10 Results on Pre-classified Data Sets

This chapter has three sections. In the first section, the results of the Databionic swarm (DBS) clustering framework are compared with the given prior classifications for data sets from the Fundamental Clustering Problems Suite (FCPS) [Ultsch, 2005a]. The results for nine data sets analyzed using common clustering algorithms are compared in the first subsection. In the second subsection, the results for data sets with no natural clusters are compared (e.g., Golf Ball). Neighbor Retrieval Visualizer (NeRV) projection and Ward clustering indicate the presence of clusters, whereas DBS does not.
The second section compares Pswarm with other common projection methods using the Delaunay clustering error (DCE). The third section compares emergent self-organizing map (ESOM), swarm-organized projection (SOP) and Pswarm using topographic map visualizations based on the generalized U-matrix for the Wine, Iris, and Swiss Banknotes data sets as well as several FCPS data sets.

10.1 Comparison with Given Classifications

The FCPS [Ultsch, 2005a] is a repository consisting of ten data sets with known classifications. These data sets are intentionally simple enough to be visualized (in 2D or 3D) but nevertheless present a variety of problems that offer good tests of the performance of clustering algorithms [Ultsch/Lötsch, 2016]. The first Figure (10.1) shows the performance of several common clustering algorithms compared with DBS based on 100 trials. The performance is depicted using boxplots of the error rate, which is defined as one minus the accuracy and for which 50% is the level attributable to chance (see chapter 3, Eq. 3.1). Here, the common clustering algorithms considered are single linkage (SL) [Florek et al., 1951], spectral clustering [Ng et al., 2002], the Ward algorithm [Ward Jr, 1963], the Linde-Buzo-Gray algorithm (LBG-k-means) [Linde et al., 1980], partitioning around medoids (PAM) [L. Kaufman/Rousseeuw, 1990] and the mixture of Gaussians (MoG) method with expectation maximization (EM) [Fraley/Raftery, 2002] (also known as model-based clustering).
Aside from the number of clusters, which is given for each of the artificial FCPS data sets, only the default parameter settings of the clustering algorithms were used. ESOM/U-matrix clustering [Ultsch et al., 2016a] and DBscan [Ester et al., 1996] were omitted because no default clustering settings exist for these methods. k-means has the highest overall error rate, and spectral clustering shows the highest variance. The results for the other clustering algorithms vary depending on the data set. DBS has the lowest overall error rate. However, on the Tetra data set, it is outperformed by PAM and MoG; on the EngyTime data set, it is outperformed by MoG; and in the case of the Wing Nut data set, it is outperformed by spectral clustering. Additional statistical tests to Fig 10.1 can be found in supplement I. With the help of insights from chapter 3, Tab. 3101 lists the FCPS cluster structures alongside the algorithms with the best results in terms of the lowest error rate and variance for each data set.

M. C. Thrun, *Projection-Based Clustering through Self-Organization and Swarm Intelligence*, https://doi.org/10.1007/978-3-658-20540-9_10

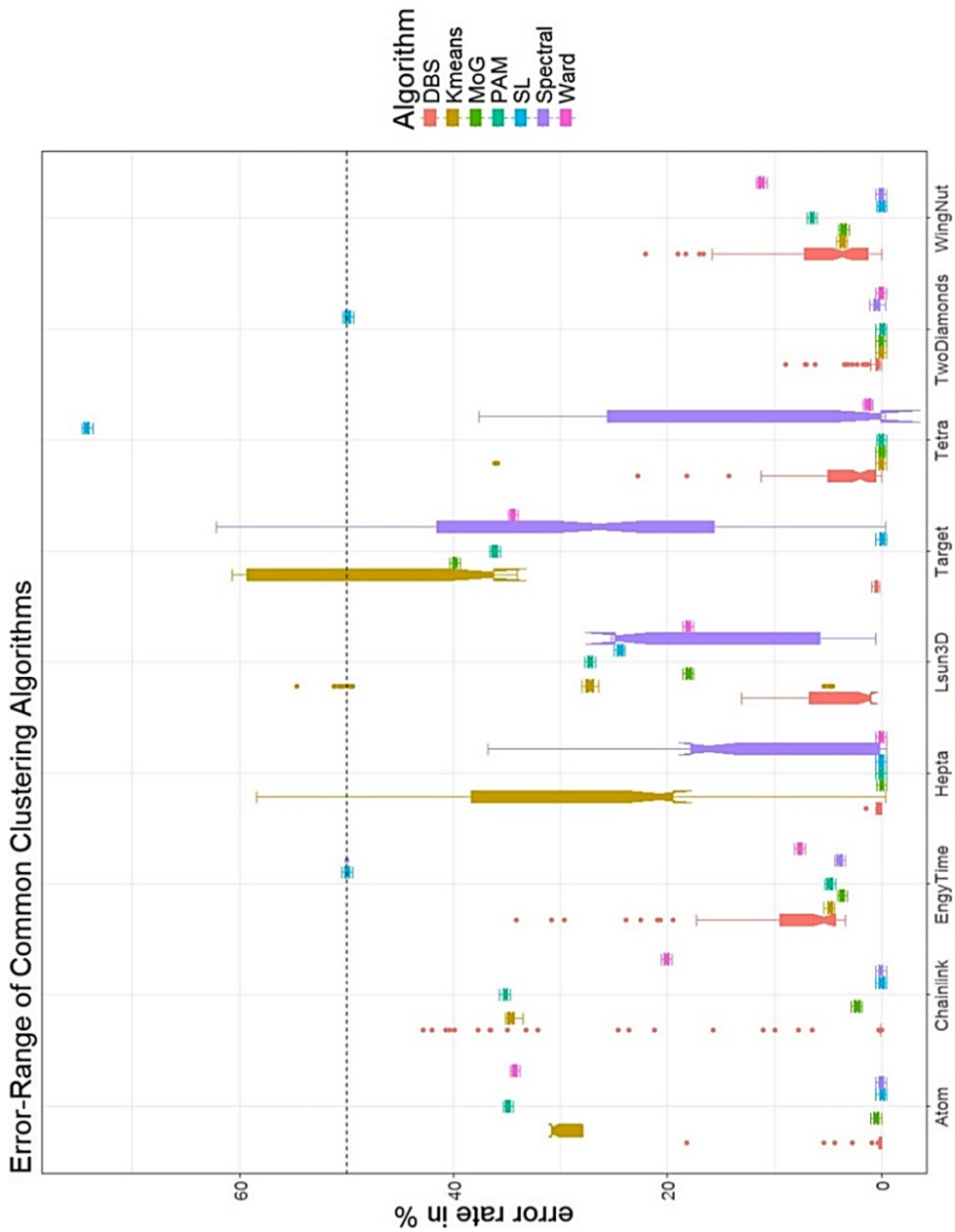

Figure 10.1: Error rate (see p. 107) of 100 trials of common clustering algorithms on nine FCPS data sets, shown as boxplots with the notch as median; chance level at 50%. The interactive clustering approach of DBS was not used here. Abbreviations: single linkage (SL), Linde-Buzo-Gray algorithm (LBG-k-means), partitioning around medoids (PAM), mixture-of-Gaussians clustering (MoG), Databionic swarm (DBS). Additional statistical tests can be found in supplement I.

10.1.1 Recognition of the Absence of Clusters

The Golf Ball data set (see chapter 9) does not exhibit natural clusters. Therefore, it is analyzed separately because, with the exception of SL and the Ward algorithm, the common clustering algorithms give no indication regarding the existence of clusters. This "cluster tendency problem has not received a great deal of attention but is certainly an important problem" [Jain/Dubes, 1988, p. 222]. Reproducing the results of [Ultsch/Lötsch, 2016], the Ward algorithm indicates six clusters, whereas SL indicates two clusters (Figure 10.2). As seen from the two dendrograms generated using DBS, the connected approach does not indicate any clusters, whereas the compact approach indicates four clusters (Figure 10.3). However, the presence of four clusters is not confirmed by the topographic map of DBS.

In Figure 10.4, the topographic maps of DBS with the NeRV are compared. The NeRV projection of the Golf Ball data set with $\lambda = 0.5$ (for the other parameters, see the R package projections), i.e., with precision and recall weighted equally, is shown in Figure 10.4 (top). The visualization of the NeRV projection strongly indicates a two-cluster structure, whereas the DBS projection does not (Figure 10.4, bottom). The compact DBS clustering divides the data points lying in valleys into different clusters and merges the data points into clusters through hills, resulting in cluster borders that are not defined by mountains.

The topographic map of DBS of the S-shape data set and the uniform and Gaussian Cuboid data sets (see chapter 9) are also shown in supplement D, Figure D.19. Neither data set contains any natural clusters; this is correctly visualized using the DBS approach.

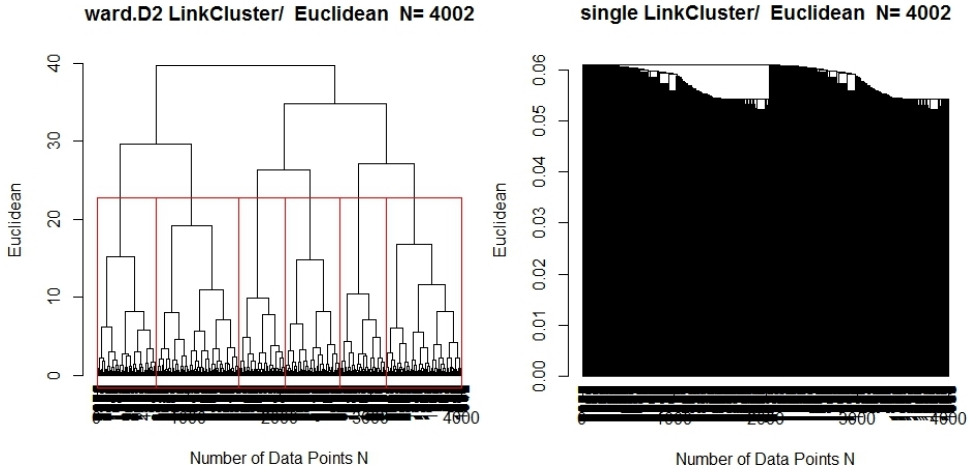

Figure 10.2: The dendrogram generated using the Ward algorithm indicates at least two clusters with a high intercluster distance. The SL dendrogram could indicate two clusters with a very low intercluster distance.

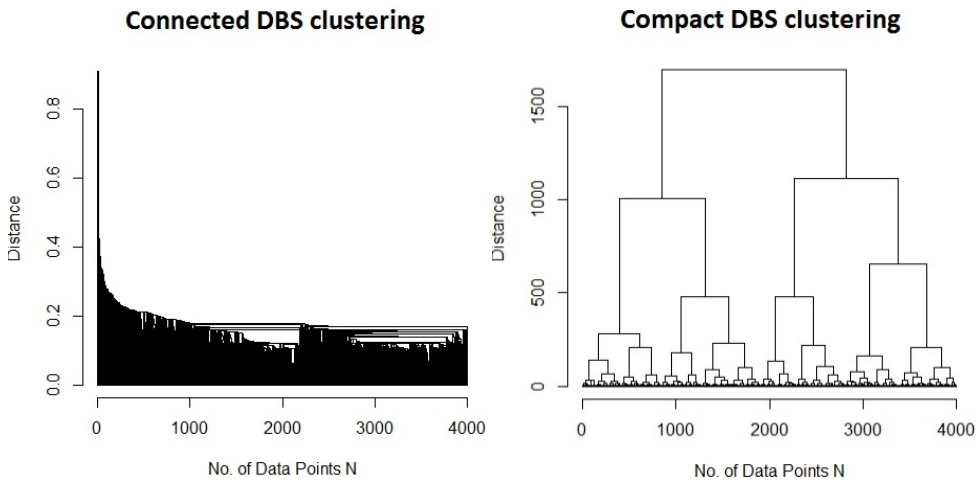

Figure 10.3: The two dendrograms generated using DBS. The connected DBS clustering does not indicate any structure whereas the compact DBS clustering indicates two or four clusters. The connected approach does not indicate any clusters, whereas the compact approach does indicate four clusters. However, Figure 10.4 shows that these clusters are inconsistent with the visualization.

10.2 Evaluation of Projections Using the Delaunay Classification Error (DCE)

Figure 10.5 shows the results for the DCE measure, relative to the baseline, for 100 trials of the common projection methods ESOM, NeRV, Sammon mapping (a multidimensional scaling (MDS) technique), curvilinear component analysis (CCA), principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). Positive values indicate higher errors compared with the baseline, whereas negative values indicate lower errors. The baseline is the NeRV projection with $\lambda = 0.5$ and PCA initialization; this baseline was chosen because the outcome of this initialization is deterministic (for the other parameters, see the R package projections). The parameter setting $\lambda = 0.5$ indicates that precision and recall are weighted equally. Every subfigure shows a robust mean estimate M and a robust standard deviation estimate S for the 100 relative DCEs. Notably, it is claimed that t-SNE projections are similar to NeRV projections with $\lambda = 1$ [Venna et al., 2010].

The linear method PCA and the MDS technique of Sammon mapping are unable to separate the connected structures of the Chainlink and Atom data sets based on their assumed neighborhood relations. This result confirms the assumptions made in chapter 4. By contrast, the CCA projections have difficulty separating compact structures based on intra- versus intercluster distances. However, not all focusing projection methods are able to separate connected structures, e.g., the t-SNE projections of Chainlink.

Without the U-matrix, the ESOM projection method distributes the points uniformly, which results in a higher DCE. The projections generated by t-SNE, Pswarm and NeRV with their default settings show high variances, although the variance in the accuracy of the DBS clustering results for these data sets is low (Figure 10.1).

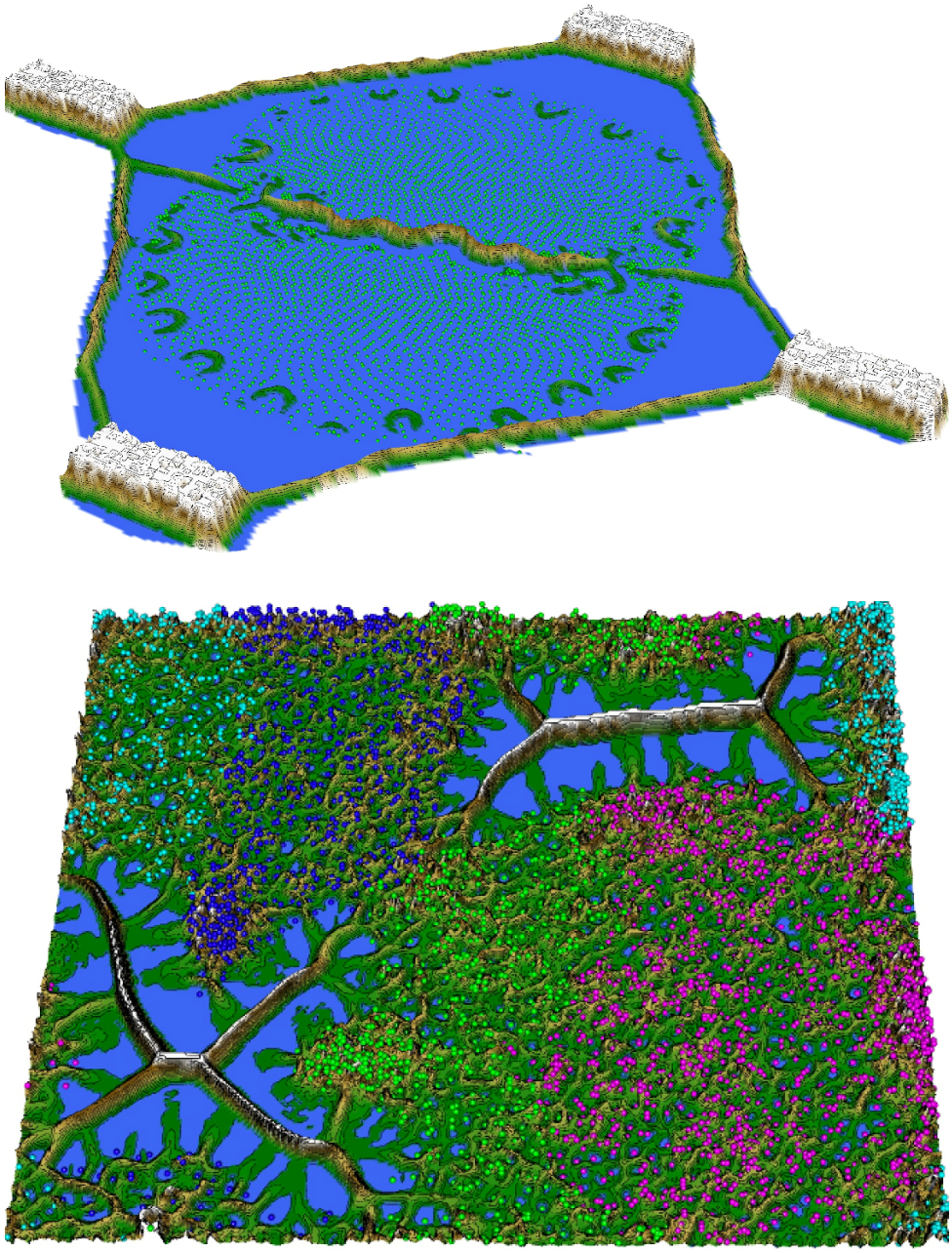

Figure 10.4: **Top**: Topographic map of the NeRV projection ($\lambda = 0.5$) of the Golf Ball data set indicates two well-separated clusters.
Bottom: The topographic map of the DBS projection and (compact) clustering of the Golf Ball data set. The projection does not indicate a cluster structure. The DBS clustering generates clusters that are not separated by mountains. No island can be extracted from the toroidal visualization.

Statistical testing was performed using the two-sample, one-sided Wilcoxon rank sum test with continuity correction [Hollander/Wolfe, 1973, pp. 68–75]. The DCE values for the Pswarm projections were compared with the projections obtained using the other methods with the "nearest"[70] ranges of DCE values "above" and "below" those of Pswarm (visually in the 90° rotated figures). In the former case, means that the DCE values of Pswarm are more negative (shifted to the left) compared with the DCE values of the projection method with the nearest range of values. Consequently, a significant result means that Pswarm's performance is considerably better. In the latter case, the DCE values of Pswarm are more positive (shifted to the right), and a significant result means that Pswarm's performance is worse than that of the projection method with the nearest range of DCE values "below" those of Pswarm. Statistical results regarding the performance of Pswarm in Figure 10.5 are as follows.

1.) **Atom**: The performance of Pswarm is significantly better than that of NeRV, with $W(100) = 1675$, $p < 0.001$, and worse than that of t-SNE, with $W(100) = 5795$, $p = 0.026$.
2.) **Hepta**: The performance of Pswarm is significantly better than that of CCA, with $W(100) = 1855$, $p < 0.001$, and worse than that of NeRV, with $W(100) = 8941$, $p < 0.001$.
3.) **Lsund3D**: The performance of Pswarm is significantly better than that of t-SNE, with $W(100) = 4145$, $p < 0.02$, and not significantly worse than that of CCA, with $W(100) = 5444$, $p = 0.14$. However, the performance of Pswarm is significantly worse than that of NeRV, with $W(100) = 7969$, $p < 0.001$.
4.) **Chainlink**: The performance of Pswarm is significantly better than that of NeRV, with $W(100) = 2472$, $p < 0.001$, and worse than that of CCA, with $W(100) = 6270$, $p = 0.001$.
5.) **Tetra**: The performance of Pswarm is significantly better than that of CCA, with $W(100) = 2879$, $p < 0.001$, and not significantly worse than that of ESOM, with $W(100) = 5000$, $p = 0.5$.

10.3 Topographic Maps with Hypsometric Colors

To compare Pswarm as a projection method with SOP and ESOM, the data sets of [Herrmann, 2011, pp. 99-100] were used. After the computation of several trials based only on the visually best[71] scatter plot, topographic maps with hypsometric colors (hypsometric tints) were generated. The Atom, Chainlink, EngyTime, Iris, Swiss Banknotes, and Wine data sets were projected using SOP, ESOM and Pswarm and visualized using the U-matrix or generalized U-matrix approach.

Figure 10.6 shows that only the colored labels corresponding to the prior classification separate the two clusters of EngyTime. The topographic map is inconsistent with the projected points in terms of lattice locations. Moreover, the separation is blurry, and several points are misplaced. Notably, the cardinality of the data set is 4096, and there are only 4096 positions on a 64x64 lattice. However, the visualization presented in Figure 10.6 shows many empty positions. Consequently, there are many positions at which more than one DataBot is located; therefore, the colored labels could be misleading, and the quality measures of [Herrmann, 2011] could be incorrect.

[70] With the highest overlap in $M \pm S$. It is assumed that non-overlapping ranges of DCE values are always statistically significant.
[71] In the sense that the structures defined by the prior classification were preserved.

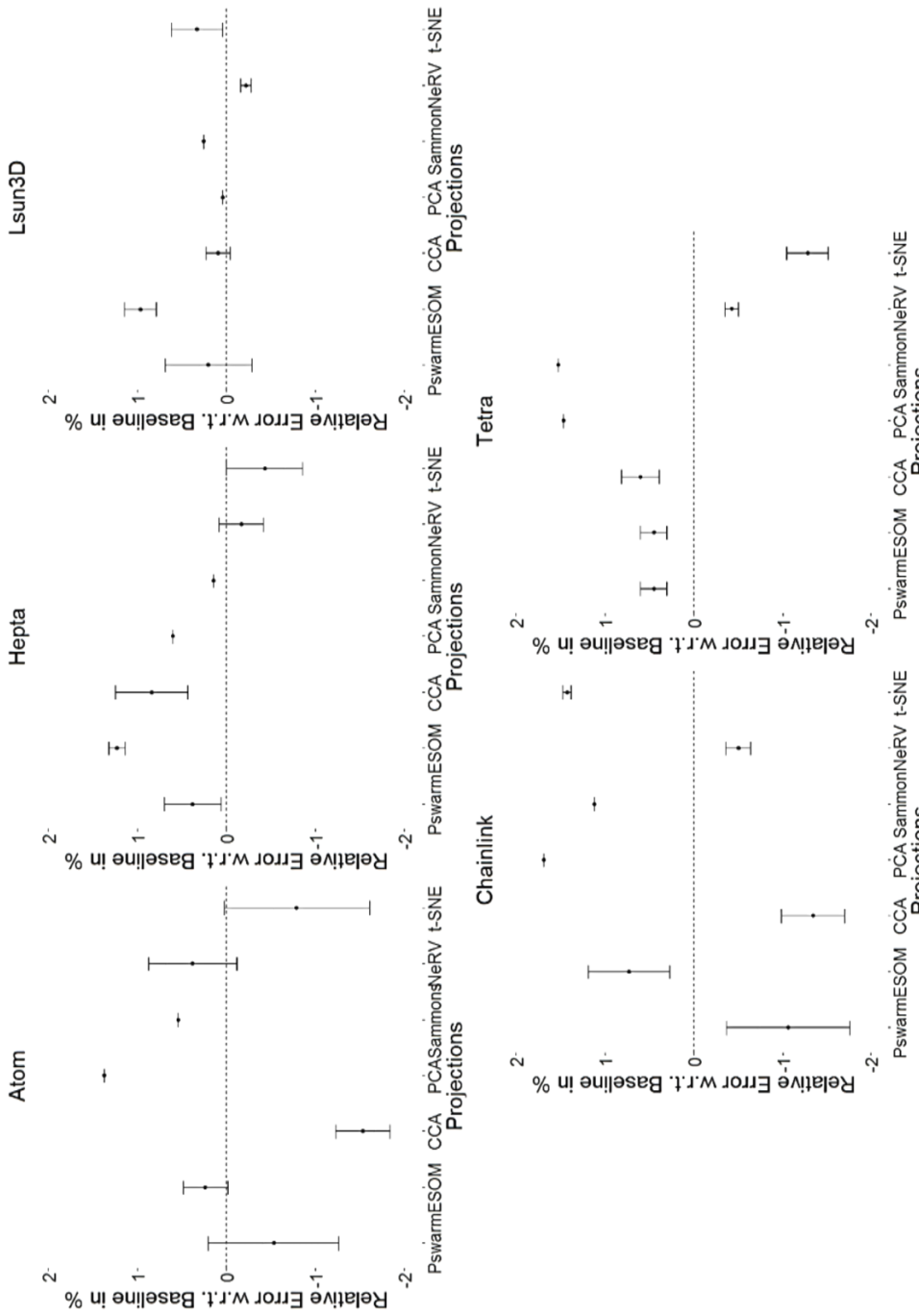

Figure 10.5: Relative DCE values for projections of the Atom, Hepta, Lsun3D, Chainlink and Tetra data sets. The following seven methods are compared: Pswarm ESOM, CCA, PCA, Sammons mapping, NeRV and t-SNE. The most structure-preserving projections have the lowest negative values. No projection method is able to outperform any other projection method on five all data sets.

Table 10.1: Cluster structures in the artificial benchmark sets of the FCPS [Ultsch, 2005a], as defined in chapter 2. The clustering algorithms with the lowest error rate and variance in Figure 10.1 are listed for each data set. These results confirm the assumptions discussed in chapter 3 regarding the cluster structures sought by common clustering algorithms. On the right the projection methods who were unable to find the structure are listed for the three-dimensional data sets. ESOM method is omitted, because it distributes the projected points uniformly. Additional statistical tests can be found in supplement I.

Data Set	**Cluster Structure**	**Clustering Algorithms that Found this Structure with a Small Variance in the Results**	**Projection Methods that did not Found this Structure**
Atom	Connected, direction-based, varying density, non-linear separable	DBS, MoG, SL, Spectral	NeRV, Sammon's mapping and PCA
Chainlink	Connected, direction-based, non-linear separable	DBS, SL, Spectral, (MoG)	t-SNE, Sammon's mapping and PCA
EngyTime	Connected, unidirectional, varying density	All except SL	
Hepta	Compact, spherical, high intercluster distance	DBS, MoG, PAM, SL, Ward	CCA
Lsun3D	Compact, ellipsoidal, outliers	DBS	t-SNE
Target	Connected, direction-based, outliers	DBS, SL, Spectral	
Tetra	Compact, spherical, low intercluster distance	All except SL and Spectral	PCA and Sammons mapping
Two Diamonds	Compact, spherical, borders defined by discontinuity	All except SL	
Wing Nut	Connected, direction-based, linear separable	DBS, SL, Spectral	
Golf Ball	No natural clustering tendency	DBS	

By contrast, in the topographic map of the Pswarm projection shown in Figure 10.7, the clusters are clearly separated by both the positions of the projected points and the high-dimensional distances and densities of the generalized U*-matrix. Here, only one DataBot is allowed per grid position. In comparison to Figure 10.7, the planar ESOM/U*-matrix projection presented in Figure 10.8 does not clearly show the border between the two clusters. As shown in Figure 10.9, when the default settings (toroidal) are used, it is difficult to distinguish between the two clusters. Because the extraction of an island was not possible, a tiled display is shown in Figure 10.9. Likewise, for the Wing Nut data set, the topograpic map of the Pswarm projection shows a clear cluster structure, whereas the toroidal ESOM/U-matrix projection does not (Figure 10.10 and supplement E, Figure E.23) when the P-matrix and U*-matrix visualization is not used.
On the Iris data set, the topographic map of the generalized U*-matrix of the SOP result shows three clusters that are clearly separated by hills, but these clusters do not match the colored labels of the prior classification (supplement C, Figure C.13). By contrast, the Pswarm projection visualized using the generalized U*-matrix approach does show these clusters, one of which is defined by its density (supplement C, Figure C.14). Five points are misplaced. The ESOM/U-matrix method is unable to separate two of the three clusters (supplement E, Figure E.22).

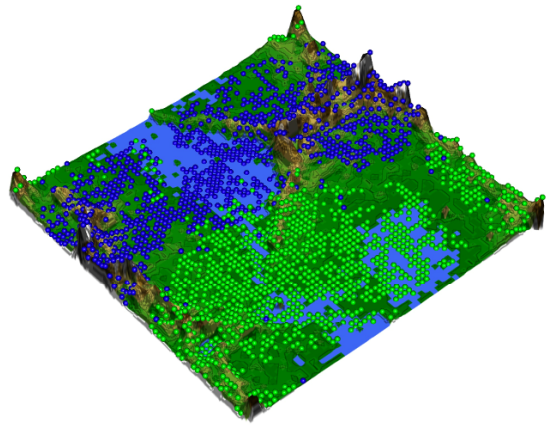

Figure 10.6: Topographic map of the EngyTime data set projected using SOP with the default parameters: The two clusters are mixed and difficult to separate without the colored labels corresponding to the classification. The radius of the P-matrix was automatically chosen to be 1.38. No island could be extracted.

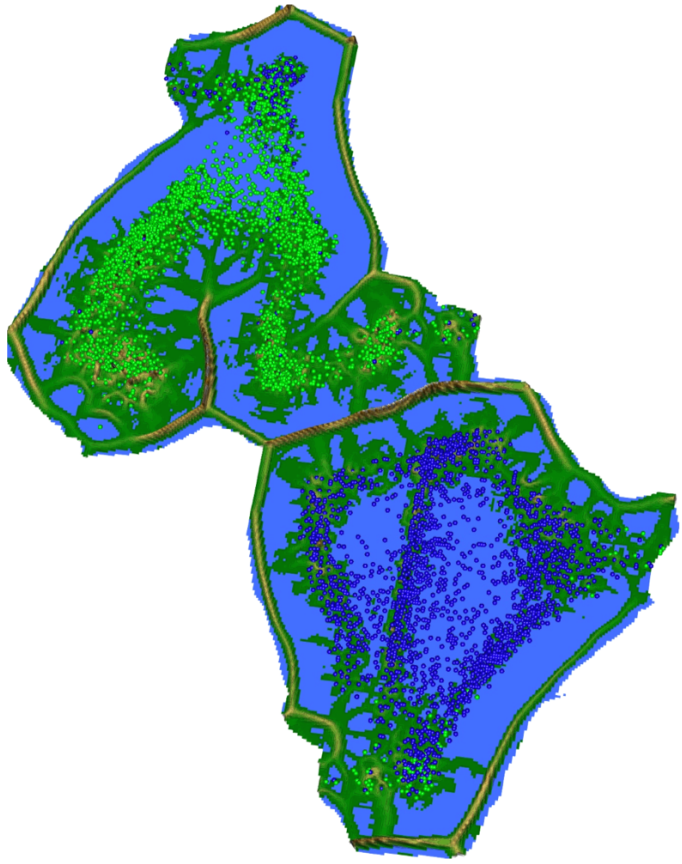

Figure 10.7: Topographic map of the EngyTime data set projected using DBS (196x220) with an automatically chosen lattice size: There are clearly two clusters with an accuracy of the DBS clustering of 95%

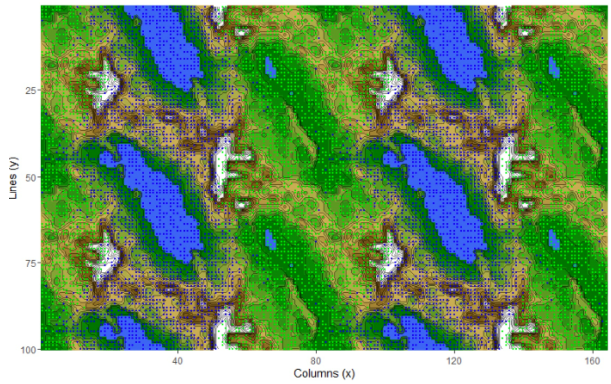

Figure 10.8: U*-matrix visualization of the toroidal ESOM projection of the EngyTime data set: The data set contains 4096 observations, and the lattice contains 4096 neurons. As shown, not every neuron is a best matching unit (BMU); therefore some BMUs include more than one observation, and the colored labels are misleading. The clusters are mixed, and no border between the green and blue BMUs can be found.

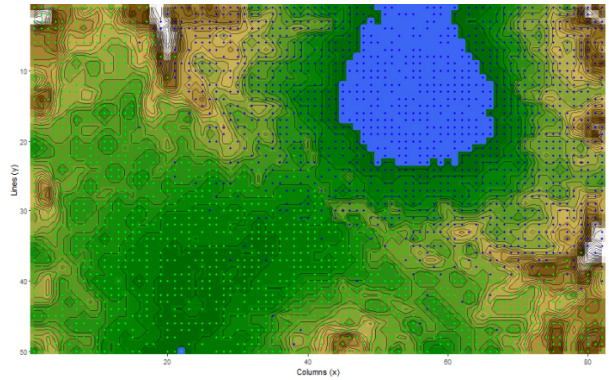

Figure 10.9: U*-matrix visualization of the planar ESOM projection of the EngyTime data set: The data set contains 4096 observations, and the lattice contains 4096 neurons. As shown, not every neuron is a best matching unit (BMU); therefore, some BMUs include more than one observation, and the colored labels are misleading. The clusters are mixed, and a border between the green and blue BMUs is difficult to locate.

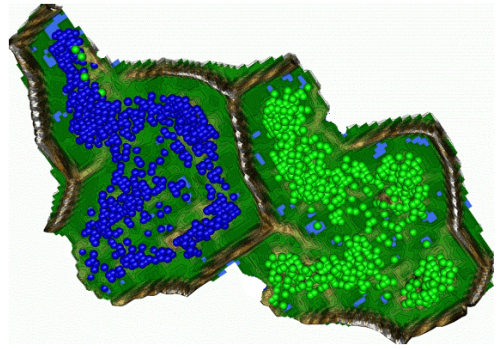

Figure 10.10: Topographic map of the DBS projection of the Wing Nut data set with Generalized Umatrix (64x68). Both clusters are clearly separated, but four points are misplaced.

The topograpic map of the Swiss Banknotes data set as projected using SOP shows three clusters based on high-dimensional distances in the generalized U-matrix, with one misplaced point (supplement C, Figure C.9). Without the topographic map, a scatter plot of the projected points would not lead the reader to the conclusion that the data set consists of separate clusters because the projected points defined by the DataBots are uniformly distributed. By comparison, Pswarm reveals two unambiguously separated clusters with two misplaced points (supplement C, Figure C.10). In the ESOM/U-matrix projection, one best matching unit is misplaced. The cluster of blue best matching unit could be interpreted as two clusters, one small and one large, based on the high hills in between (supplement E, Figure E.21).An interpretation of the uniformly distributed projected points of the Wine data set, as generated via SOP, does not allow the number of clusters to be determined (supplement C, Figure C.11). The generalized U-matrix shows no clear borders between projected points with differently colored labels. Several points are misplaced. By contrast, the topographic map of the Pswarm projection explicitly shows three clusters (supplement C, Figure C.12). — one triangular, one rectangular and one square — but six points are misplaced. In the ESOM/U-matrix projection, the clusters in the Wine data set are difficult to separate without their colored labels (supplement E, Figure E.20). Again, in the SOP result for the Atom data set, the clusters are distinguished only by the borders of the generalized U-matrix and the colored labels corresponding to the prior classification because the points are uniformly distributed (supplement C, Figure C.15). However, the visualization could also be misleading in suggesting that the data set consists of three clusters. The topographic map of the Pswarm projection explicitly shows two clusters (supplement C, Figure C.16).The projections of the Chainlink data set obtained using both SOP and Pswarm are similar (supplement C, Figure C.17) but the Pswarm visualization is smoother in terms of intracluster structure (supplement C, Figure C.18).