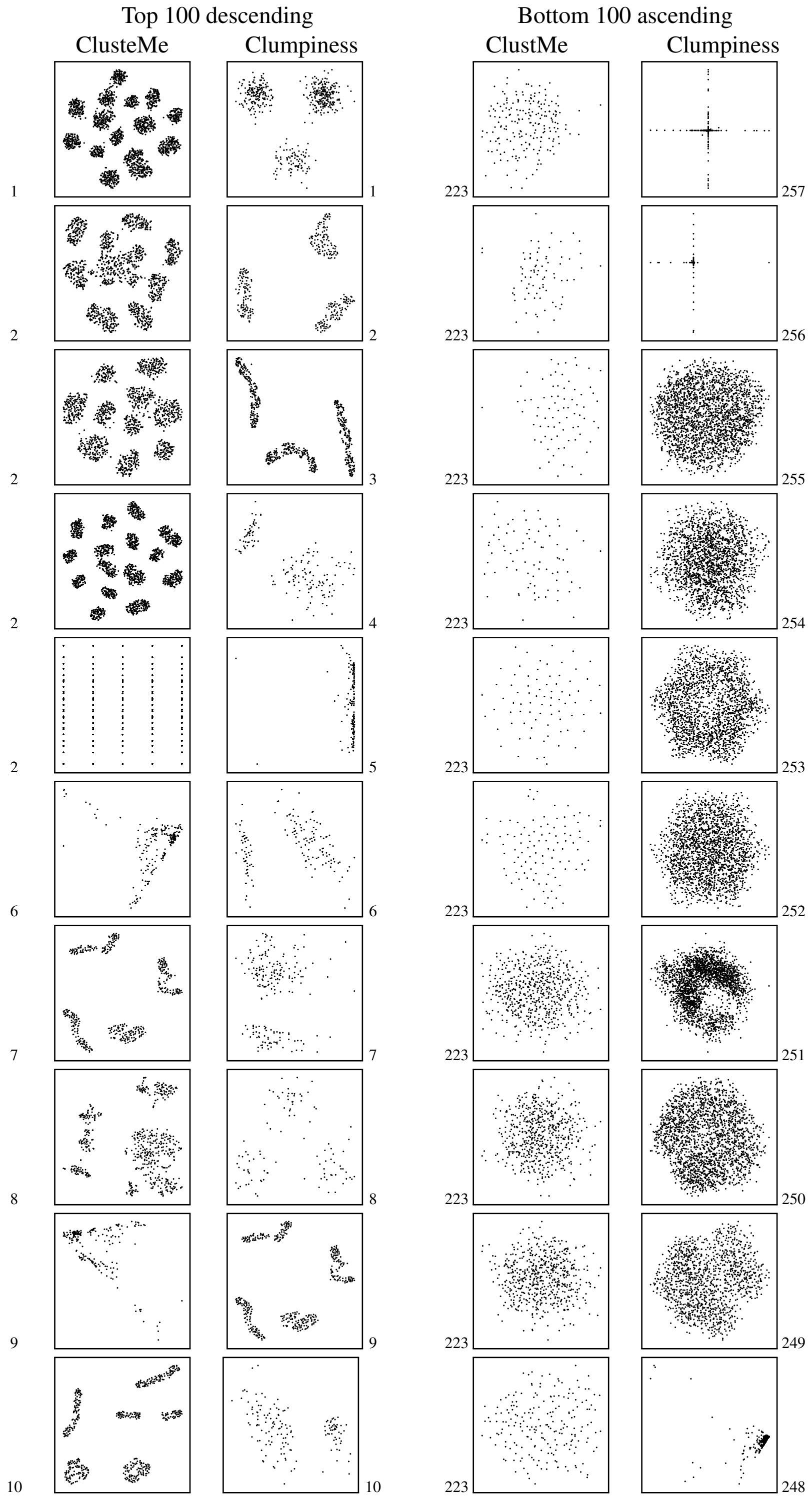

Figure 5: *Series 1/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

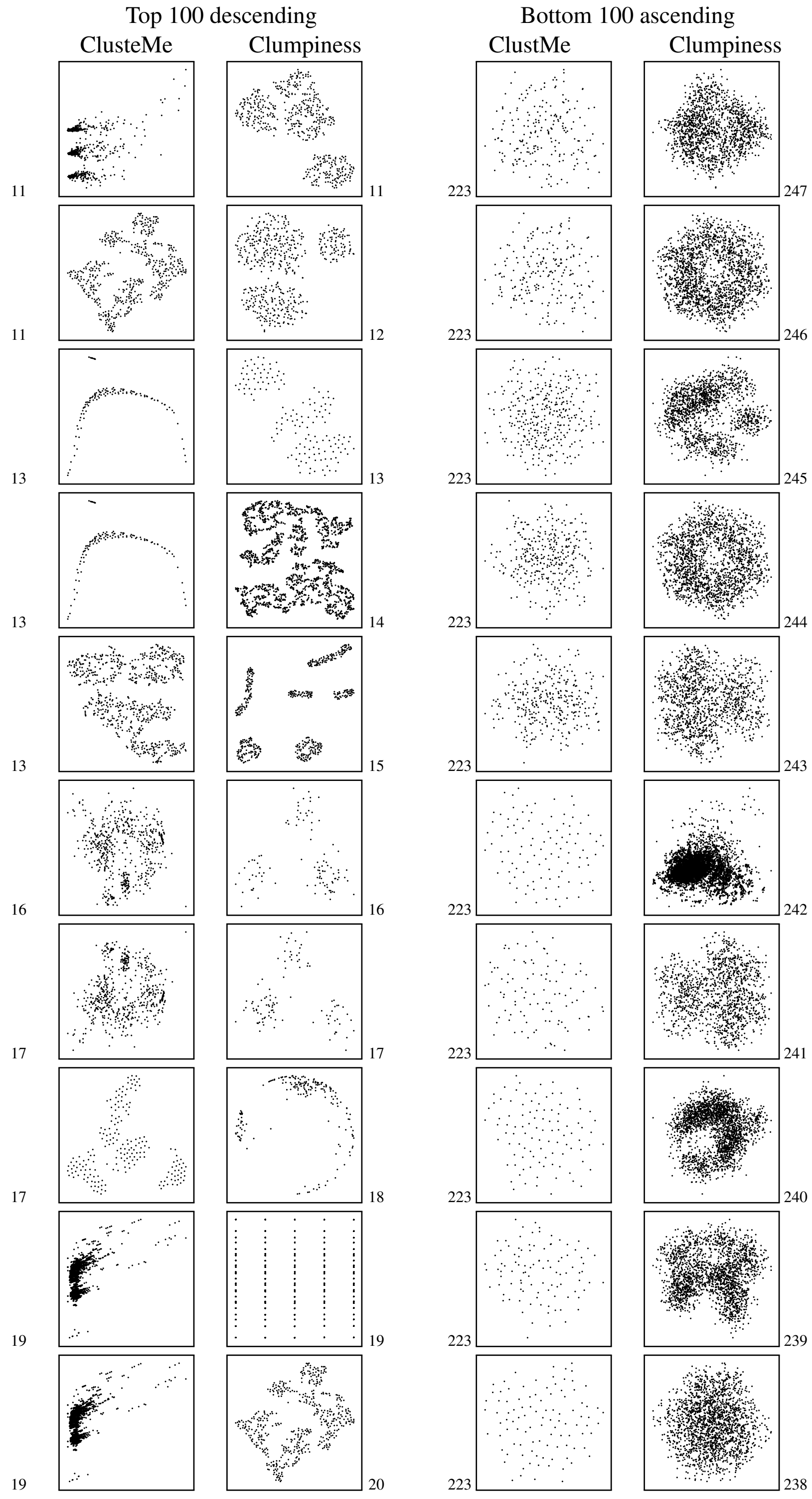

Figure 6: *Series 2/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

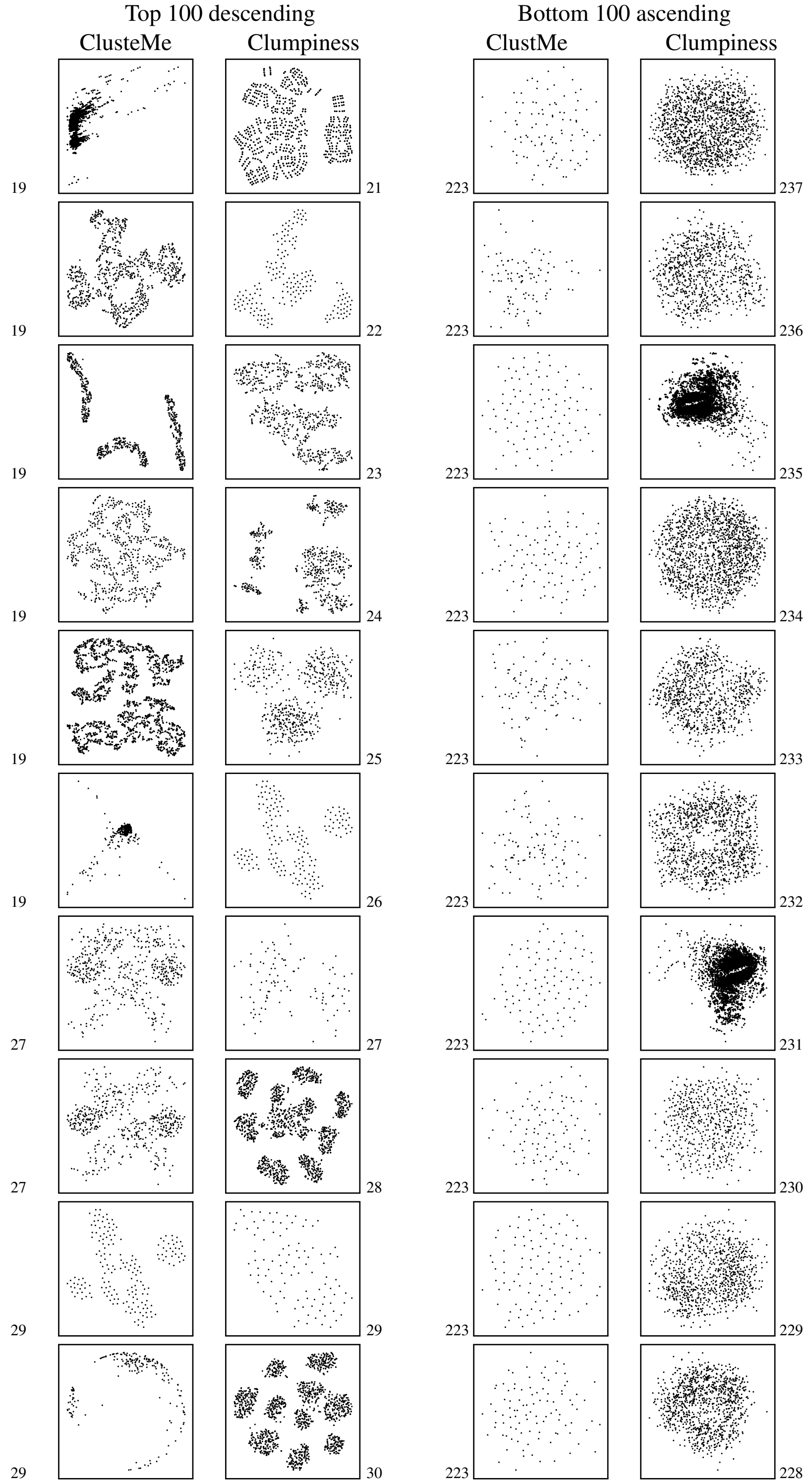

Figure 7: *Series 3/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

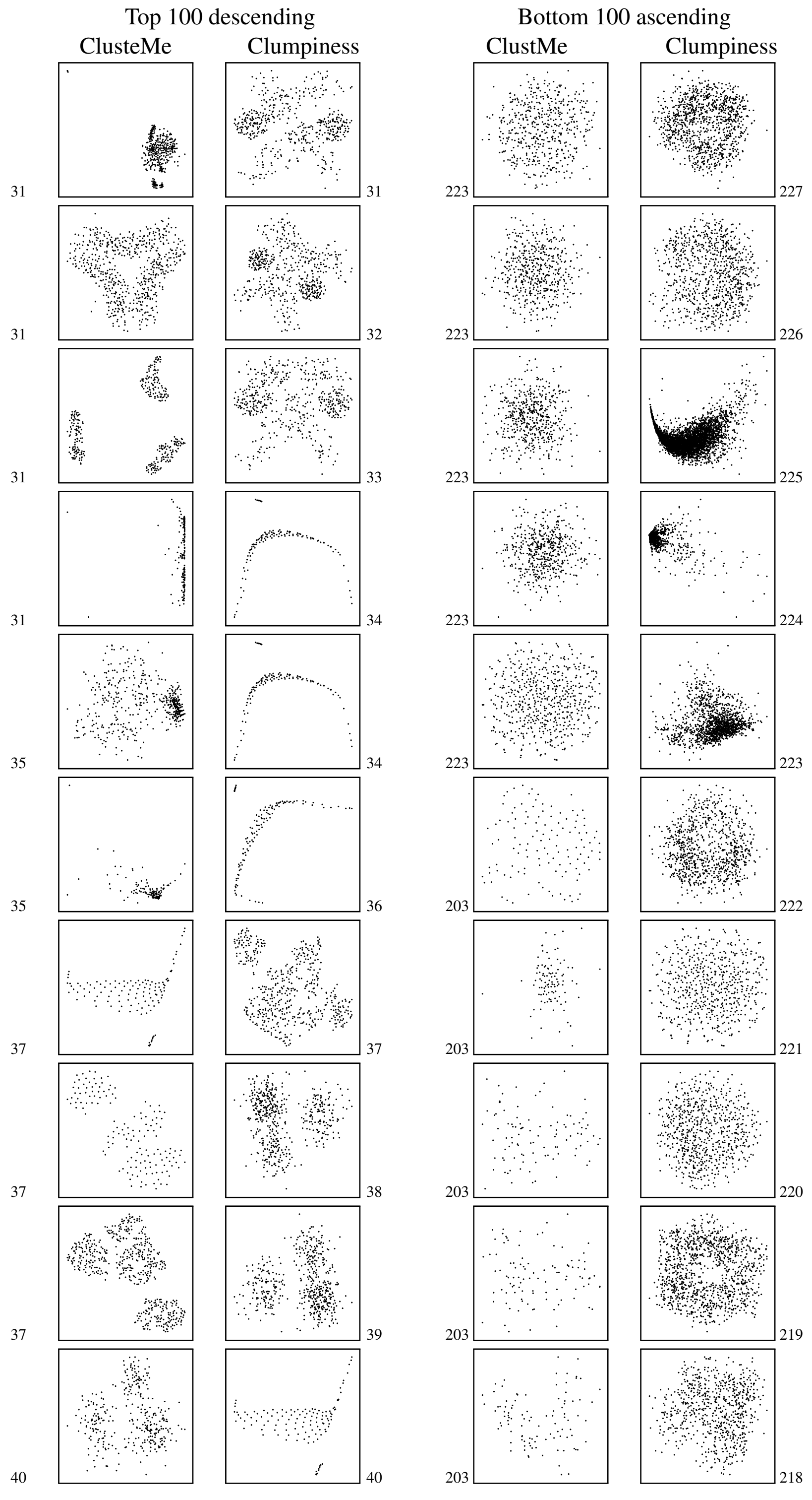

Figure 8: *Series 4/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

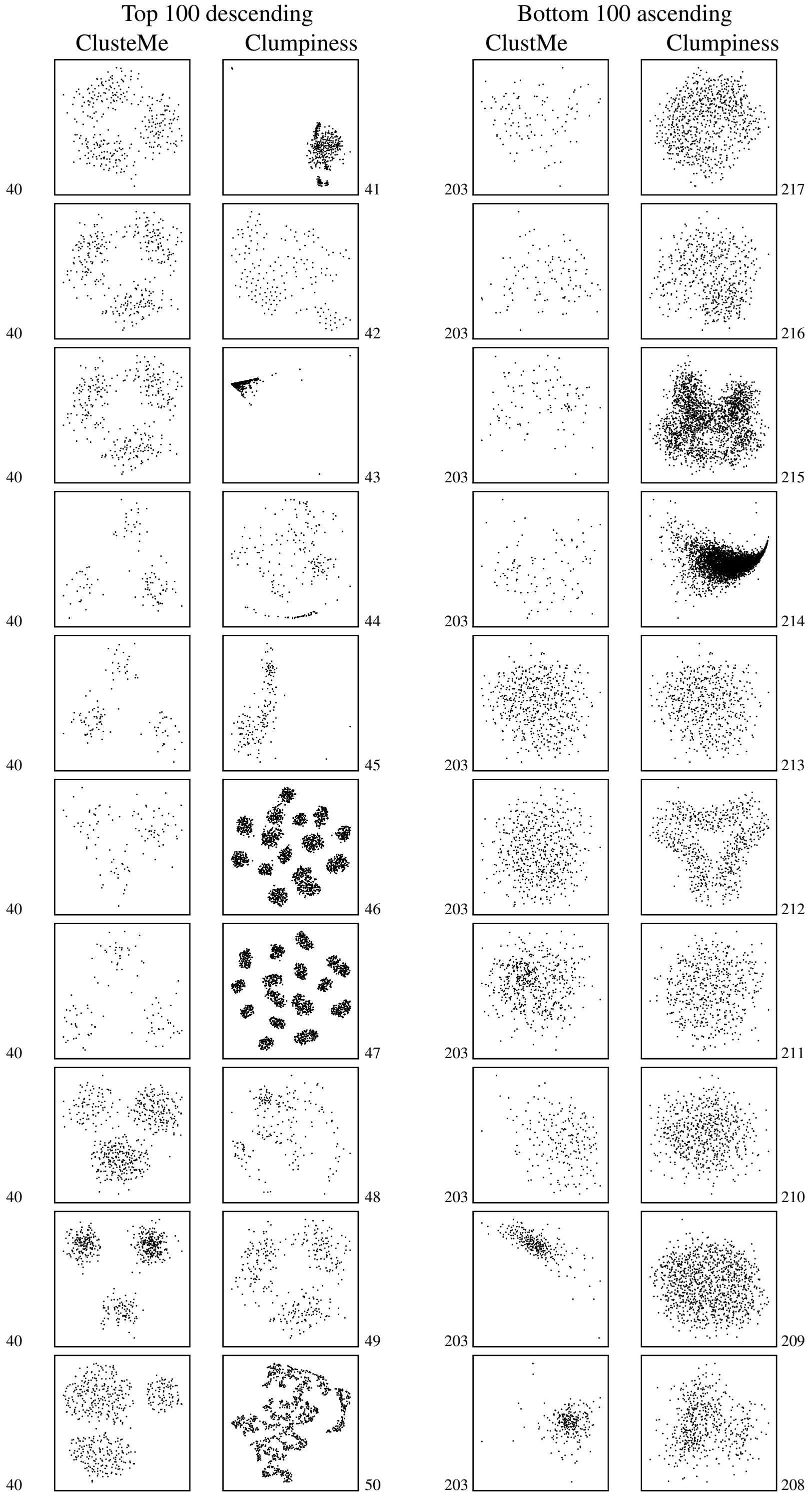

Figure 9: *Series 5/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

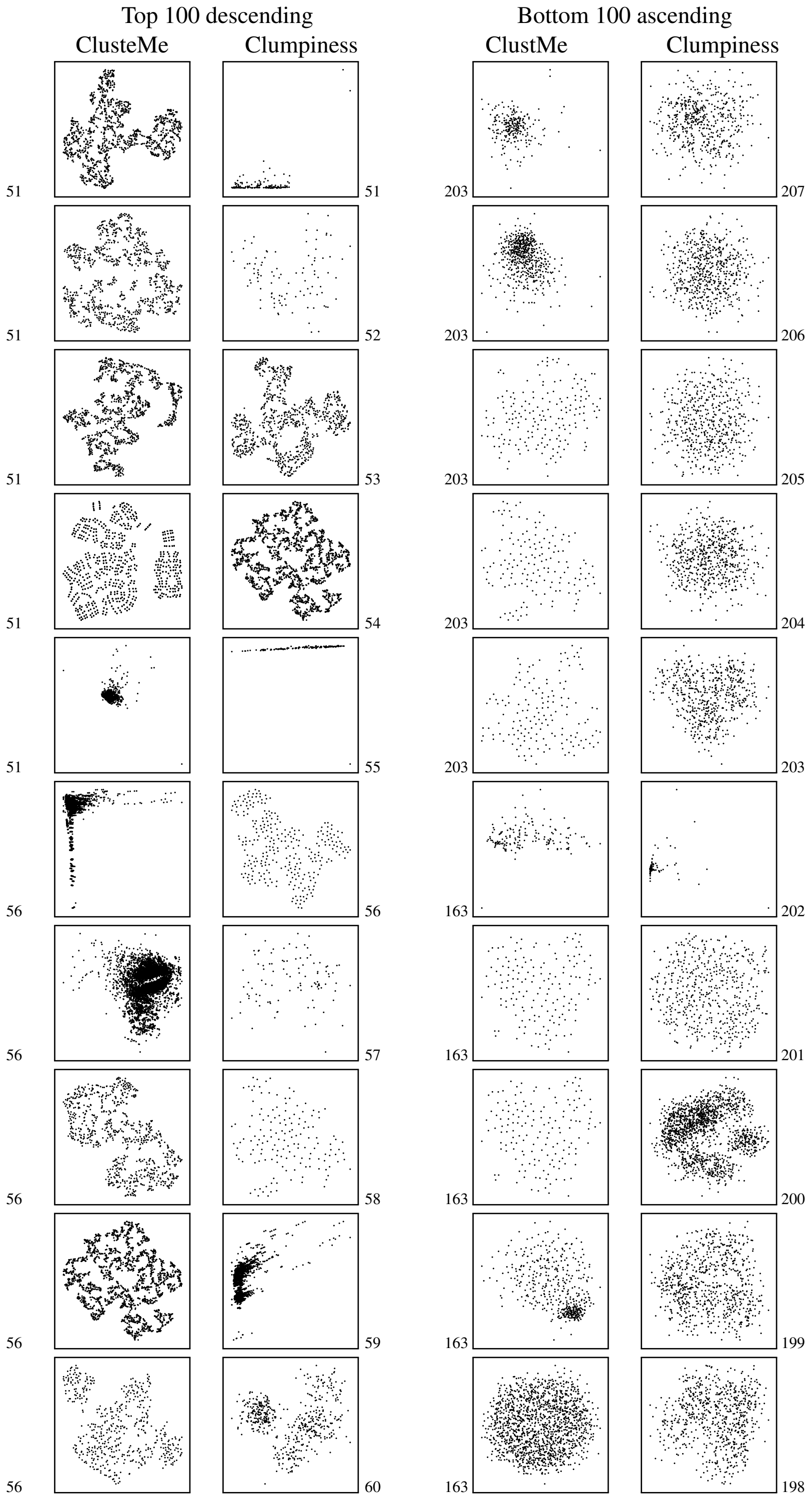

Figure 10: *Series 6/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

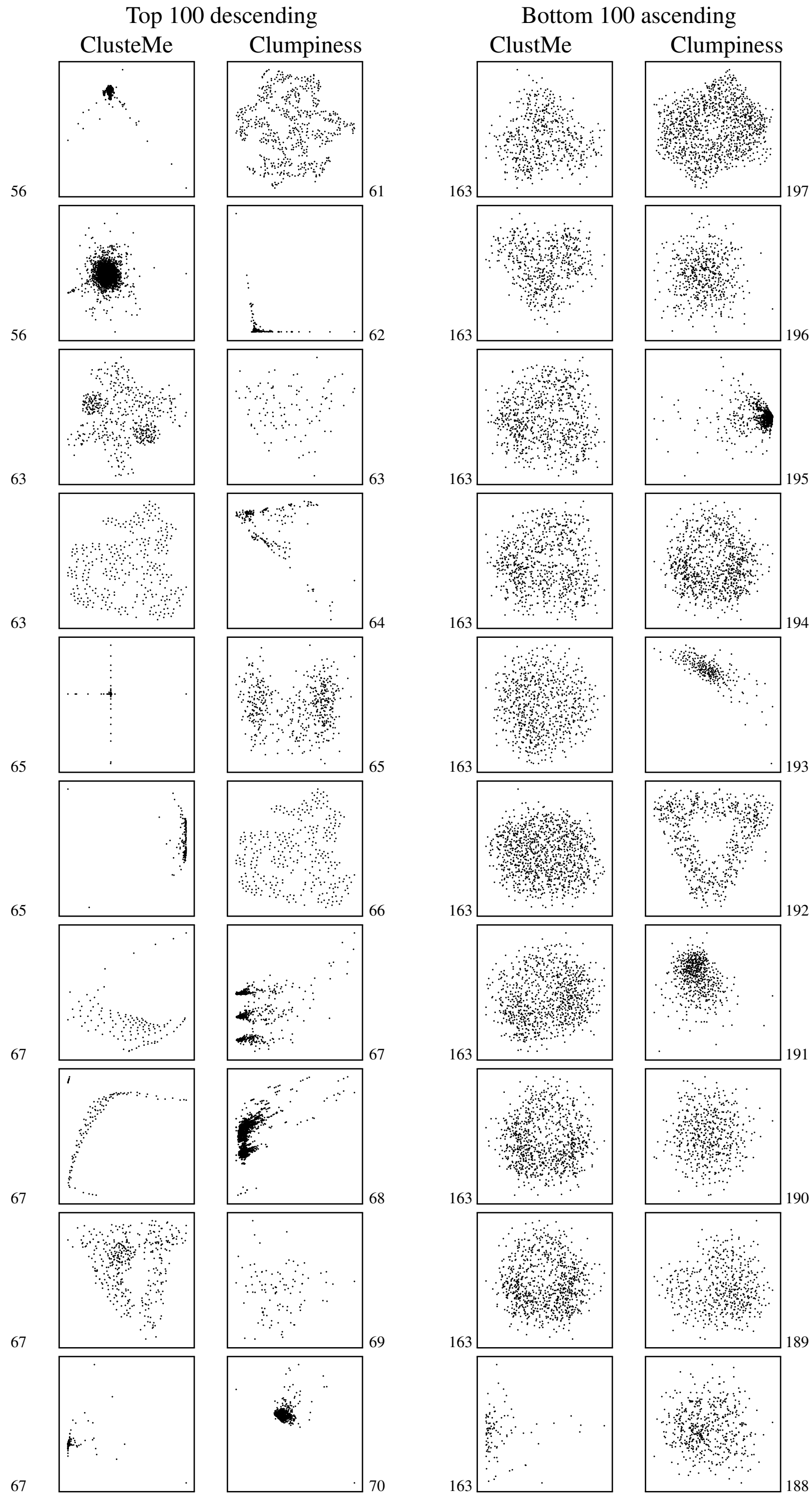

Figure 11: *Series 7/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

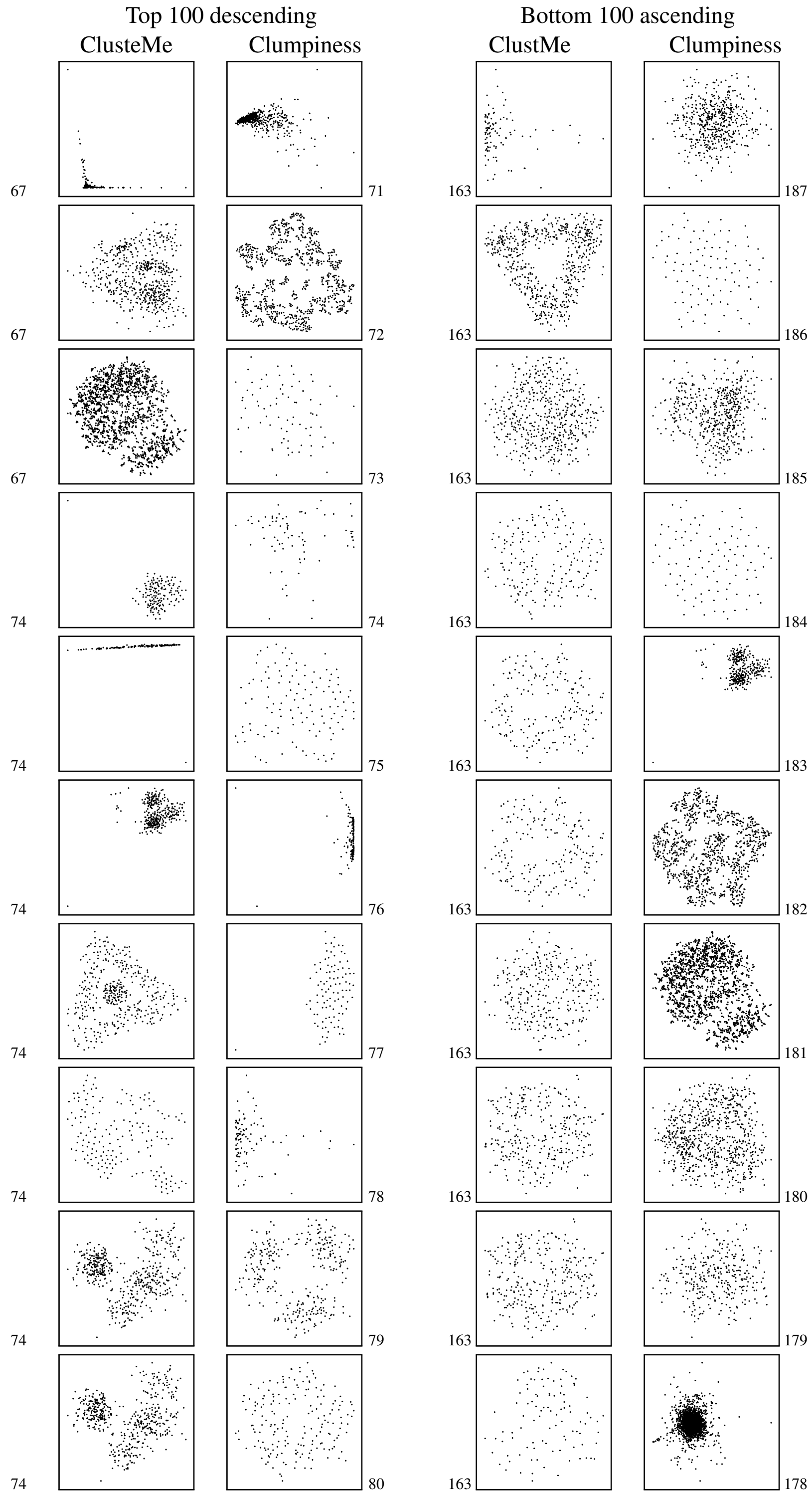

Figure 12: *Series 8/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

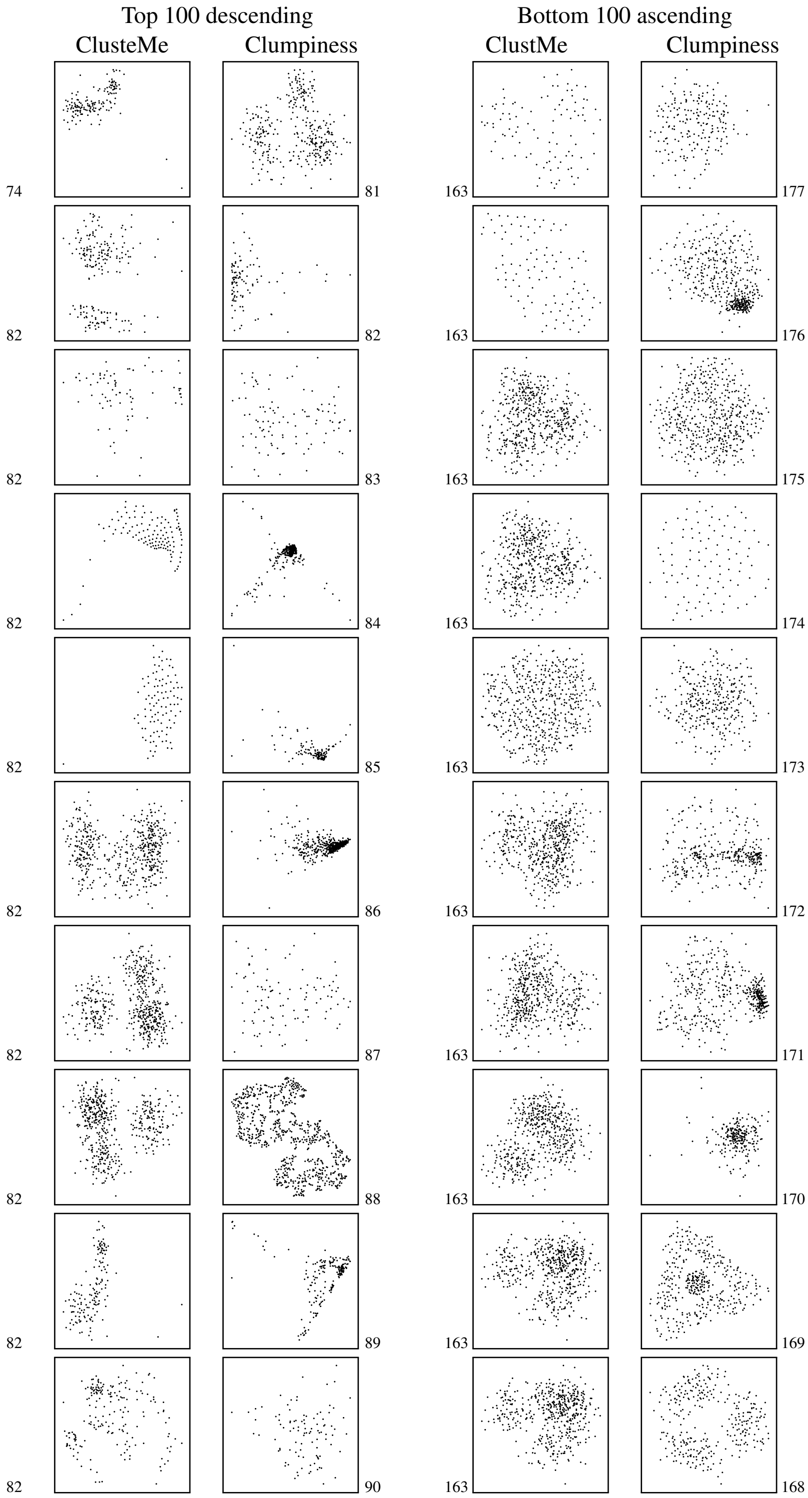

Figure 13: *Series 9/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*

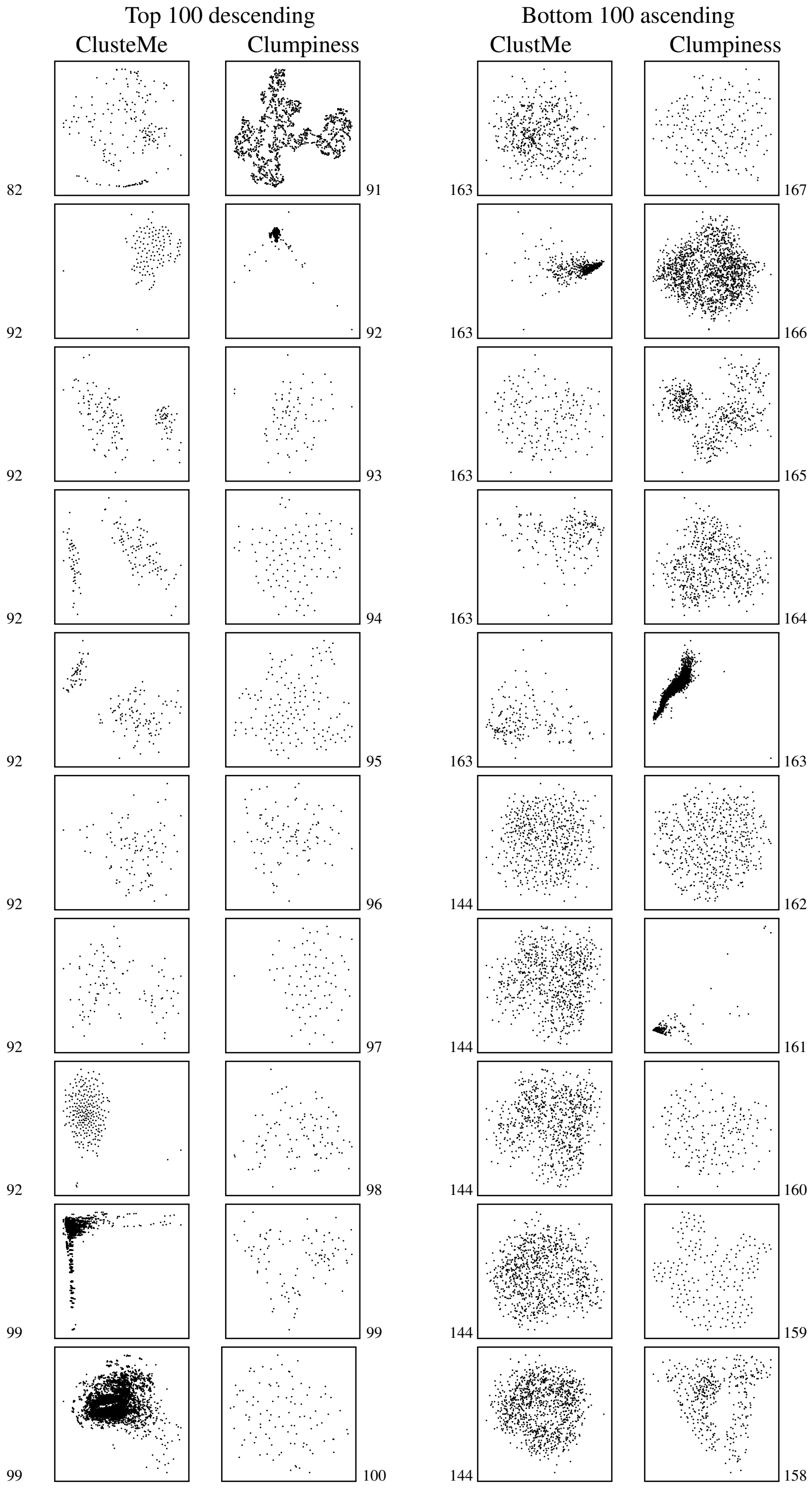

Figure 14: *Series 10/10: Top 100 and bottom 100 scatterplots for the* 257 *benchmark data ranked with ClustMe and Clumpiness.*