The Concentration Centroid Minimum Distance Clustering Criterion

Matthias Zerbst; Lars Tschiersch

Department of Statistics, University of Dortmund, Germany

Abstract: Building homogenous classes is one of the main goals in clustering. Homogeneity can be measured by the intra-class variance (Bock, 1998). Especially in erosion projects but in other applications as well the separation between the built classes is as important as the homogeneity of the classes. Special clustering methods can be used to reach this aim, for instance the *Maximum Linkage Algorithm* (Zerbst, 2001) or the *Advanced Maximum Linkage Algorithm* (Tschiersch, 2002). To judge the separation quality of such clusterings, the shortest distances between all centroids is considered. Zerbst (2001) shows that the arithmetic mean over all distances isn't good enough for judging selectivity. Therefore the concentration centroid minimum distance criterion is proposed in this paper. This criterion is based on the ratio of weighted symmetric mean over the minimal distances and the *Gini* coefficient over the minimal distances. It also judges the class separation independent of the underlying data situation.

1 Introduction

Clustering methods are widely used in many applications, for instance in data preparation. While clustering information is gathered for further analysis. Foregone the building of homogenous classes is essential. The homogeneity is indicated by a small intra-class variance of the clustered data. In some area of application a high selectivity between the classes is essential, too. Examples of these applications are pixel based image clustering of erosion studies or gene expression data analysis, respectively. Beside the need of special algorithms for clustering under the aspect of high selectivity between the classes, such as the *Maximum Linkage Algorithm* (Zerbst, 2001) or *Advanced Maximum Linkage Algorithm* (Tschiersch, 2002), there is a demand for a criterion which judges the selectivity of a clustering. As abbreviation for *Maximum Linkage Algorithm* we use *MLA* and for *Advanced Maximum Linkage Algorithm* we use *AMLA*. An attempt in developing such a criterion is based on the minimal distance of each centroid to its nearest neighbors and can be read by Zerbst et al (2000). This criterion is called average minimal distance (*AvgMinDist*). The use of the minimal distance of each centroid to its nearest neighbor is very suggestive. Though, Zerbst (2001) shows that the simple arithmetic mean over these distances is only a good criterion under special conditions. Especially in the case of outliers in the data the criterion can come to a wrong decision. Even the use of a censored or winsorized mean can not guarantee an improvement. The determination of the grade of censoring is a not trivial problem in such a case. A solution of such a problem is given by the *Concentration Centroid Minimum Distance* criterion (*CCMD*), which will be introduced next. The presented examples demonstrate the functionality of the criterion. Two worst-case examples show its behavior in such cases. Further a possibility is presented to determine the number of needed clusters.

In the outlook chapter the results will be summarized. As well the determination of the number of cluster is considered.

2 Judging criterion for selectivity between classes

2.1 Clustering and known clustering criteria

A clustering separates a set of elements in as homogenous classes as possible. Let $\Omega \subseteq I\!R^m$ be the feature space and $O = \{x_1, ..., x_n\}, \quad x_i \in \Omega, i = 1, ..., n$ a set of vector values. We sought after a clustering *C(O)* of the values $x_1, ..., x_n$ from *O*. The number of classes q, in which the data will be separated, has to be established a priori. A clustering is given by

$$C(O) = \{c_1, ..., c_q\} \quad \text{with } c_i = \{x_{i1}, ..., x_{in_i}\}, \; i = 1, ..., q. \tag{2.1}$$

For a meaningful clustering we need two further assumptions:

1. To ensure that all classes are not empty, let $n_i > 0$, $i = 1, ..., q$.
2. To reach information reduction through a clustering, let $q \ll n$.

Moreover, we sought after additional class representatives for the q classes from (2.1). These class representatives $z_1, ..., z_q \in I\!R^m$ will be needed to describe the classes and for further analysis. The easiest way of building such centroids is:

$$z_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad .$$

To judge the clustering quality some criteria are necessary. Thereby the homogeneity of the classes has to be to the fore. This will be judged by the intra-class variance *g*(*C*(*O*)). According to Bock (1998) this can be formalized by:

$$g(C) := \frac{1}{n}\sum_{i=1}^{q}\sum_{k\in c_i}\left\| x_k - \bar{x}_{c_i}\right\|_2 \ , \tag{2.2}$$

Equation (2.2) shows that a better clustering implies a smaller intra-class variance. In many ecological and biological applications an additional requirement is needed. We ought to have a high selectivity between the classes. This feature should not be mixed up with a maximized inter-class variance. We are much more interested in the actual distance between the classes. Therefore the distances between the centroids or class representatives $z_1,..., z_q \in IR^m$ build the basis for judging the selectivity. The decisive distances are given by:

$$d_i = \min_{j\in\{1,...,q\}\setminus j}\left\| x_i - x_j\right\|_2 \ , i = 1,\ldots,q. \tag{2.3}$$

According to Zerbst (2001) the criterion of the average minimal distances with respect to equation (2.3) are derived by

$$AvgMinDist \quad (C(O)) = \frac{1}{q}\sum_{i=1}^{q} d_i \ .$$

The separation of the classes is better if the above criterion has larger values. The following section shows the drawbacks of this criterion with some examples.

2.2 Misjudging of *AvgMinDist* based on examples

The *AvgMinDist* has one essential disadvantage. It is founded on the arithmetic mean; hence it is sensitive to outliers. This can be lead to misinterpretation, if only one large minimal distance exists. This clarifies to the following example.

<u>Example 1</u>

Let two clusterings A and B be carried out. The ordered distances between each centroid to his next neighbor are given in Tab. 1. Notice, that $d_{(i)}$ is the ordered distance.

	$d_{(1)}$	$d_{(2)}$	$d_{(3)}$	$d_{(4)}$	$d_{(5)}$	$d_{(6)}$	$d_{(7)}$	$d_{(8)}$	$d_{(9)}$	$d_{(10)}$
Clusterung A	3.2	3.2	4.4	5.0	6.7	7.2	7.6	8.1	8.6	22.0
Clusterung B	4.7	4.7	5.1	7.0	7.0	8.0	8.0	9.0	9.5	9.5

Tab. 1: Distances of each centroid to his next neighbor of the two clusterings A and B.

To get further knowledge about the quality of separation of the classes, we'll have a look on the following diagram.

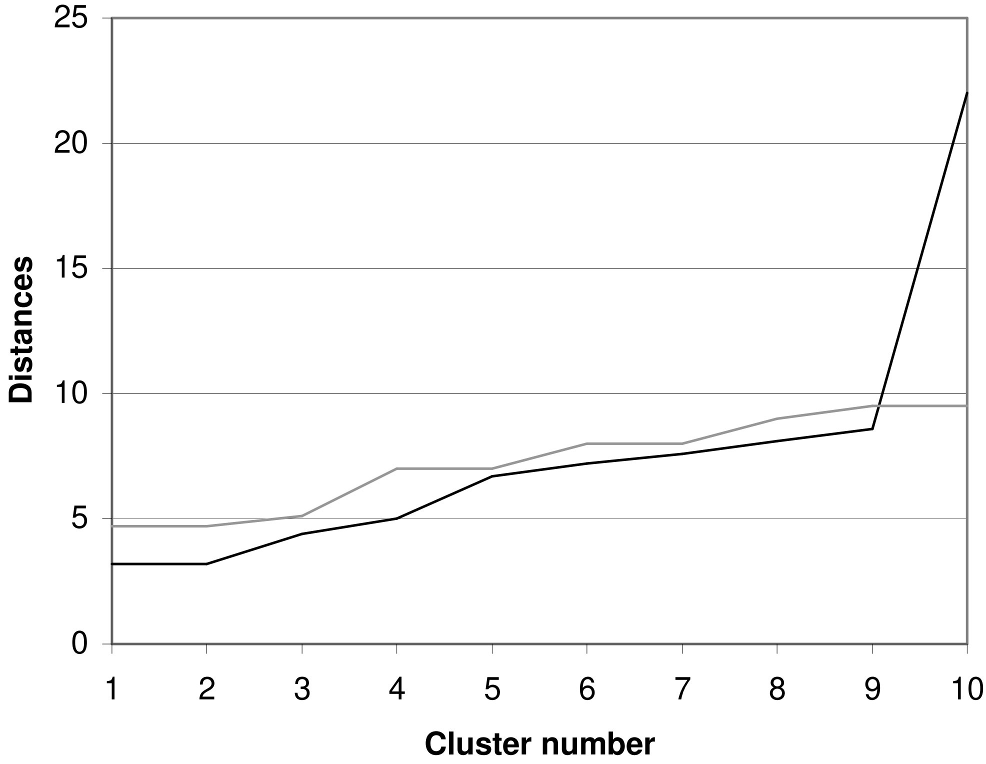

Fig. 1: Diagram of the minimal distances from Tab. 1. The black line represents clustering A and the gray line clustering B.

It can be seen from the above Tab. 1 that the clustering A (*AvgMinDist*(C_A) = 7.6) is better than Clustering B (*AvgMinDist*(C_B) = 7.25) due to the one really large distance (22.0). It can be seen at once from Fig. 1 that the judgment should be just the other way round, because the distances of B are larger for all values exceptionally the last one. The wrong judgment has its reason in building the mean over the distances with respect to the arithmetic mean.
Basically, we can see an additional problem. The judgment doesn't only depend on the mean value, but on the concentration as well. We will consider this in the next example.

<u>Example 2</u>

Let us consider two further clusterings A and B again. The ordered minimal distances between the centroids and their next neighbor are listed in Tab. 2.

	$d_{(1)}$	$d_{(2)}$	$d_{(3)}$	$d_{(4)}$	$d_{(5)}$	$d_{(6)}$	$d_{(7)}$	$d_{(8)}$	$d_{(9)}$	$d_{(10)}$
Clus. A	0.50	0.50	0.75	0.79	0.84	19.80	20.10	20.40	21.70	22.10
Clus. B	7.00	7.00	7.40	7.90	8.20	10.00	11.00	11.50	12.30	13.10

Tab. 2: Distances of each centroid to his next neighbor of the clusterings A and B of example 2.

When looking on Tab. 2, we see that the distances can be blocked in two groups for each clustering. In clustering A the first five values are very small, the last five essentially larger. Clustering B behaves similar. The distinction between the blocks is smaller.

The judgement of both clusterings leads to the following results. Clustering A, with an *AvgMinDist* of 10.75, is better than clustering B with an *AvgMinDist* of 9.54. This holds because five large values of clustering A dominate the criterion. Figure 2 shows the extreme differences within clustering A and the differences between clustering A and B as well. Therefore clustering B should be preferred, because of having no extreme small values (high selectivity) and no extreme shifts in distances (equality of distances).

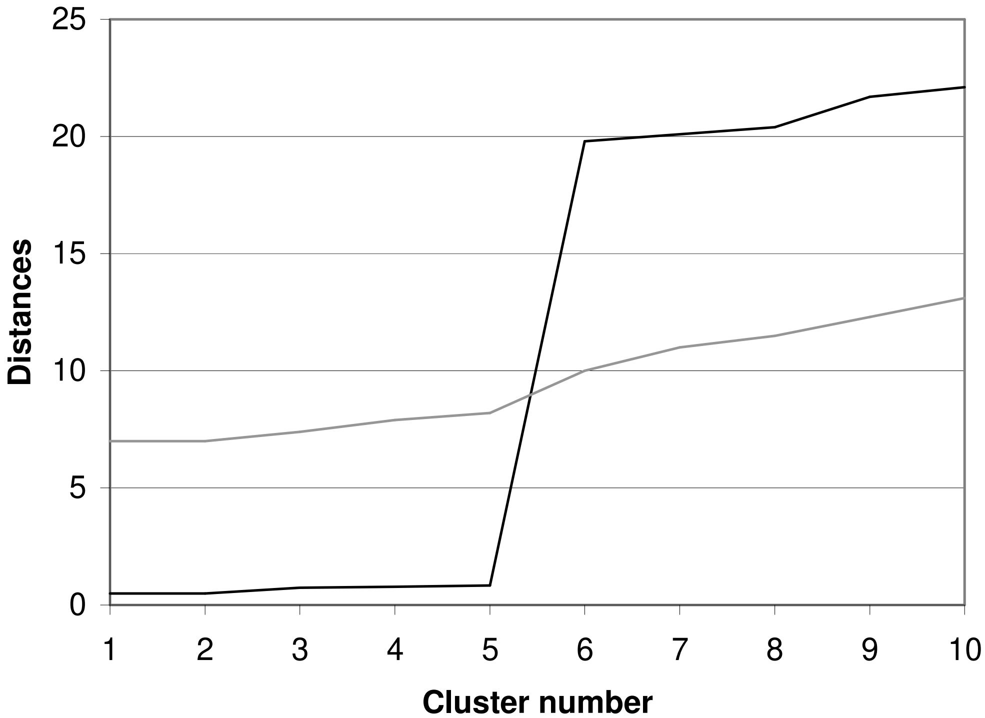

Fig. 2: Diagram of the minimal distances of example 2 from Tab. 2. The black line belongs to clustering A and the gray line represents clustering B.

In the following section we will present a new criterion which overcomes the disadvantages showed in example 1 and 2. It will also lead to a clustering with an even better selectivity.

2.3 The Concentration Centroid Minimum Distance criterion (*CCMD* criterion)

The *CCMD* criterion is founded on two ideas. Let $d_i,\ i = 1...,q$ be as in (2.3). The first idea is to use weighted symmetric mean over the ordered d_i instead of using the usual arithmetic mean. The weighting is performed by weighting values lying at the edge less than those lying in the middle of the ordered distances. Thereby the effect of outliers is weakened in all directions. The weighting is done according to the following equations.

$$\omega(i) = \begin{cases} \dfrac{\dfrac{1-a}{\frac{q+1}{2}-1}\cdot i + a - \dfrac{1-a}{\frac{q+1}{2}-1}}{D_{num}} & ,\quad i \le \dfrac{q}{2} \\[2ex] \dfrac{\dfrac{a-1}{q-\frac{q+1}{2}}\cdot i + 1 - \dfrac{q+1}{2}\dfrac{a-1}{q-\frac{q+1}{2}}}{D_{num}} & ,\quad i > \dfrac{q}{2} \end{cases} \tag{2.4}$$

with

$$D_{num} := \sum_{j\le q/2}\left(\frac{1-a}{\frac{q+1}{2}-1}\cdot j + a - \frac{1-a}{\frac{q+1}{2}-1}\right) + \sum_{j>q/2}\left(\frac{a-1}{q-\frac{q+1}{2}}\cdot j + 1 - \frac{q+1}{2}\frac{a-1}{q-\frac{q+1}{2}}\right),$$

where $0 \le a < 1$. In our case a will be chosen equal to zero. The structure of the weights defined by (2.4) is not as complex as it seems when looking on the equation. Fig. 3 contains two examples for (2.4) with $q = 10$ and $a = 0$ (black line) and $a = 0.5$ (grey line). Beside the fact, that the $\omega(i)$ are discrete values, the single points are connected to get a better impression of them.

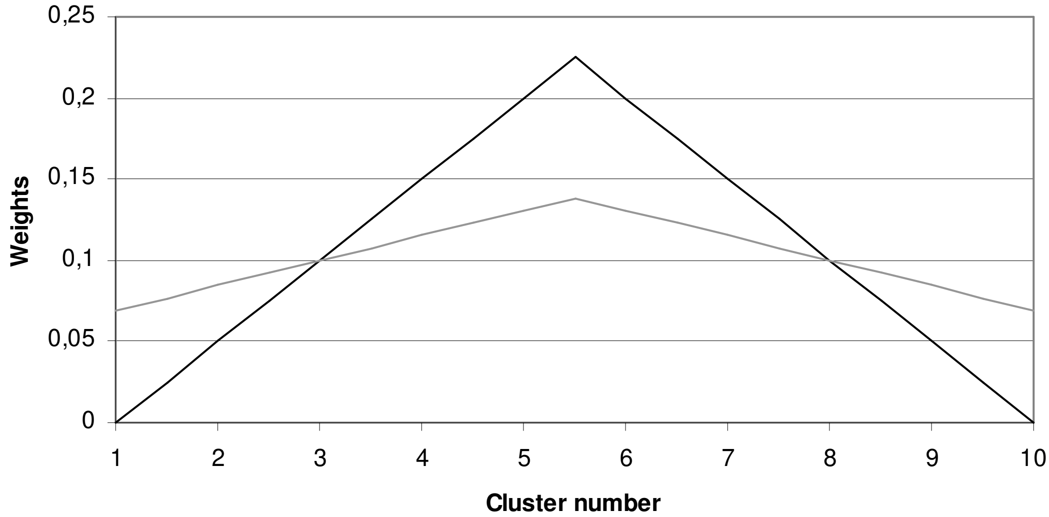

Fig. 3: Diagram of function (2.4) for $q = 10$ distances and a = 0 (black line) and $a = 0.5$ (grey line).

Based on this weighting we define the *Weighted Average Minimal Distance* (*WAMD*) by

$$WAMD_{d_i} = \sum_{i=1}^{q} \omega(i) \cdot d_{(i)} \,. \tag{2.5}$$

The interpretation follows the principle: A larger value for $WAMD_{C(O)}$ implies a better clustering. The negative effects of extreme unequal distances, as in example 2, will not be considered, though. The solution of this problem has been partly mentioned above. We have to take into account the growing of the ordered distances, moreover the concentration of the distances.
The consideration of the distance concentration leads to the Lorenz curve. We also can consider the strongly related *Gini* coefficient *G*. As known from the literature a small *Gini* coefficient implies less concentration, hence with respect to the minimal distances a better clustering.

Therefore we use the $WAMD_{C(O)}$ over the distances relative to the *Gini* coefficient to judge a clustering. So $CCMD_{C(O)}$ has the form

$$CCMD_{C(O)} = \frac{WAMD_{C(O)}}{\left(G_{C(O)}\right)^{1/2}} \,,$$

where $G_{C(O)}$ is the *Gini* coefficient. This coefficient is clearly the double area between the Lorenz curve and the 45° line. In our case the Lorenz curve has $q + 1$

points of support. When joining the points of support with a straight line we get the Lorenz curve.

In our case the points of support for the abscissa (k_i) and ordinate (l_i), $i = 0,\ldots,q$ are given by:

$$k_i = \frac{i}{q} \quad \text{and} \quad l_i = \frac{\sum_{j=1}^{i} d_{[j]}}{\sum_{j=1}^{q} d_j}, \text{ with } i = 0,\ldots,q, \quad \text{respectively.}$$

The *Gini* coefficient has the form

$$G_{C(O)} = \sum_{i=1}^{q} \left(k_{i-1} + k_i\right) \frac{d_{(i)}}{\sum_{j=1}^{q} d_i} - 1 .$$

To calculate the *CCMD* criterion the following equation is used:

$$CCMD_{C(O)} = \frac{\sum_{i=2}^{q} \omega(i) \cdot d_{(i)}}{\left(\sum_{i=1}^{q} \left(\frac{i-1}{q} + \frac{i}{q} \right) \frac{d_{(i)}}{\sum_{j=1}^{q} d_i} - 1 \right)^{1/2}} .$$

In contrast to the *AvgMinDist* is this parameter able to identify "better" clusters.

Example	**Clustering**	***AvgMinDist***	**WAMD**	***Gini***	***CCMD***
1	A	7.60 *	6,51	0.154	16,60
	B	7.25	7,37 *	0.069 *	28,02 *
2	A	10.75 *	10,49 *	0.241	21,36
	B	9.54	9,33	0.065 *	36,66 *

Tab. 3: Parameters for judging the clusterings of example 1 and 2. The grey lines represents the better clusterings with respect to the judgement based on the optical impression of the diagrams of minimal distances. The "*" mark the values which belong to the better clustering for the corresponding criterion.

The parameter of judging the selectivity of two clustering of example 1 and 2 are given in

Tab. 3. The grey rows show the "best" clustering for each example. The parameter marked with * represent the best parameter within the example. The *Gini* coefficient identifies the best clustering in each case. After a further look onto

Tab. 3 the question rises whether the judging according to the *Gini* coefficient is sufficient. As one can see from the following example this holds not in every case.

Example 3:

Consider the clustering A and B. The minimal distances between the centroids to its nearest neighbors are listed in Tab. 4.

	$d_{(1)}$	$d_{(2)}$	$d_{(3)}$	$d_{(4)}$	$d_{(5)}$	$d_{(6)}$	$d_{(7)}$	$d_{(8)}$	$d_{(9)}$	$d_{(10)}$
Clustering A	5,8	5,8	5,9	6,1	7.2	7.6	8.1	8.4	8.8	9.2
Clustering B	5,3	5,3	6.7	8.4	10.5	11,8	12,5	13,1	14,5	16,1

Tab. 4: Example 3: Distances of each centroid to his next neighbor of two clusterings A and B.

Clustering A starts with a larger minimal distance as B does. Considering only this we would prefer clustering A. Even the rise of the minimal distances is smaller than in B. The uniformly concentration of the distances considered alone is an argument for A as well. When considering the minimal distances of B, which are lower than the distances of A in the beginning and larger in the end one will decide to choose clustering B. The diagram underlines this fact.

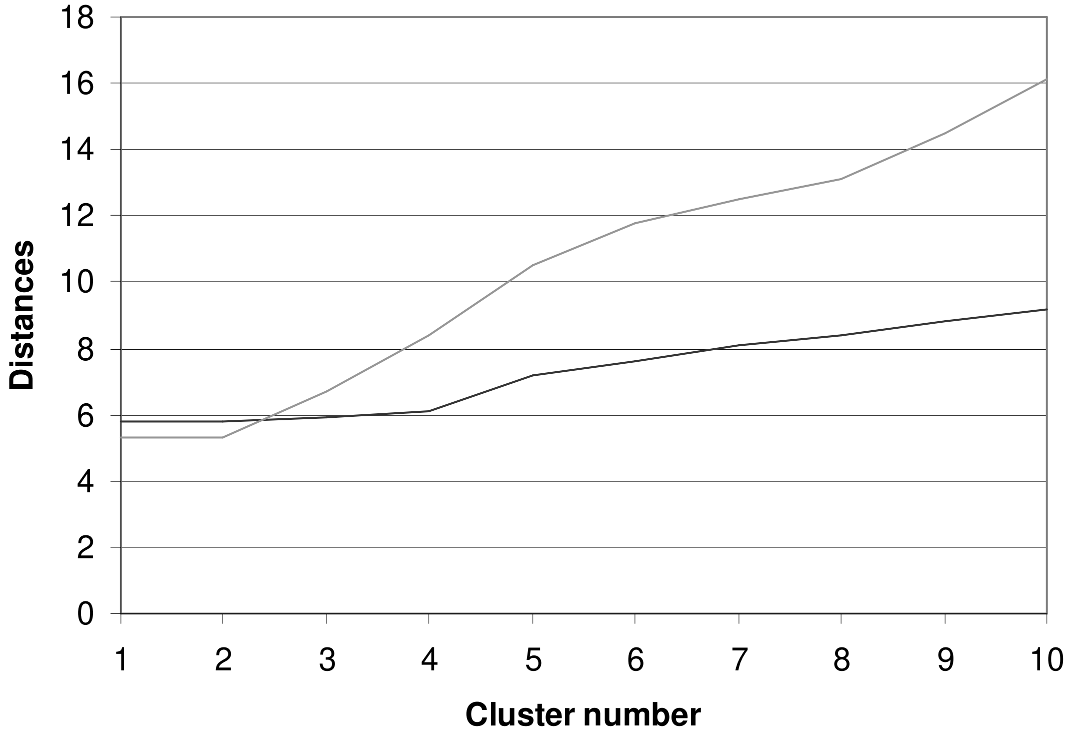

Fig. 4: Diagram of the minimal distances of example 3 from Tab. 4. The black line belongs to clustering A and the gray line represents clustering B.

The results are summarized in Tab. 5.

Clustering	*AvgMinDist*	WAMD	*Gini*	*CCMD*
A	7,29	7,25	0,048 *	32,97
B	10,42 *	10,57 *	0,099	33,51 *

Tab. 5: Parameters for judging the clusterings of example 3.

The example shows: The *Gini* coefficient does not detect the "best" clustering in this case, but the *CCMD* criterion does detect the "best" clustering.

Due to the presentment of the *CCMD* criterion for judging the selectivity of a clustering this criterion is superior to the *AvgMinDist*. The criterion yields the same results as an expert when looking at the visualization of the graphical tools. The *CCMD* criterion is suitable for judging the selectivity of a clustering, regardless which method for clustering is used.
The next chapter shows how to interpret the graphical output of the minimal distances for the *MLA* and *AMLA* methods with respect to find the optimal number of clusters for clustering.

3 A graphical method for determination of the number of clusters

When considering the selectivity of a clustering we can use a graphical tool for *MLA* and *AMLA* for finding the right amount of classes. Analog to the considered graphics of the minimal distances we will create some graphic tools as well. For this new graphic we do not need a clustering. The method can also be used for the calculated centroids of *AMLA*. In this graphic the distances of each centroid to its nearest neighbor will be mapped. But the values will not be sorted in ascending order than in descending order. A typical graphic is shown in Fig. 4.

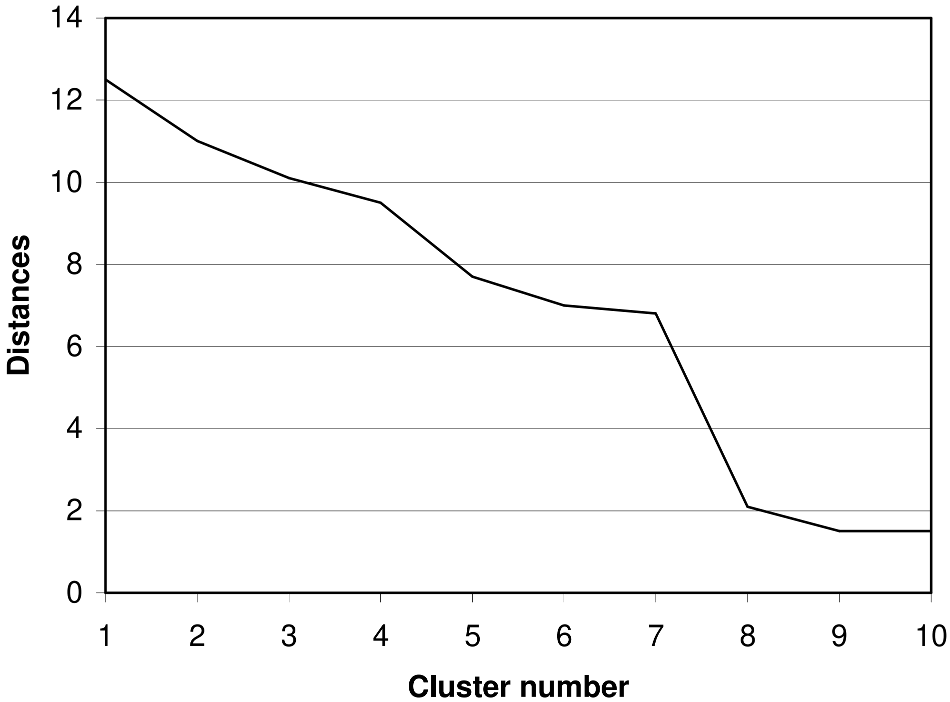

Fig. 5: Diagram of minimal distances in descending order

Interpretation of Fig 5.: Is the curve to low or does the curve show an extreme downward buckle, as between centroid seven and eight, we have a hint that the number of clusters is to large for a clustering with high selectivity. Only a clustering with fewer clusters can reach a higher selectivity. The reason for this is that the selectivity is chosen maximal with this method.

At this point we can use an advantage of the *MLA* / *AMLA* method. A clustering with i clusters has the same 1 to i centroids as a clustering with $i + 1$ clusters. The $i + 1$ cluster is only an additional centroid. If we would test which number of classes for a clustering is necessary the centroids which are calculated by *MLA / AMLA* can be used. Suppose we have clustering with a maximum of *j* clusters. Therefore, *j* centroids have been calculated by *MLA / AMLA*. When the selectivity of the clustering with *j* clusters isn't large enough the following process is started: Remove the last calculated centroid and draw the diagram of the minimal distances. By leaving out the last centroid two things are archived. The centroid with the smallest distance value is skipped due to the fact that *MLA / AMLA* chooses the centroid with the maximal minimal distance in each step. The value is the last in the diagram. When calculating new minimal distances with a reduced number of centroids by one, the diagram of minimal distances rises. At least the value of the now last and smallest distances rises to the value of the former third last value. In most cases even the values of further minimal distances become larger. This effect is shown in Fig. 6. However, the diversification depends on the site of the data.

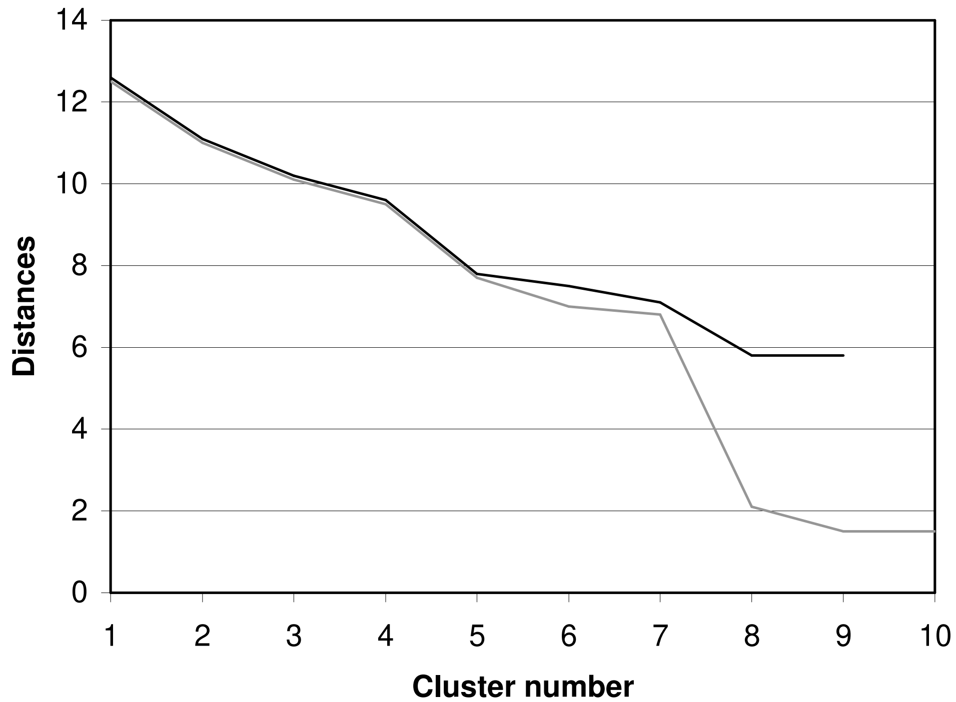

Fig. 6: Diagram of minimal distances in descending order. The gray line belongs to the original clustering. The black line represents the minimal distances after skipping one centroid.

Omitting the last centroid is carried out as long as the smallest value is conform with the users requirement. Subsequent the intra-class variance has to be checked. The variance should not be "to large". To large depends on the requirement of the user and the demand of the problem. If the intra-class variance is too large, we have to re-implant the omitted centroids step by step until the combination of small intra-class variance and large minimal distances reach the users requirements. The stepwise adding of centroids is much more time consuming than omitting a centroid. This holds because there is a need for a new classification to calculate the intra-class variance.

The graphical selection procedure is not restricted to *MLA* / *AMLA*. Furthermore it can be used for other clustering methods as well. But the expenditure of omitting and re-implanting centroids will be greater with other methods.

Outlook

The introduced *CCMD* criterion is excellent applicative in finding the clustering with the intuitive "best" selectivity. The identification is achieved independently of two problematic effects. First a low regime of the minimal distances of the centroids to its nearest neighbor and second a large contrast of the minimal distances. This connotes that the curves have a larger difference in altitude. Therefore, this criterion is more sensitive than the *AvgMinDist*. Moreover, in extreme cases like outliers, we wouldn't come to the wrong decision.

Nevertheless, there are still two criteria needed for judging a clustering: The *CCMD* for the selectivity and the intra-class variance for the homogeneity of the classes. An optimal innovation would be the conflation of these two criteria. The problem which arises is that the criteria measure different quality properties. The conflation would involve a number of assumptions and pre-considerations.

The graphical interpretation for the determination of the number of classes is a byproduct of *CCMD*. This approach is insightful, but the use of the method will probably be too much time consuming. Furthermore, when calculating a clustering we have to take the underlying problem into account as well. Therefore, the choice of the number of classes is not arbitrary. It is conceivable to convert the graphic tool into a parameter like the conversion of the minimal distance diagram to the *CCMD* criterion. This supplement of the algorithm could lead to a parallel computation of the number of clusters and of the clustering itself. But even for this a couple of pre-considerations are necessary.

Acknowledgement

This research is supported by the German Research Council (Deutsche Forschungsgemeinschaft – DFG) through the Collaborative Research Centre 'Reduction of Complexity for Multivariate Data Structures' (SFB 475) at the University of Dortmund.

5 References

Anderberg, M. R. (1973): *Cluster Analysis for Applications*. Academic Press, New York.

Bock, H.-H. (1998): *Clustering and neural networks.* In: Rizzi, A.; Vichi, M.; Bock, H.-H. (Eds): *Advances in data science and classification*. Springer, Heidelberg, 265-278.

Cressie, N.A. (1993): *Statistics for spatial data.* Revised Edition Wiley, New York, New York.

Hartigan, J.A. (1975): *Clustering algorithms.* Wiley, New York, New York.

Johnson, R.A.; Wichern, D.W. (1992): *Applied multivariate statistical analysis.* 3rd-Edition, Prentice Hall, Engelwood Cliffs, California.

Kohonen, T. (2001): *Self-organizing maps, 3rd Edition.* Springer, Heidelberg.

Myers, W.; Patil, G.P.; Taillie, C. (1997): *PHASE formulation of synoptic multivariate landscape data.* Technical report Number 97-1102. Center for statistical ecology and enviromental statistics, Department of Statistics, PennState University, Pennsylvania.

Tou, J.T.; Gonzalez, R.C. (1974): *Pattern recognition principles.* Addison-Wesley, Reading, Mass.

Tschiersch, L.; Zerbst, M. (2002): *The Advanced Maximum Linkage Clustering Algorithm*. Technical Report 10, Department of Statistics, University of Dortmund.

Zerbst, M.; Tschiersch, L.; Guimarães, G.; Talbi, M.; Urfer, U. (2001): *On clustering of aerial photographs and high resolution satellite images.* Handed to ENVIRONMETRICS, Canada.

Zerbst, M. (2001): *Die pixelbasierte Clusterung von Luftaufnahmen im Rahmen* von *Erosionsuntersuchungen.* Dissertation, Universität Dortmund, Dortmund.