data sets the optimal number of clusters could be reliably determined, but for a high percentage of missing values in data sets the clustering structure of data could be determined only for data set with three clusters.

Comparing cluster validity functions with each other a crucial distinction emerges for data with missing values NMAR and MAR. On average, the optimal number of clusters determined by NPC and S is smaller than the actual number of clusters. In contrast, FHV and PD determined a greater number of clusters than the optimum. The reason for this is concerned with the monotonicity properties of these cluster validity functions for increasing number of clusters.

5. Conclusions and Future Work

In this paper, we analysed different cluster validity functions for fuzzy clustering in terms of determining the optimal number of clusters on incomplete data. We proposed some adaptions for cluster validity performance measures involving data matrix for the calculation. Furthermore, in experiments on several data sets we analysed to what extent the clustering results produced by clustering methods for incomplete data reflect the distribution structure of original data. The experimental results have shown that the best performance in terms of cluster tendency assessment was achieved by clustering algorithms OCSFCM and NPSFCM, which estimate missing values by values close to cluster prototypes. In this way, they preserve and strengthen clustering structure of data. In contrast, the clustering results obtained by the partial distance strategy FCM were instable regarding the determining the optimal number of clusters. However, the basic OCSFCM and NPSFCM leave clustering structure (e.g. cluster sizes) out of consideration while estimating missing values. Due to this these methods are instable in assigning data object to clusters on data with differently sized clusters [5]. Therefore, in our future research, we plan to continue working on the improvement of clustering algorithms for incomplete data using cluster dispersion.

Comparing cluster validity functions regarding finding the optimal number of clusters, NPC obtained slightly better results than FHV and PD. However, summarising the results of our study, we do not conclude that NPC is the better cluster validity index for determining the optimal number of clusters. It might produce better results on data with simple distribution structure compared to other indices but, as already mentioned in [14], NPC ignores the geometric structure of clustering. Cluster validity indices like FHV and PD overcome this problem involving the data matrix for calculation. Since not all clustering methods for incomplete data complete the data matrix and the adaptions pursuing the available case approach for FHV and PD do not provide satisfying results, in future we also plan to develop a better adaption of cluster validity functions using volume and density of clusters to incomplete data.

References

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, 1981.

[2] R. J. Hathaway and J. C. Bezdek. Fuzzy c-means Clustering of Incomplete Data, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 31, no. 5, pp. 735–744, 2001.

[3] M. Sarkar and T.-Y. Leong. Fuzzy k-means Clustering with Missing Values. In *Proceedings of American Medical Informatics Association Annual Symposium (AMIA)*, pp. 588–592, 2001.

[4] H. Timm, C. Döring, and R. Kruse. Different Approaches to Fuzzy Clustering of Incomplete Datasets, *International Journal of Approximate Reasoning*, vol. 35, pp. 239–249, 2004.

[5] L. Himmelspach and S. Conrad. Fuzzy Clustering of Incomplete Data Based on Cluster Dispersion. In Proceedings of the 13^{th} *International Conference on Information Processing and Management of Uncertainty* (IPMU 2010), Lecture Notes in Computer Science 6178, pp. 59-68, Springer-Verlag, 2010.

[6] L. Himmelspach and S. Conrad. Clustering Approaches for Data with Missing Values: Comparison and Evaluation. In *Proceedings of the Fifth IEEE International Conference on Digital Information Management* (ICDIM 2010), 2010.

[7] J. K. Dixon. Pattern Recognition with Partly Missing Data, *IEEE Transactions on System, Man and Cybernetics*, vol. 9, pp. 617–621, 1979.

[8] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[9] E. Backer and A. K. Jain. A Clustering Performance Measure based on Fuzzy Set Decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3 (1), pp. 66–74, 1981.

[10] X. L. Xie and G. Beni. A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 (8), pp. 841–847, 1991.

[11] I. Gath and A. B. Geva. Unsupervised Optimal Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, pp. 773–781, 1989.

[12] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*, John Wiley & Sons, 2002.

[13] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*, Wiley, 1999.

[14] J. C. Bezdek, W. Li, Y. Attikiouzel, and M. P. Windham. A Geometric Approach to Cluster Validity for Normal Mixtures, *Soft Computing*, vol. 1, no. 4, pp. 166–179, 1997.