Theor. Appl. Climatol. 64, 143–150 (1999)

Theoretical and Applied Climatology

[1]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany
[2]Meteorological Institute, University of Hamburg, Germany

Applying Non-Hierarchical Cluster Analysis Algorithms to Climate Classification: Some Problems and their Solution

F.-W. Gerstengarbe[1], **P. C.Werner**[1], and **K. Fraedrich**[2]

With 4 Figures

Received September 3, 1998
Revised July 5, 1999

Summary

Extended non-hierarchical cluster analysis is improved by deriving the initial cluster number and estimating the outliers in the final cluster set. These improvements are tested and compared with an established cluster algorithm using a toy example. Applying the improved cluster analysis to a classification of the European climates shows that the proposed techniques can be of great practical relevance.

1. Introduction

The aim of the cluster analysis is the separation of several elements into homogeneous groups. Two main techniques are possible: Using hierarchical methods (see, for example, Bacher, 1996), different sequences of groups on different levels may be constructed. The result is a hierarchy of clusters in a ''tree structure''. This method is commonly used in most of the existing statistical software tools (i.e., SAS, 1990; StatSoft, 1994; SPSS, 1999). The disadvantage of this technique lies in the fact that an exchange of elements between the groups is impossible when the ''tree structure'' is building up. With non-hierarchical methods, this disadvantage vanishes because the elements are simultaneously partitioned into a given number of clusters (see, for example, Steinhausen and Langer, 1977; StatSoft, 1994). Jahnke (1988) showed for independent samples that the minimum-distance methods fulfill the consistence criterion of statistical estimation procedures which, therefore, appear to be an ideal tool for climate analysis to objectively classify regions of similar climate.

In the following analysis we apply the minimum-distance method (Forgy, 1965) of the non-hierarchical cluster analysis. The starting condition is to attribute an equal number L of elements e_i from a total of M to the initial number of K_0 clusters (initial partition) so that each cluster receives $L = M/K_0$ elements as follows:

$$\begin{array}{ll} e_1, \ldots, & e_L \in c_1 \\ e_{L+1}, \ldots, & e_{2L} \in c_2 \\ \vdots & \vdots \\ e_{(k-1)L+1}, \ldots, & e_{kL} \in c_k \end{array} \qquad (1)$$

where c_i, $i = 1, \ldots, k$ represents the cluster.

A so-called group centroid $\bar{e}_k$ is then calculated for each k of the K_0 clusters:

$$\bar{e}_k = \frac{1}{L} \sum_{i=(k-1)L+1}^{kL} e_i \qquad (2)$$

The Euclidean distance between the elements and the group centroid $\bar{e}_k$ defines the following target function $a(g)$ at each grouping step g:

$$a(g) = \sum_{k=1}^{K} \sum_{i \in k} \left| e_i - \bar{e}_k \right|^2 \tag{3}$$

In this sense each grouping step can be seen as displacement of the element e_i into that cluster of the nearest centroid. Thus the target function can be minimized:

$$a(g) \; \forall \; g \rightarrow \; \min \tag{4}$$

This procedure is repeated until a local minimum of the target function is reached. Note that in the remainder of this paper the procedure described above is referred to as ''standard method''. The initial and final number of clusters are the same and subjectively defined when applying the ''standard'' non-hierarchical cluster analysis algorithm. If, for example, the initial number of clusters is too small, the number of elements within a single cluster is relatively large. Consequently, possible internal structures of an initially identified cluster cannot be considered any further. Problems associated with this procedure are the subjectively defined number of clusters and the unknown statistical significance of the cluster separation. Solutions of the problems have been suggested by Gerstengarbe and Werner (1997), but two additional difficulties arise which require attention; that is (a) the choice of an optimal initial number of clusters from which the iteration commences, and (b) the appropriate cluster separation.

Section 2 presents the theoretical basis for improving on these two points. Section 3 shows a climate classification for Europe utilising the improved cluster analysis.

2. Theoretical Basis

Gerstengarbe and Werner (1997) have developed a procedure to test the quality of cluster separation as follows: After having reached the local minimum each cluster is equipped with a varying number of elements. Each element is defined by n parameters, that is, it is located in an n-dimensional parameter space. Each cluster consists of a certain number of elements representing a scatter plot of elements in the parameter space. If the clustering leads to a minimum of the target function (Eq. 4), overlaps may occur between the scatter plots of individual clusters. This means that the parameter space of a cluster a passes into that of cluster b and vice versa and the number of parameters in the common space of the two clusters can be defined as overlaps of cluster a with respect to cluster b. The maximum possible number of overlaps between two clusters a and b is $O = NL_aL_b$ (N-number of parameters, L_a-number of elements in cluster a, L_b-number of elements in cluster b). This number is reached if both clusters cover the same region within the n-dimensional parameter space. The following χ^2-test can be derived introducing the maximum possible number of overlaps $O_{a,b}^{\max}$, the actual number of overlaps $O_{a,b}$, and the mean over all actual numbers of overlaps $\bar{O}$ or all combinations of cluster pairs:

$$\chi^2 = \frac{(O_{a,b} - \bar{O})^2 \cdot (2O_{a,b}^{\max} - 1)}{(O_{a,b} + \bar{O}) \cdot (2O_{a,b}^{\max} - O_{a,b} - \bar{O})} \tag{5}$$

with one degree of freedom. Using this method the statistical confidence of separation can be determined and the optimum number of clusters which gives the best separation between all clusters. This number is, in general, not identical with the given initial number of clusters. The following steps need to be performed to achieve the optimum separation:

- Apply the cluster algorithm up to the point that one cluster is fully separated from all others.
- Reduce the initial data set by the separated cluster elements.
- Repeat the algorithm until all clusters are seperated in a statistically reliable sense.

Practical applications require further improvements for (a) the choice of an optimal initial number of clusters starting the iteration and (b) the cluster separation.

2.1 Optimal Initial Number of Clusters

The initial number of clusters can influence the cluster result. Therefore, it is necessary to estimate an optimum initial number of clusters. The following procedure is suggested.

The starting point for the calculation of the initial cluster number is the target function $a(g)$

defined for the "standard method". We know that the target function is constructed in such a way that the partition for which the function reaches a minimum defines the most favourable grouping of the clusters. Now we calculate this target function for an increasing number of initial clusters (for $p = 2, 3, 4, \ldots, K_0$) so that we get a sequence of K_0 target function values. This sequence can be incorporated in the following estimation of an optimum initial number of clusters. Realising that each value of the target function is equivalent to a specific initial number of clusters, we define the optimal initial number as the inflection point within the sequence of target function values where the trend of the target function values disappears and no further significant changes occur. This idea can be solved practically by:

- calculation of the differences between consecutive values of the target function sequence and creation of a difference series d_i $(i = 1, \ldots, m)$ with $m = K_0 - 1$ values and
- applying the Pettitt-test (Pettitt, 1979) to estimate the beginning of a trend (or inflection point) within the difference series.

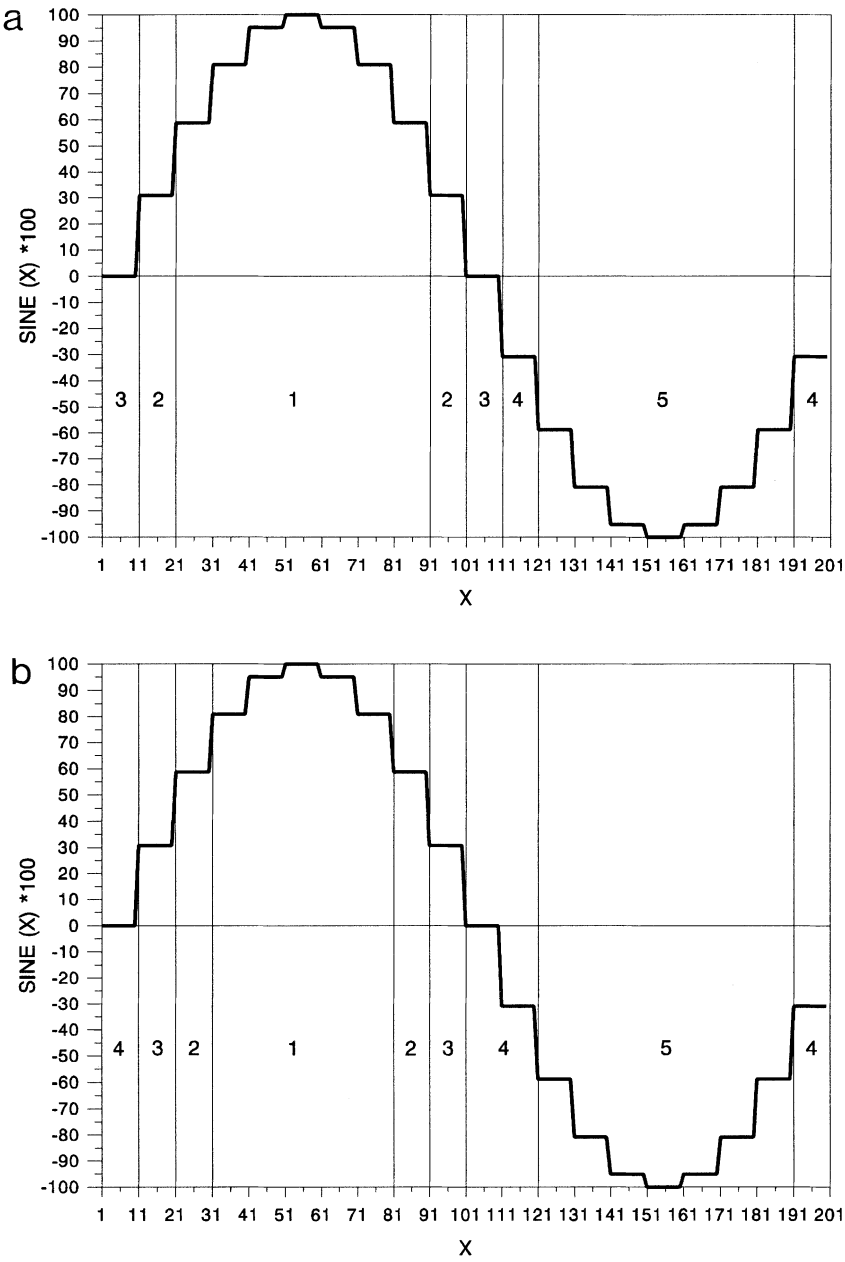

Fig. 1. Cluster analysis applied to a one parameter oscillation: (a) an optimal initial number of clusters, $k_0 = 5$, and (b) a defined initial number of clusters, $k_0 = 5$

The Pettitt-test can be derived from the U-test (Mann-Whitney, 1947), which is based on the rank values of the sequence. The inflection point is defined as that point for which the absolute value of the sequence of differences, d_i, has reached a maximum with

$$X_p = 2 \cdot R_p - p \cdot (m + 1) \tag{6}$$

where p is the position within the difference series d_i, m is the number of values of the difference series, and R_p is the sum of the ranks of the difference series d_i of the target function values. Continuously increasing the initial number of clusters, the Pettitt-test finally defines that position within the difference series d_i (of the target function values) which divides this series into one part with significant changes values and the other one without changes.

2.2 Cluster Separation

The proposed cluster separation algorithm leads to a number K of separated clusters, the significance test of which is connected with a defined error probability (level of significance $\alpha = 0.01$ or 0.05). Note that a statistically significant separation of two clusters allows a small number of overlaps. Thus, some clusters may contain ''strange'' elements are identified as outliers. In the statistical sense of significance, this case is without any consequence. But in some cases, such outliers can have a negative influence on the clustering (which will be shown in some detail in Section 3), and if such outliers exist, they may be better assigned to another cluster.

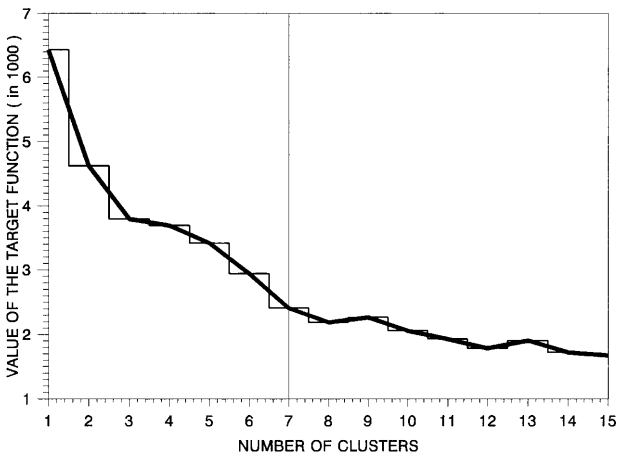

Fig. 2. Result of the Pettitt-test for the estimation of the initial number of clusters (climate classification)

An outlier test provides the solution for this problem because it identifies a value deviating significantly from the basic sample (Ferguson, 1961). For each element of a cluster, we calculate its Euclidean distance to the group centroid which, for each cluster, leads to a sample of Euclidean distances $x_i (i = 1, \ldots, k; k$-number of elements within the investigated cluster). Using the Thompson-rule (Müller et al., 1973) we can identify the outliers of the sample with the following test value:

$$t_i = \frac{x_i - \bar{x}}{s^*} \tag{7}$$

where $\bar{x}$ and s^* are the arithmetic sample mean and the standard deviation. Outliers are all values $x_i (i = 1, \ldots, k)$ for which $|t_i| > z_{f;\alpha}$ is valid, with $f = k - 2$ ($f =$ degree of freedom; $z_{f;\alpha} =$ critical value; s. statistical table). In this sense the Thompson rule is a two-sided test to examine the hypothesis H_0:''The sample has no outliers for a chosen level of significance α''. If outliers exist a better assignment can be reached by arranging the outliers into other clusters with smaller Euclidean distances between the outlier elements and the group centroids. This procedure can be continued until no outliers exist.

2.3 A First Example: A One-Parameter Oscillation

The solutions suggested in sections 2.1 and 2.2 are demonstrated by a first example. A one parameter sine-oscillation is selected described by 200 values between 0°–360°. Its regular course is replaced by 10-value steps. Correct clustering of the oscillation is achieved if (i) the boundaries between the clusters are identical with those between the steps of the oscillation and (ii) the partition of the clusters is symmetric in the following two aspects: First, the positive (and the negative) parts of the oscillation from 0°–180° (and 180°–360°) must be symmetric about 90° (and 270°). Second, the positive region 0°–180° must mirror the negative region 180°–360° symmetrically. Two variants of clustering procedures are applied:

a) The statistically significant cluster separation (section 2.1) and the calculation of the

optimal initial number of clusters (section 2.2) and

b) the ''standard method'', the defined number of clusters is set to $K = 5$. Five clusters are chosen because of the same calculated optimal number in variant (a).

For variant a) we start with the initial number of clusters $K_0 = 5$ computed as described in section 2.1. The number of separated clusters in this case is also $K = 5$. All conditions of a correct clustering discussed above are fulfilled (Fig. 1a). This example shows that a correct solution is achieved by the clustering, if the proposed improvements are incorporated. The boundaries of the clusters coincide with the steps of the oscillation and the symmetry is fulfilled both within the positive and negative part. Figure 1b shows the result of variant b). Note that the positive and negative parts are asymmetric with

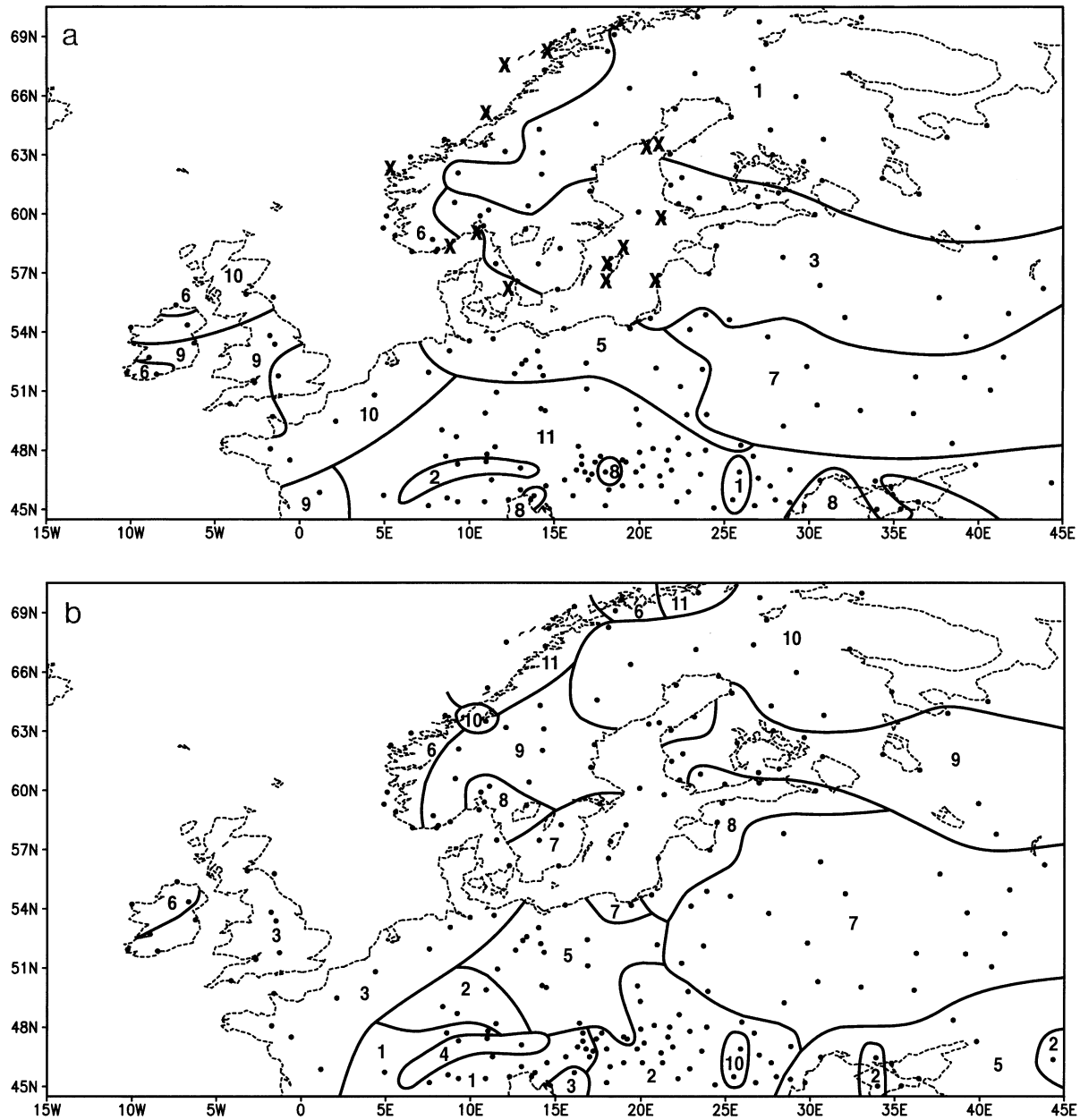

Fig. 3. Climate classification with a) the optimal initial number of clusters and statistically separated clusters (x–climate type 4; dots–positions of the stations), and b) the defined number of clusters $k = 11$ calculated by the ''standard'' non-hierarchical cluster analysis algorithm

respect to each other. That is, the positive part of the oscillation includes three clusters (1–3), whereas the negative one contains two clusters (4–5). Additionally, cluster 4 contains the zero-level. That is, the ''standard method'' leads to a significant error in the clustering procedure. For comparison, a standard hierarchical procedure (WARD, Steinhauser and Langer, 1977) is tested. It shows that the algorithm used identifies 161 secondary minima during the whole iteration process which leads to an incomplete distribution of the sample values into the calculated clusters because ambiguous minima occur. Therefore, this hierarchical procedure will not be used for further investigations.

3. The Climate of Europe

The aim of this section is to classify European climate (between 45° and 70° N, 12° W and 45° E) by regional climate types using monthly and annual means of precipitation, surface air temperature, and the monthly means of the daily temperature range for the time period 1979–1992 at 228 meteorological stations (locations see Fig. 3a). The stations are part of the PIK climatological data bank system (Potsdam Institute for Climate Impact Research). Again, the improved clustering variant a) is compared with variant b) applying the ''standard method'' prescribing $K = 11$ clusters (derived in variant a)). The following results are summarised:

For variant a) the calculated optimal initial number of clusters is $K_0 = 7$. Figure 2 shows two parts in the course of the target function with growing cluster number. In the first part, the values decrease continuously with an increasing number of clusters; in the second, one observes only random oscillations of the target function values. The clustering with statistically significant cluster separation (see Section 2) yields 11 climate types (clusters) shown in Fig. 3a; eight of the 228 stations are marked as outliers seven of which can be attributed to other climate types (see Section 2.2):

- Eleven climate types classify the whole region of Europe neither too subtly nor too coarsely.
- All climate types (clusters) are represented by a sufficient number of stations (between 8 and 60, except for the Alps).
- The three mountain stations (Saentis, Sonnblick, Zugspitze) of the Alps fall into one cluster (cluster 2).

Table 1. *Selected Clusters of Variant (b)-Cluster 3 and of Variant (a)-Cluster 9 and 10 (D–Germany; F–France; GB–Great Britain; I–Italy; IR–Ireland; CR–Croatia)*

Variant (b) = ''standard method'' Cluster 3		Variant (a) = improved method Cluster 9	Cluster 10
Dublin	IR	Dublin	
Sheffield	GB	Sheffield	
Bradford	GB	Bradford	
Cherbourg	F	Cherbourg	
Long Asthon	GB	Long Asthon	
Plymouth	GB	Plymouth	
Shannon	IR	Shannon	
Portoroz	CR	Portoroz	
Limoges	F	Limoges	
Durham	GB		Durham
Oxford	GB		Oxford
Edinburgh	GB		Edinburgh
Beauvais	F		Beauvais
Angers	F		Angers
Renns	F		Renns
Uccle	B		Uccle
Münster	D		Münster
Armagh	GB		Armagh
Hamburg	D		
Trieste	I		

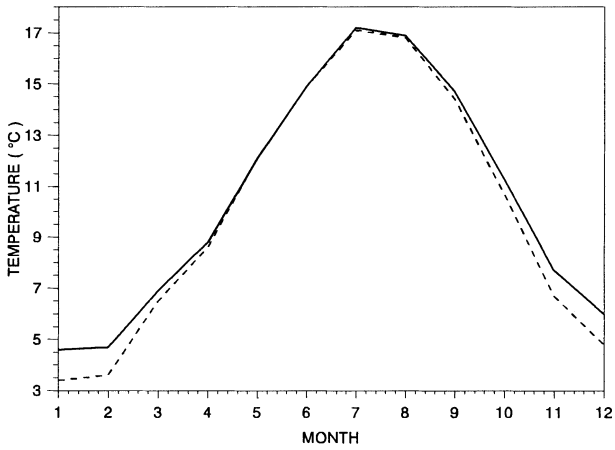

Fig. 4a. Monthly mean air temperatures–variant a): climate type 9 (full) and climate type 10 (dashed)

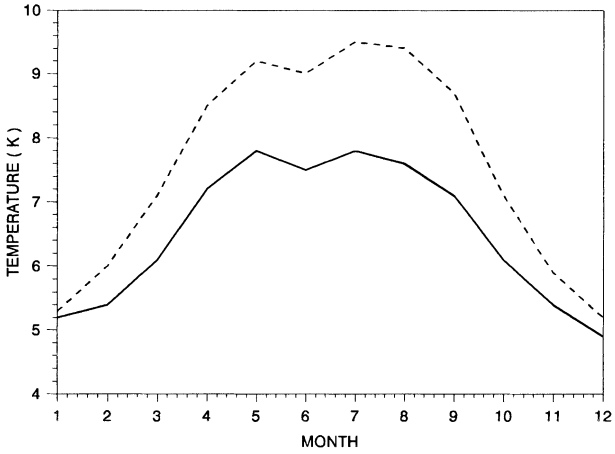

Fig. 4b. Monthly mean daily ranges of air temperature–variant a): climate type 9 (full) and climate type 10 (dashed)

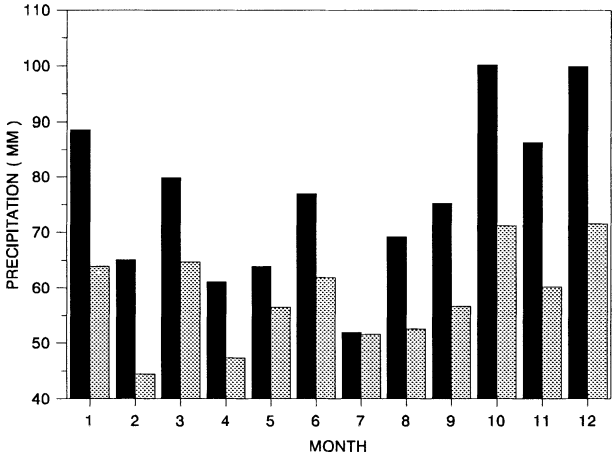

Fig. 4c. Monthly sum of precipitation–variant a): climate type 9 (black) and climate type 10 (grey)

– The stations of each cluster represent connected areas.

In variant b) the "standard method" is applied with the subjective initial number $K = 11$ clusters as determined by variant a). Without the use of the statistical cluster separation technique (section 2) the $K = 11$ clusters obtained are not significantly separated. This leads to the differences with variant a) when comparing Fig. 3a and 3b. An example (see Table 1) shows that these differences are not negligible. Table 1 contains the stations attached to cluster 3 of variant b) and those of clusters 9 and 10 of variant a). We can see that the stations in cluster 3 are the same as those in clusters 9 and 10, except for two stations: Hamburg moves to cluster 5, Trieste to cluster 11. It is obvious that the new classification of Hamburg and Trieste is climatologically more plausible. Furthermore, the question arises whether the differences between clusters 9 and 10 are climatologically significant. If there are differences, cluster 3 of variant b) does not represent an optimal classification. To answer this question, the annual cycles of the parameters of the two clusters are compared (Fig. 4a to c). Large differences are evident for the daily temperature range and the monthly sums of precipitation but for air temperature exist only during winter months. That is, the "standard" cluster algorithm does not lead to an optimal climate classification, despite the use of the optimal number of clusters.

4. Conclusions

The following improvements are suggested when applying non-hierarchical cluster analysis methods. Implementation of these improvements leads to a cluster analysis with an optimum multivariate classification:

(1) As the "standard method" without statistical significant cluster separation can produce grouping errors it is desirable to apply a statistical test for cluster separation quality.
(2) The choice of the initial number of clusters is of great importance for the optimum cluster separation. This number is calculated by estimating a trend-change within the sequence of target function values.

(3) The separation quality can be improved by identifying outlier elements within each separated cluster. These outliers are then sorted into that cluster which reveals the smallest distance between the outliers' parameters and the respective cluster centroid.

References

Bacher, J., 1996: *Clusteranalyse*. München: Oldenbourg, 424 pp.

Ferguson, Th. S., 1961: Rules for rejection of outliers. *Ref. Inst. Internat. Statist.,* **29**, 29–43.

Forgy, E. W., 1965: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics,* **21**, 768.

Gerstengarbe, F.-W., Werner, P. C., 1997: A method to estimate the statistical confidence of cluster separation. *Theor. Appl. Climatol.,* **57**, 103–110.

Jahnke, H., 1988: *Cluster Analysis as a Procedure in Inferential Statistics. On a Graphic Consistency Conception for Cluster Analysis Procedures*. Göttingen: Vandenhoeck & Ruprecht, 168 pp.

Mann, H. B., Whitney, D. R., 1947: On a test of whether one of two random variables is stochastically larger than other. *Ann. Math. Statist.,* **18**, 52–54.

Müller, P. H., Neumann, P., Storm, R., 1973: *Tafeln der mathematischen Statistik*. Leipzig: VEB Fachbuchverlag.

Pettitt, A. N., 1979: A non-parametric approach to the change-point problem. *Applied Statistics,* **28**, 126–135.

SAS, 1990: *SAS/STAT User's Guide. Version 6*. Fourth Edition, Vol. 1, Cary: SAS Institute Inc., NC, 890 pp.

SPSS, 1999: *SPSS for Windows Manual*. Chicago: SPSS Inc.

StatSoft, 1994: *Statistica*. Vol. III, Statistics II. Tulsa: StatSoft Technical Support.

Steinhausen, D., Langer, K., 1977: *Clusteranalyse–Einführung in Methoden und Verfahrender automatischen Klassifikation*. Berlin: Walter de Gruyter, 411 pp.

Authors' addresses: Dr. habil. F.-W. Gerstengarbe, Dr. habil. P. C. Werner, Potsdam Institute for Climate Impact Research, P. O. Box 601203, D-14412 Potsdam, Federal Republic of Germany. Prof. K. Fraedrich, Meteorological Institute, University of Hamburg, Bundesstrasse 55, D-20146 Hamburg, Federal Republic of Germany.