M. C. Thrun, F. Pape, A. Ultsch

Benchmarking Cluster Analysis Methods using PDE-Optimized Violin Plots

R packages at www.deepbionics.org

European Conference on Data Analysis (ECDA), Germany, 4th - 6th July, 2018

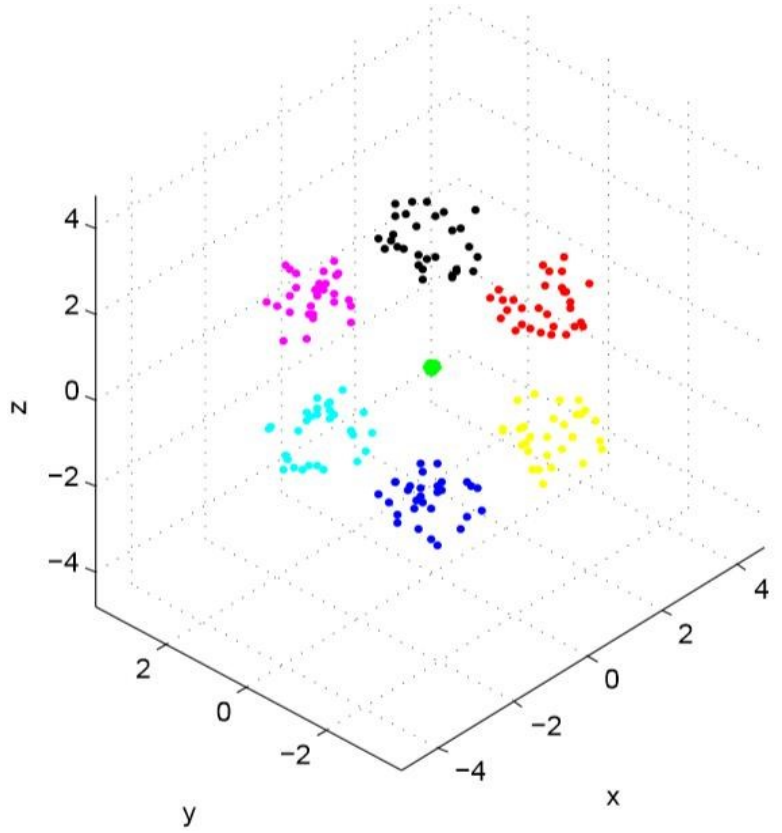
Motivation

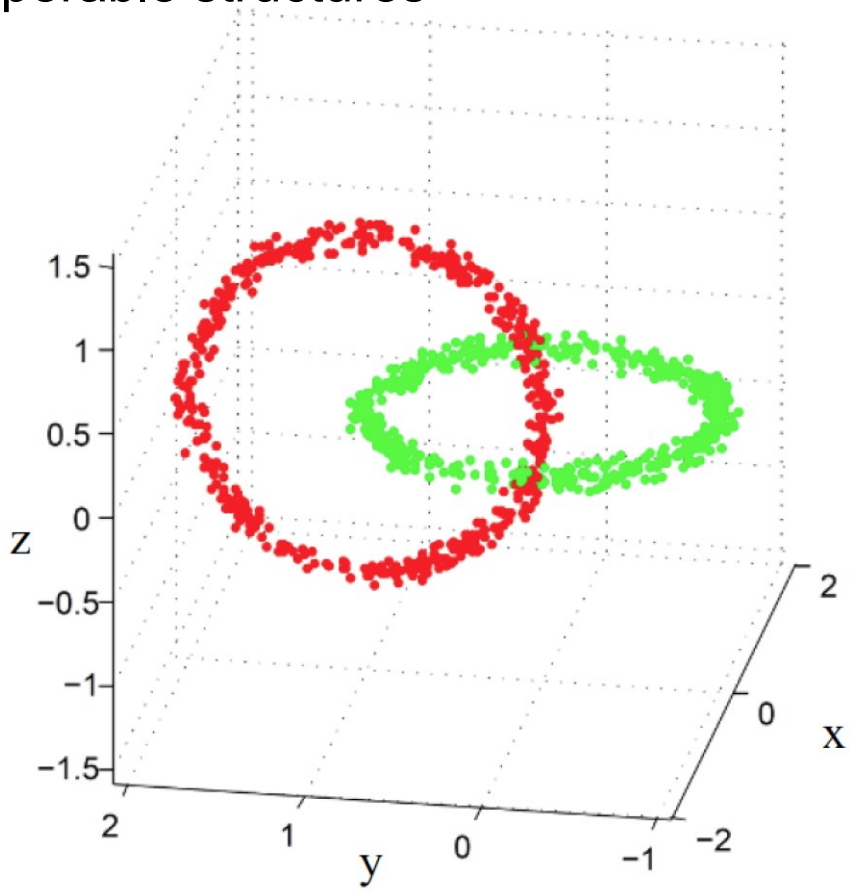
Separate data into similar groups -> **Clustering**

-> detect meaningful cluster structures defined by a clustering

-> Interested in distance and density-based structures

-> Linear seperable and linear non-seperable structures





Background

- Clusters can be of arbitrary shapes (structures) (1)
  - No generally accepted definition of clusters exists in the literature (2)
  - Number of clusters difficult to estimate

Implicit assumptions about structures of data are made by

- Clustering criterions (3)
- Projection methods (besides ESOM and Pswarm of DBS) (5)
- Quality measures (QMs) for projection methods
  - My other talk: Investigating Quality Measures of Projections for the Evaluation of Distance and Density-based Structures of High-Dimensional Data
- Quality assessments for clustering methods in the case of unknown class labeles (4)

(1) [Jain/Dubes, 1988]; (2) [Hennig et al., 2015, p. 705]; (3) [Duda et al., 2001; Everitt et al., 2001; Handl et al., 2005; Theodoridis/Koutroumbas, 2009; Ultsch/Lötsch, 2016]; (4) [Handl et al., 2005]; (5) [Thrun, 2018]

Challenges of Cluster Analysis

In this talk:

1. How reproducible are the structures a clustering algorithm finds?
2. Can any cluster algorithm find any structure type in data?
3. How to chose the right parameter settings?
   - e.g. Spectral Clustering
     - State of the art: test all possible parameter settings (1)

Example in my other Talk: Knowledge discovery from low-frequency stream nitrate concentrations: hydrology and biology contributions

4. How can a cluster analysis be performed on a data set of unknown structures without prior assumptions?
5. Does the structure defined by a cluster algorithm lead to plausible insights?

(1) [Wiwie et al., 2015]

First step: Benchmarking

- Start with artificial datasets with ground truth
  - Define structures priorly using 2D and 3D datasets

  -> FCPS provides a good start of datasets (1)
- Should be done unbiased
  - Use default settings
  - Use an as simple as possible evaluation method
    - Experience shows that more elaborate quality measures are often biased

      => State of the art: Use all supervised indices available (2)
  - Compare many trials per algorithm to each other
- Why not natural and high-dimensional datasets?
  - Structures are difficult to know beforehand
  - May have more than one clustering solution depending on
    - Domain expert
    - Application based

(1) [Ultsch, 2005], (2) [Wiwie et al., 2015]

Defining Unbiased Quality Measure

$$Accuracy = \frac{\text{No. of true positives}}{\text{No. of cases}}$$

1. Calculate 100 trials per clustering method
   - For each trial
     - The best of all permutation of labels with the highest accuracy is selected
       - because algorithms arbitrarily define the labels of a clustering
2. Apply Distribution analysis
   - Univariate
     - QuantileQuantile plot
     - Histogram
     - Evaluation of cdf or pdf
   - Multivariate
     - Methods above are difficult to visualize
     - Box-Whisker diagrams (box plot)
     - Violin plots

Why not box plot?

- Visualizes the number of values in a specific range
  - End of the two whiskers are proportional to the interquartile range (often 1.5*IQR), (1)
  - The box marks 25 and 75% percentile
- Does not indicate multimodality or if median is valid
- Estimates of underlying distribution quantiles based on one or two order statistics
- At least nine different approaches for estimation (2)
  - Assumption about distribution are made

(1) [Tukey, 1977]; (2) [Hyndman/Fan, 1996]

Why not violin/bean plot (1)?

- Univariate density estimation is trying
- Clear model behind density estimation required
  - Emphasis on multimodality

-> Estimation of pdf called “Pareto Density Estimation (PDE), (2)

  - Kernel density estimation with variable radius
    - Representing the relative likelihood of a given variable taking on specific values
    - Slivered in kernels with a specific width
      - this width, and therefore the number of kernels, depends on the data
  - Particularly suitable for the discovery of structures in continuous data
    - Allows the discovery of mixtures of Gaussians

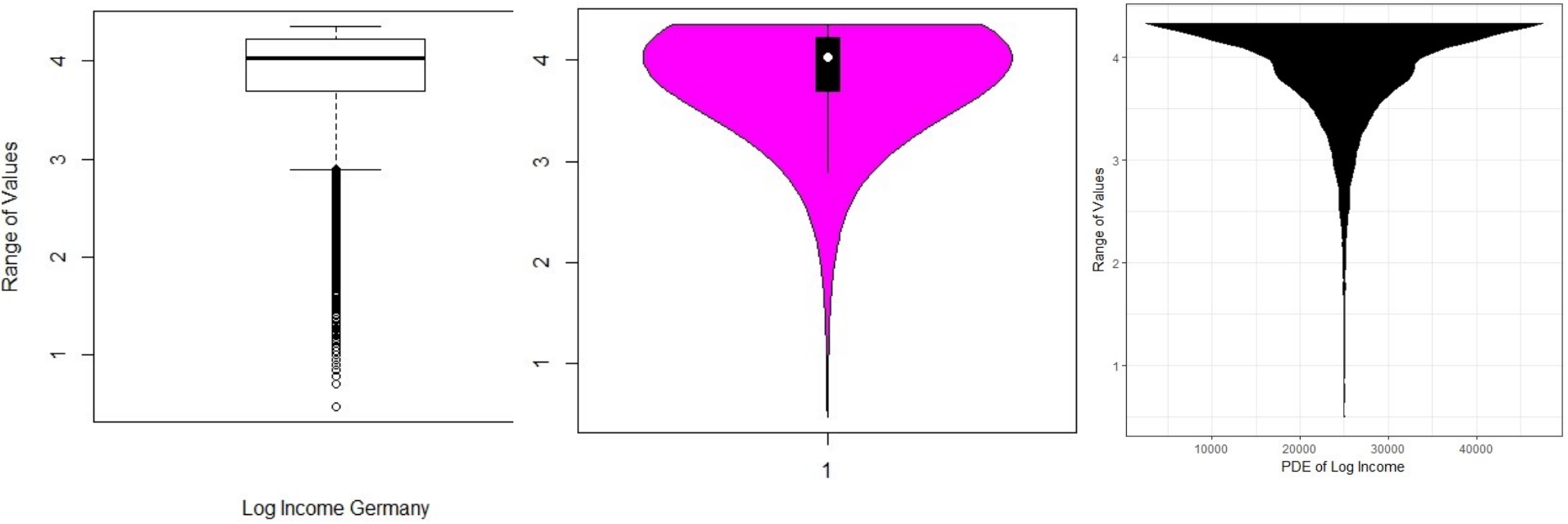
-> Pareto density estimation (PDE) is used to improve the violin, or the so-called bean plot

(1) [Hintze/Nelson, 1998]; (2) [Ultsch, 2005b]

Skewed Distribution

- Data: Peoples income in Germany (1)
- Left: Boxplot
- Middle: Violin plot (2)
- Right: PDE-optimized violin plot

-> Boxplot and violin plot underestimate skewness of distribution

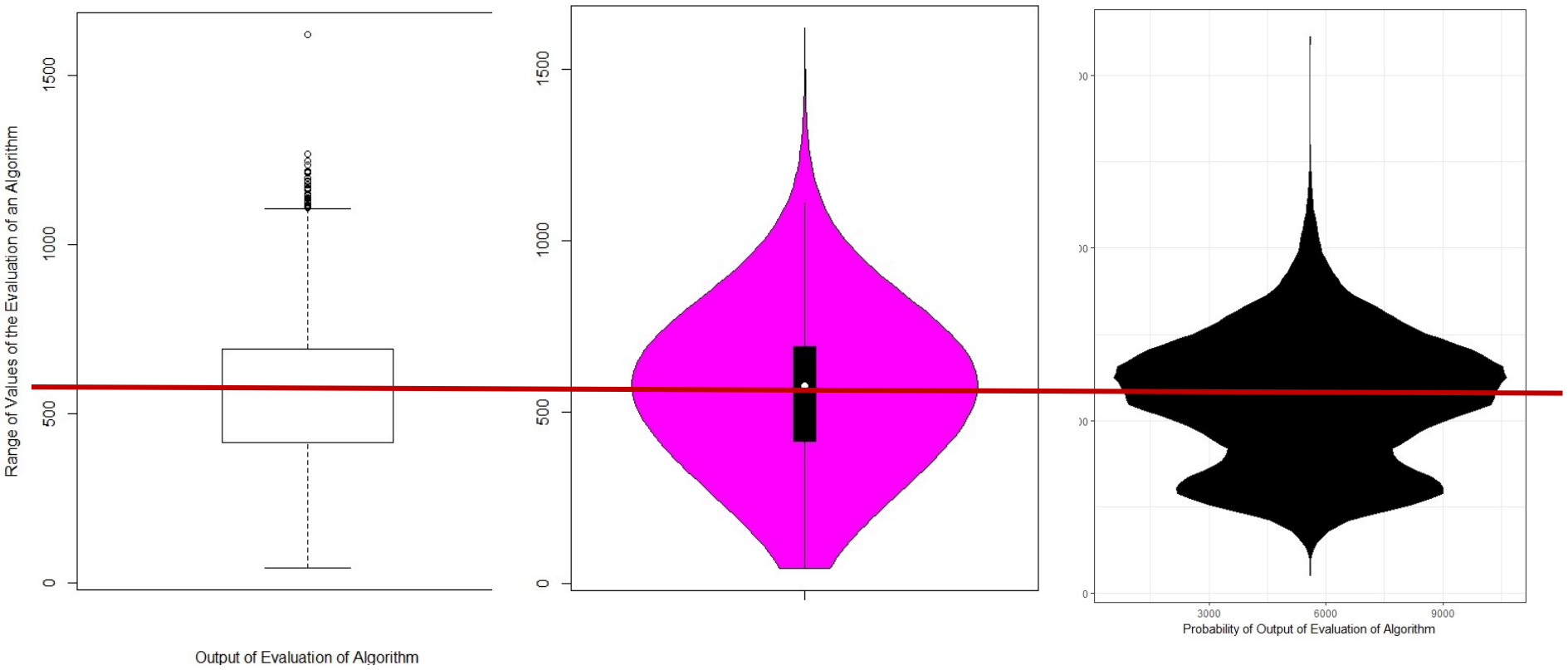


(1) [Thrun/Ultsch, 2015]; (2) [Hintze/Nelson, 1998];

Multimodal Distribution

- Data: Income Tax Share of German municipalities (1)

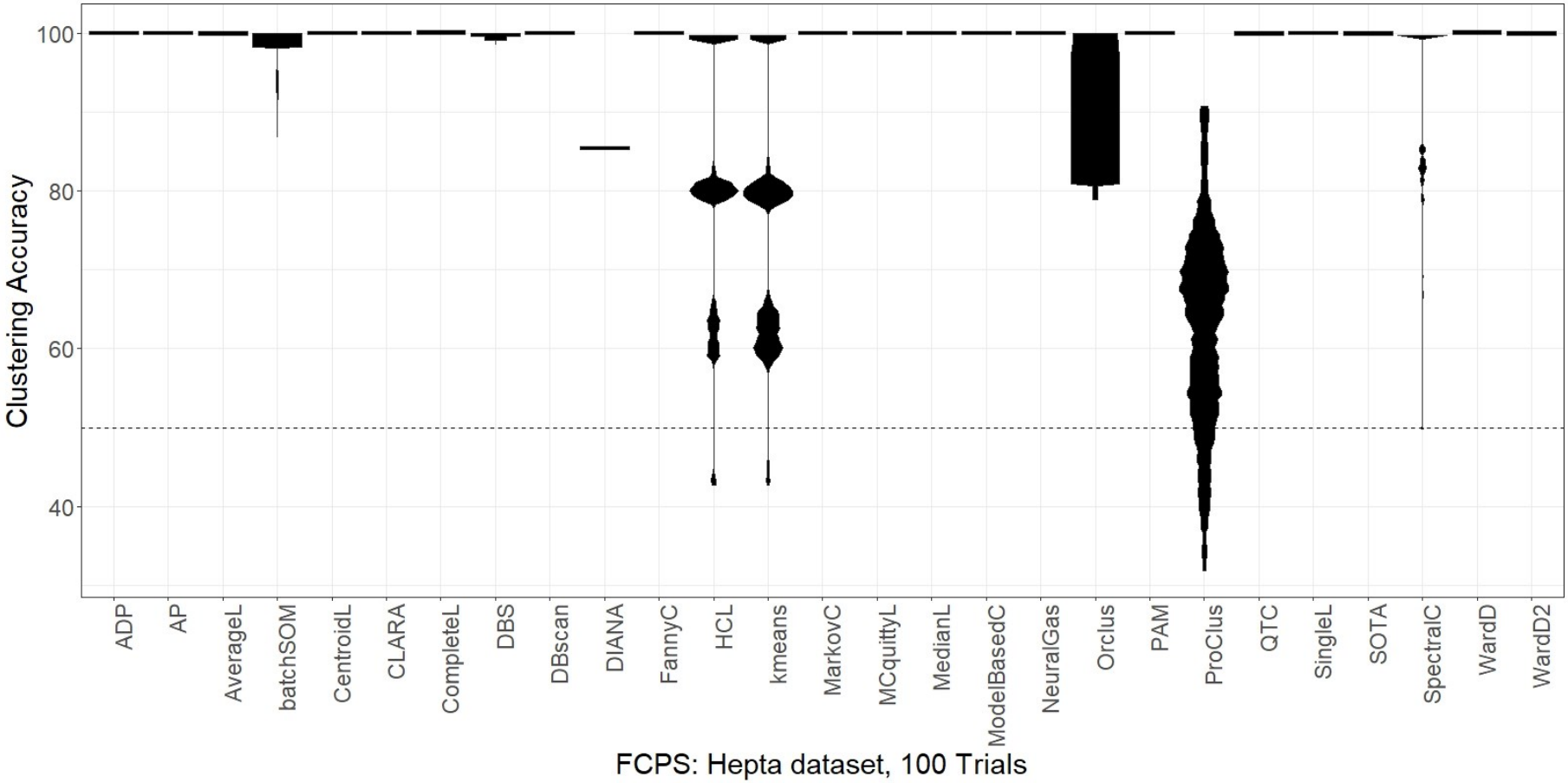
=> Multimodality is given (2), but only PDE-optimized violin plot finds it!



(1) [Ultsch/Behnisch, 2017], (2) [Thrun/Ultsch, 2018]

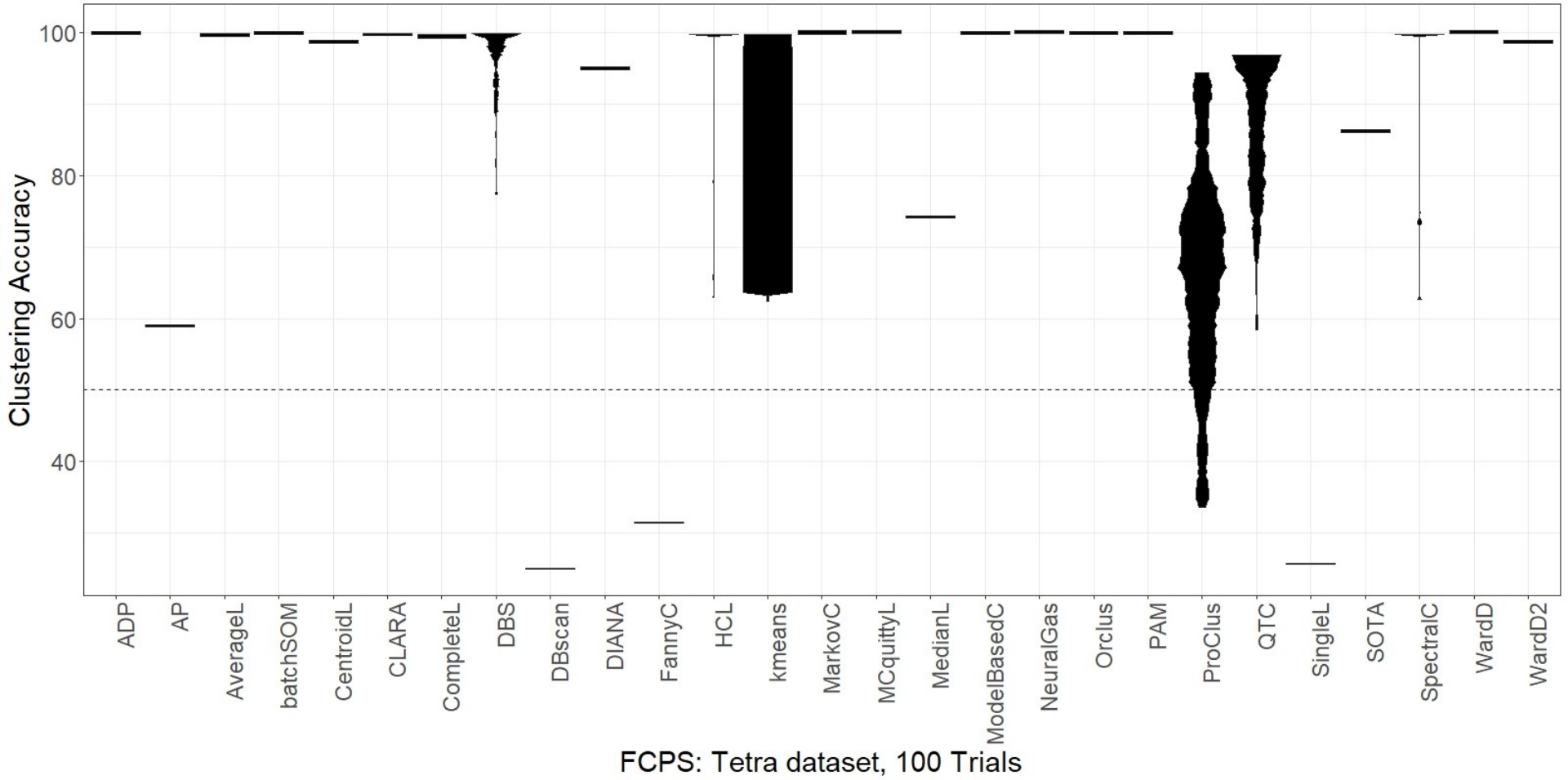
Distance-Based Structures

- Hepta: Cluster structures based only on spatial relationships between data points leading to 7 spherical Clusters
  - Spectral clustering, HCL and k-means have various modes
    - Probability states with very varying results



Linear Separable Structures

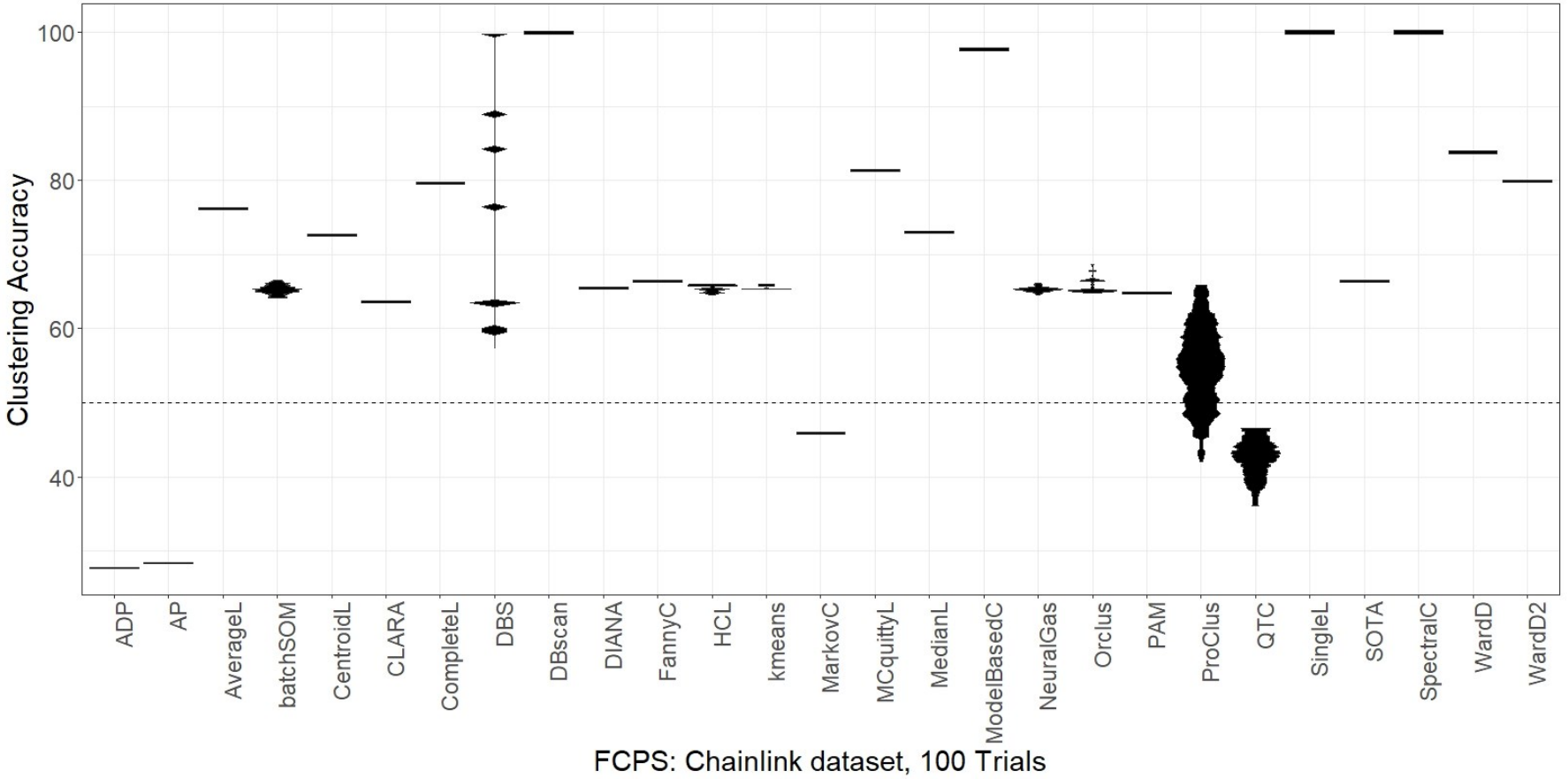
- Clusters defined by structures which can be separated by lines
  - Dataset Tetra has 4 almost touching cluster of equal variance



Linear Non-Seperable structures

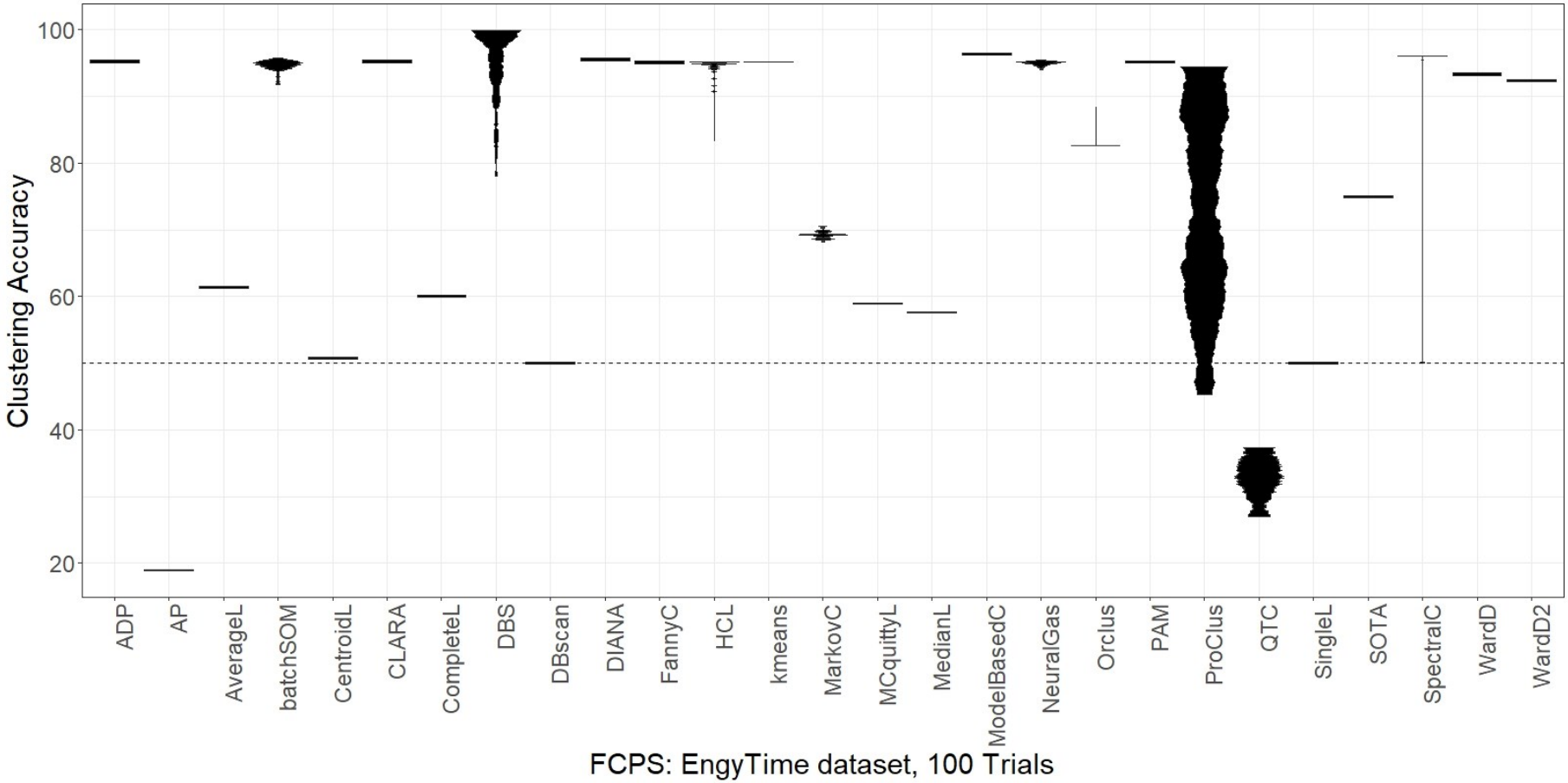
- Clusters defined by structures which cannot be separate by a line
  - Here Chainlink – two intertwined rings
- Most algorithms are unable to untangle such structures, e.g. model based clustering
  - Boxplot showed only outliers in DBS but PDE-optimized violin plot shows 6 modes

=> In praxis DBS is never used automatically, it uses the topographic map of generalized Umatrix to verify result



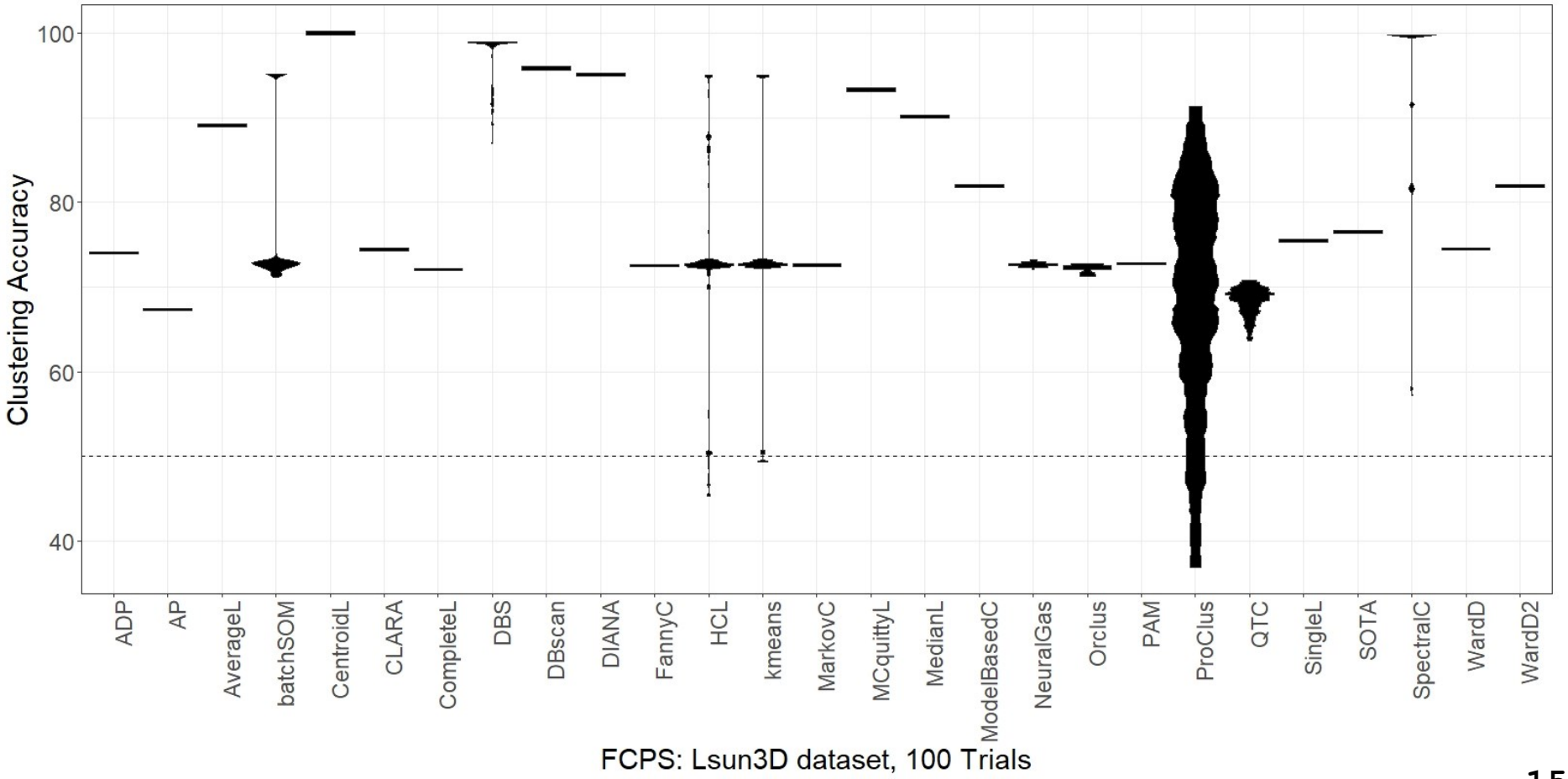
Density-Based Structures

- EngyTime dataset is defined by two 2D Gaussians with varying variance
- Only DBS captures structures completely



Combinations of Different Types of Structures

- For example, Outliers+Distance based structures (Lsun 3D)
  - Most algorithms are unable to catch different types of structures in one dataset



Discussion

- 27 clustering algorithms compared with default parameter settings
  - For density estimation the PDE was used leading to PDE-optimized violin plots
    - Can outperform violin plots and boxplots
- No Clustering algorithm is always able to reproduce all type of structures
- But some will more probable reproduce structures
  - e.g. Databionic Swarm (DBS), (1)
- Often algorithms produce results depending on the trial
  - Depends on the dataset

⇒ Do not compare only one trial per algorithm

(1) [Thrun, 2018]

Conclusion

- Use artificial datasets to compare clustering results with clearly predefined cluster structures
- PDE-optimized violin plot with an unbiased supervised index are a good approach to evaluate algorithms
  - Available in the R-package DataVisualizations on CRAN

- Are natural high-dimensional dataset useful to serve for benchmarking algorithms?
  - In our opinion: only if structures are known beforehand and prior classification is unambiguous

Thank you for listening.

Any questions?

Feel free to contact me through www.deepbionics.org

References

[Duda et al., 2001] Duda, R. O., Hart, P. E., & Stork, D. G.: *Pattern Classification*, (Second Edition ed.), Ney York, USA, John Wiley & Sons, ISBN: 0-471-05669-3, 2001.

[Everitt et al., 2001] Everitt, B. S., Landau, S., & Leese, M.: *Cluster analysis*, (McAllister, L. Ed. Fourth Edition ed.), London, Arnold, ISBN: 978-0-340-76119-9, 2001.

[Handl et al., 2005] Handl, J., Knowles, J., & Kell, D. B.: Computational cluster validation in post-genomic data analysis, *Bioinformatics, Vol. 21*(15), pp. 3201-3212. 2005.

[Hennig, 2015] Hennig, C., et al. (Hg.): *Handbook of cluster analysis*, New York, USA, Chapman&Hall/CRC Press, ISBN: 9781466551893, 2015.

[Hintze/Nelson, 1998] Hintze, J. L., & Nelson, R. D.: Violin plots: a box plot-density trace synergism, *The American Statistician, Vol. 52*(2), pp. 181-184. 1998.

[Hyndman/Fan, 1996] Hyndman, R. J., & Fan, Y.: Sample quantiles in statistical packages, *The American Statistician, Vol. 50*(4), pp. 361-365. 1996.

[Jain/Dubes, 1988] Jain, A. K., & Dubes, R. C.: *Algorithms for Clustering Data*, (Vol. 3), Englewood Cliffs, New Jersey, USA, Prentice Hall College Div, ISBN: 9780130222787, 1988.

[Theodoridis/Koutroumbas, 2009] Theodoridis, S., & Koutroumbas, K.: *Pattern Recognition*, (Fourth Edition ed.), Canada, Elsevier, ISBN: 978-1-59749-272-0, 2009.

[Thrun, 2018] Thrun, M. C.: *Projection Based Clustering through Self-Organization and Swarm Intelligence*, (Ultsch, A. & Hüllermeier, E. Eds., 10.1007/978-3-658-20540-9), Doctoral dissertation, Heidelberg, Springer, ISBN: 978-3658205393, 2018.

[Thrun/Ultsch, 2015] Thrun, M. C., & Ultsch, A.: Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis, DOI 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.

[Thrun/Ultsch, 2018] Thrun, M. C., & Ultsch, A.: Effects of the payout system of income taxes to municipalities in Germany, in Papież, M. & Śmiech, S. (eds.), Proc. 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, pp. 533-542, Cracow: Foundation of the Cracow University of Economics, Cracow, Poland, 2018.

[Tukey, 1977] Tukey, J. W.: *Exploratory data analysis*, (1 edition ed.), United States Pearson Education, ISBN: 978-0201076165, 1977.

[Ultsch, 2005a] Ultsch, A.: Clustering wih SOM: U* C, Proc. Proceedings of the 5th Workshop on Self-Organizing Maps, Vol. 2, pp. 75-82, 2005a.

[Ultsch, 2005b] Ultsch, A.: Pareto density estimation: A density estimation for knowledge discovery, In Baier, D. & Werrnecke, K. D. (Eds.), *Innovations in classification, data science, and information systems*, (Vol. 27, pp. 91-100), Berlin, Germany, Springer, 2005b.

[Ultsch/Behnisch, 2017] Ultsch, A., & Behnisch, M.: Effects of the payout system of income taxes to municipalities in Germany, *Applied Geography, Vol. 81*, pp. 21-31. 2017.

[Ultsch et al., 2016] Ultsch, A., Behnisch, M., & Lötsch, J.: ESOM Visualizations for Quality Assessment in Clustering, In Merényi, E., Mendenhall, J. M. & O'Driscoll, P. (Eds.), *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, (10.1007/978-3-319-28518-4_3pp. 39-48), Cham, Springer International Publishing, 2016.

[Wiwie et al., 2015] Wiwie, C., Baumbach, J., & Röttger, R.: Comparing the performance of biomedical clustering methods, *Nature methods, Vol. 12*(11), pp. 1033. 2015.