PIK Report

No. 50

THE COMPLETE NON-HIERARCHICAL CLUSTER ANALYSIS

by

F.-W. Gerstengarbe & P. C. Werner

POTSDAM INSTITUTE
FOR
CLIMATE IMPACT RESEARCH (PIK)

Corresponding author:
Dr. F.-W. Gerstengarbe
Potsdam Institute for Climate Impact Research
P.O. Box 60 12 03, D-14412 Potsdam, Germany
Phone: +49-331-288-2586
Fax: +49-331-288-2695
E-mail: gerstengarbe@pik-potsdam.de

Herausgeber:
Dr. F.-W. Gerstengarbe

Technische Ausführung:
U. Werner

POTSDAM-INSTITUT
FÜR KLIMAFOLGENFORSCHUNG
Telegrafenberg
Postfach 60 12 03, 14412 Potsdam
GERMANY

Tel.: +49 (331) 288-2500
Fax: +49 (331) 288-2600
E-mail-Adresse: pik-staff@pik-potsdam.de

POTSDAM, JANUAR 1999

Abstract

Cluster analysis contains several multivariate methods for the separation of patterns (clusters). Definition of the optimum, or globally best, cluster analysis is an unresolved issue. Two methods are of special importance: 1. The statistical security of cluster separation. 2. The definition of the optimal number of clusters. On the basis of non-hierachical minimum-distance cluster analysis a new method is described that allows a separation of clusters in a statistically well-founded way. Applying this extended non-hierarchical cluster analysis algorithm, the following additional problems need to be solved: The generation of a suitable initial partition, the estimation of the initial number of clusters, and the error reduction by delimitation of the level of significance for cluster separation. The following solutions are proposed: Random ranking of the initial partition, derivation of the cluster number using target function and Pettitt-test, and estimation of outliers including a new classification with the clusters. The complete method is tested and discussed using a theoretical and a practical example. For the practical example, a climate classification of Europe is established which shows that the proposed improvements can be of great practical relevance.

1. Introduction

The main idea of the cluster analysis is to relate to each other an existing number M of elements e_i which are each described by N parameters p, i.e.:

$$e_i = f\,(p_{i1},\ \ldots,\ p_{iN}\,). \tag{1}$$

Two main techniques are possible:
Using hierachical methods, different sequences of groups on different levels are constructed. The result is an hierachy of clusters in a "tree structure". The disadvantage of this technique lies in the fact that an exchange of elements is impossible if the "tree structure" is built up. This disadvantage restricts the application.
With the non-hierachical methods, the elements e_i are simultaneously partitioned into a given number of clusters K : by displacing the elements between the clusters in case of a given quality criterion, a given initial partition is built up step by step, and developed into steadily improving groupings until reaching the optimum. For more details, see Steinhausen and Langer (1977). The starting point for concerning the - description of the following method is the non-hierarchical minimum-distance method according to Forgy (1965). The starting condition when applying the above method is to have the elements e_i equally distributed over a number K of given clusters (initial partition). In the case of M given elements and K clusters each cluster receives $L = M/K$ elements as follows:

$$\begin{array}{llll} e_1, & \ldots, & e_L & \ni\ c_1 \\ e_{L+1}, & \ldots, & e_{2L} & \ni\ c_2 \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ e_{(k-1)L+1}, & \ldots, & e_{kL} & \ni\ c_k \end{array} \tag{2}$$

(The number of clusters K must be defined empirically; the number of elements depends on the data series and the problem which has to be investigated.)

A so-called group centroid $\bar{e}_k$ is then calculated for each k of the K clusters (cluster mean value under consideration of those existing parameters that have to be normalized accordingly in the case of different scalings):

$$\bar{e}_k = \frac{1}{L} \sum_{i=(k-1)L+1}^{kL} e_i \tag{3}$$

By applying the Euclidean distance, the following objective function $a(g)$ for each grouping step g can be defined:

$$a(g) = \sum_{k=1}^{K} \sum_{i \in k} |e_i - \bar{e}_k|^2 \tag{4}$$

By considering the Euclidean distance, each grouping step can be seen as a displacement of the element e_i into that cluster which contains the respective nearest centroid. The objective can thus be minimized:

$$a(g) \; \forall \; g \rightarrow \min \tag{5}$$

This procedure is repeated until a local minimum of the objective function is reached. The objective function reaches a local minimum if two successive grouping steps show the same result; the iteration is in this case discontinued, i.e., the optimum classification with respect to the given number of clusters has been reached.

An important disadvantage of this method is that one does not know whether an absolute or just a secondary minimum of the objective function has been obtained (Fovell, 1993; Milligan, Cooper, 1985). That is why the quality of separation is unknown, as is the objective number of clusters. The following procedure shows a solution of this problem.

2. Definition of a quality criterion to separate clusters

The quality criterion represents the statistical security of the cluster separation. The basic idea to define this criterion can be described as follows:
After having reached the local minimum, each cluster is equipped with a generally varying number of elements. Each element is defined by N parameters, i.e., it is located in a N-dimensional parameter space. As each cluster consists of a certain number of elements, they each represent a scatterplot of elements in the above space. If the clustering leads to a local secondary minimum, overlaps occur between the scatterplots of single clusters. The principle of this method is presented in figure 1, which depicts the projection of two parameters within the N-dimensional space.

The number of overlaps O of the two clusters a and b of N parameters can accordingly be defined as follows:

$$O_{a,b} = \sum_{i_a=1}^{L_a} \sum_{i_b=1}^{L_b} \sum_{j=1}^{N} o_{i_a,\, i_b,\, j} \qquad \begin{matrix} a = 1, \ldots, k-1 \\ b = 2, \ldots, k \end{matrix} \tag{6}$$

with

$$o_{i_a,\, i_b,\, j} = \begin{cases} 1 & p_{i_b,\, j} \geq p_{i_a,\, j} \\ 0 & p_{i_b,\, j} < p_{i_a,\, j} \end{cases} \tag{7}$$

under the additional condition

$$\bar{e}_1 > \bar{e}_2 > \ldots > \bar{e}_k \tag{8}$$

If $O_{a,b} = 0$, than the clusters a and b are completely separated from each other. The maximum possible number of overlaps is

$$O_{a,b}^{\max} = NL_aL_b \tag{9}$$

This number is reached if both clusters cover the same region within the N-dimensional space.

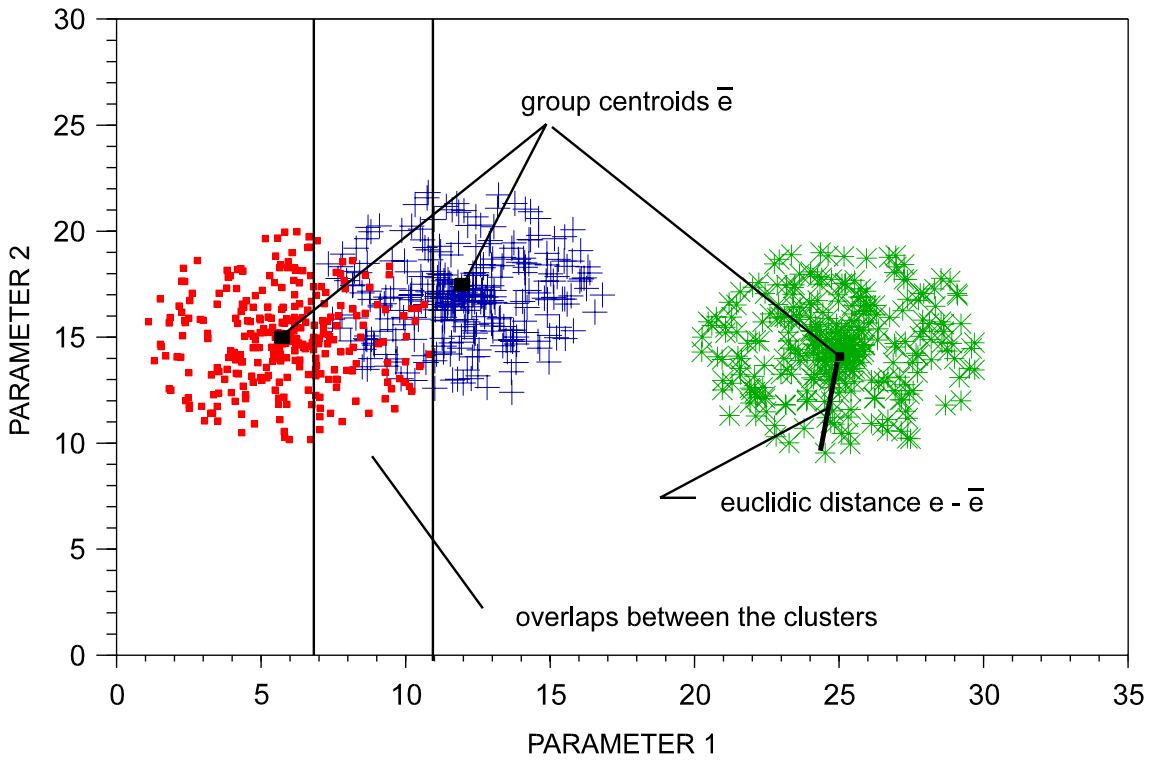

Fig. 1 Principle scheme of the description of the clustering quality (square/cross - overlapped clusters, double cross - full separated cluster)

Thus by applying the equations (6) to (9) the quality of the separation of clusters can be determined statistically by the following steps:

1. Calculation of the mean number of the maximum possible overlaps $\bar{O}^{\max}$, as well as the mean actual number of overlaps $\bar{O}$ over all combinations of cluster pairs.
2. Subsequently, a test is carried out to see whether $\bar{O}$ and $\bar{O}^{\max}$ originate from the same basic population. Assuming that there is a normal distribution, Student's t-test can be used. (Because of the necessary normalization of the parameters, a normal distribution is generally realized.) The null hypothesis implies that both mean values originate from the same population. The clusters

can be separated only when the null hypothesis is rejected. Otherwise, the procedure is as follows:

3. The ratio $\nu_{a,b}$ of the actual to the maximum possible number of overlaps is determined for each cluster pair:

$$\nu_{a,b} = \frac{O_{a,b}}{O_{a,b}^{\max}} \tag{10}$$

4. The mean value $\bar{\nu}$ over all $\nu_{a,b}$ is calculated. It is the empirical estimate of the actual occurence probability of overlaps.
5. In the case that not all mean values $\bar{\nu}$ are identical, paragraph 2. implies that there is - according to the chosen level of significance - a statistically significant separation of those clusters for which $\nu_{a,b} < \bar{\nu}$.
6. The quality of separation in the case $\nu_{a,b} > \bar{\nu}$ still needs to be determined. The point is hence to clarify whether a certain value of the number of the actual overlaps $O_{a,b}$ is compatible with the mean value of all numbers of the actual overlaps $\bar{O}$ or not. If one interprets the overlaps as empirical occurence frequencies, a statistical comparison between both is possible. This can be done for instance by the χ^2-test (e.g. Taubenheim, 1969) which can be written as follows:

$$\chi^2 = \frac{(O_{a,b} - \bar{O})^2 * (2O_{a,b}^{\max} - 1)}{(O_{a,b} + \bar{O}) * (2O_{a,b}^{\max} - O_{a,b} - \bar{O})} \tag{11}$$

with the degree of freedom $d_f = 1$.

The result of the test can be interpreted in the following way:
If the calculated χ^2-value is greater than a given threshold of significance, the frequency of overlaps exceeding the mean value $\bar{O}$ differs significantly from the χ^2-value. The separation between the clusters is hence statistically not significant, in contrast to the other case where a statistically reliable separation exists.

3. Determination of an optimum number of clusters

The optimum number of clusters is defined as that number which realize the best separation between all clusters. The method presented above allows the optimum number of clusters for the non-hierarchical clustering to be determined in the best possible way. The following procedure is required to this end:

1. If a clustering with a given initial number of clusters does not lead to a separation, then the initial number of clusters is varied until at least a single statistically reliable separation between one cluster and the rest exists.
2. If paragraph 1. is fulfilled, the elements of the separated clusters are noted as being a final partial result.

3. The initial series is reduced by the separated cluster elements.
4. This algorithm is repeated using the method presented in section 2 until all clusters are statistically reliably separated.
5. The optimum number of clusters results from the amount of clusters separated per algorithm step.

Nevertheless, some problems applying this method remain:

A) a correct provision of the initial partition (Is valid also for all other cluster analysis methods).
B) an estimation of the optimal initial number of clusters
C) a reduction of the error appearing in connection with the delimitation of the level of significance for cluster separation.

In the following, solutions to the problems A) to C) are proposed and discussed. Section 4 contains the theoretical basis of the improvements. In sections 5 and 6 two applications are discussed in detail: First, the theoretical mechanisms are discussed by a one parameter oscillation. Then, the practical application is demonstrated by a calculation of a climate classification for Europe.

4. Theoretical basis of the improvements

The structure of the initial partition (section 4.1) and the choice of the initial number of clusters (section 4.2) play an important role. After achiving a statistically significant cluster separation, an error margin remains which, in general, is of the magnitude between 1% or 5%. Normally, this error can be neglected; however, cases occur where this error needs to be considered (as shown in the application of section 5.2). A possible way to reduce the error is presented in section 4.3.

4.1 The initial partition

For each statistical investigation, the elements of the sample must be independently and identically distributed. This principle is also valid for cluster analysis. If neglected the following course of events may appear:
In the first step of the clustering, the elements of the sample are regularly distributed in the initial number of clusters. In this case, the sequence of the distribution depends on the position of each sample element. That is, in each cluster of the initial partition there is a number of elements which are sorted one following the other in the sample. Thus these elements are not neccessarily independent which means that the structure of the sample may create “pre-grouping”. As a consequence, a greater number of dependent elements must exist within a sample. Then, a secondary minimum of the target function can be reached already after only a few iterations devoid of an optimal grouping. This defect can be avoided in a simple way by a random ranking of the elements of the sample so that the persistence of the series tends to 0.

4.2 The initial number of clusters

Given a sample with a limited number n of elements and their regular distribution in the initial number of clusters. This means that a too large or too small initial number of clusters leads to a situation in which some clusters can be separated significantly before the optimum distribution of the elements has been reached. If, for example, the initial number of clusters is too small, the number of elements within a single cluster is relatively large. As a result possible internal structures of a separated cluster cannot be considered. In the other case, artificial structures can be occur. To estimate the optimum initial number of clusters the following procedure can be carried out:
The starting point for the calculation of the initial cluster number is the target function (eq. 4). 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. If, for a varying number of clusters (from 2 to m), the value of the target function is calculated, a series is obtained whose values can be included for the estimation of the optimum initial number of clusters. As each value of the target function is equivalent to a specific number of clusters, the initial number of clusters can be defined as that inflection point within the series (of target function values) where a trend disappears. From this point on significant changes within the series do not exist. This idea can be solved practically with the following steps:

- Calculation of the differences between neighbouring values of the target function series and creation of a differential series with $m1 = m - 1$ values,
- Using the Pettitt-test (Pettitt, 1979) to estimate the beginning of a trend (inflection point) within the differential series.

The Pettitt-test can be derived from the U-test (Mann-Whitney, 1947), based on the ranks of the series. The inflection point is defined as that point for which the absolute value of X_k reaches a maximum with

$$X_k = 2 \cdot R_k \quad k \cdot (m1 + 1) \tag{12}$$

where

$$R_k = \sum_{i=1}^{k} r_i \tag{13}$$

k is the position within the series, $m1$ is the number of values of the differential series, and r_i is the rank of the i th target function value. Continuously increasing the initial number of clusters, the Pettitt-test finally defines that position within the series of the target function values where the series is divided into a part with significant changes of the target function values and another one without changes.

4.3 The error margin

In general, the test described in section 2 is connected with an error probability of 1% or 5%. That is, in spite of a statistically significant separation of two clusters, a small number of overlaps can occur so that some clusters may contain “strange” elements. In a statistical sense, this case is without any consequence. In some cases, however, such outliers can have a negative influence on the clustering.

This problem can be circumvented as follows using the definition of an outlier as a value deviating significantly from the basic sample: After a significant separation of all clusters has been achieved, the distance between each element within the cluster and the group centroid is calculated. These distances within each cluster are defined as a basic sample and utilised for identifying outliers. Here we suggest the Euclidean distance (eq. 1) as the measure for the estimation of the outliers. For each element of a cluster, we calculate the sum of Euclidean distances between the single parameters and the group centroid. This leads to a sample of these sums for each cluster. Using the Thompson-rule (Müller et al., 1973) we can estimate the outliers of the clusters. The test value is defined as:

$$t_i = \frac{x_i - \bar{x}}{s^*} \quad (i = 1, \ldots, n) \tag{14}$$

where $\bar{x}$ is the arithmetical mean of the sample and s^* the standard deviation of the sample. Outliers are all values x_i $(i = 1, \ldots, n)$ for which $|t_i| > z_{m;\,\alpha}$ is valid, with $m = n - 2$ ($z_{m;\,\alpha}$ = 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, the Euclidean distance makes it possible to test a better assignment of the outlier to another cluster.

5 Examples

5.1 A one-parameter oscillation

The solutions suggested to the problems of A) to C) are demonstrated by two applications: 1.) A simple oscillation is decomposed into characteristic patterns; the different patterns make the existing difficulties visible. 2.) Of more practical relevance is the calculation of climatology for Europe.

As an example for a one parameter oscillation, a simple sine-oscillation is selected and described by 200 values. Its regular course is replaced by 10-value steps in form of stairs. In case of clustering of the new curve the boundaries between the clusters have to be identical with those between the steps of the curve. The partition of the clusters must be symmetric in two respects: First, the positive part of the oscillation must be symmetric as well as the negative. Second, the positive region must mirror the negative region symmetrically.Three variants of clustering are investigated on the basis of the

discussed procedures:

a1) The defined initial number of clusters is set to k_0 = 8; the initial partition consists of random ranked values

b1) The optimal initial number of clusters is counted; the values of the initial partition are ordered from 1 - 200 in the same course like the sine oscillation

c1) The optimal initial number of clusters is calculated; the initial partition consists of random ranked values.

Figure 2a shows the result of variant a1). One can see that the boundaries of the clusters are coincide with the spots. Additionally, the symmetry is fulfilled within the positive part as well as within the negative part. The positive and negative parts are asymmetric with respect to each other. If we define cluster 4 as "neutral", 3 clusters remain in the positive part, 4 in the negative one, while cluster 1 contains 5 steps and cluster 8 as the pendant only 3.

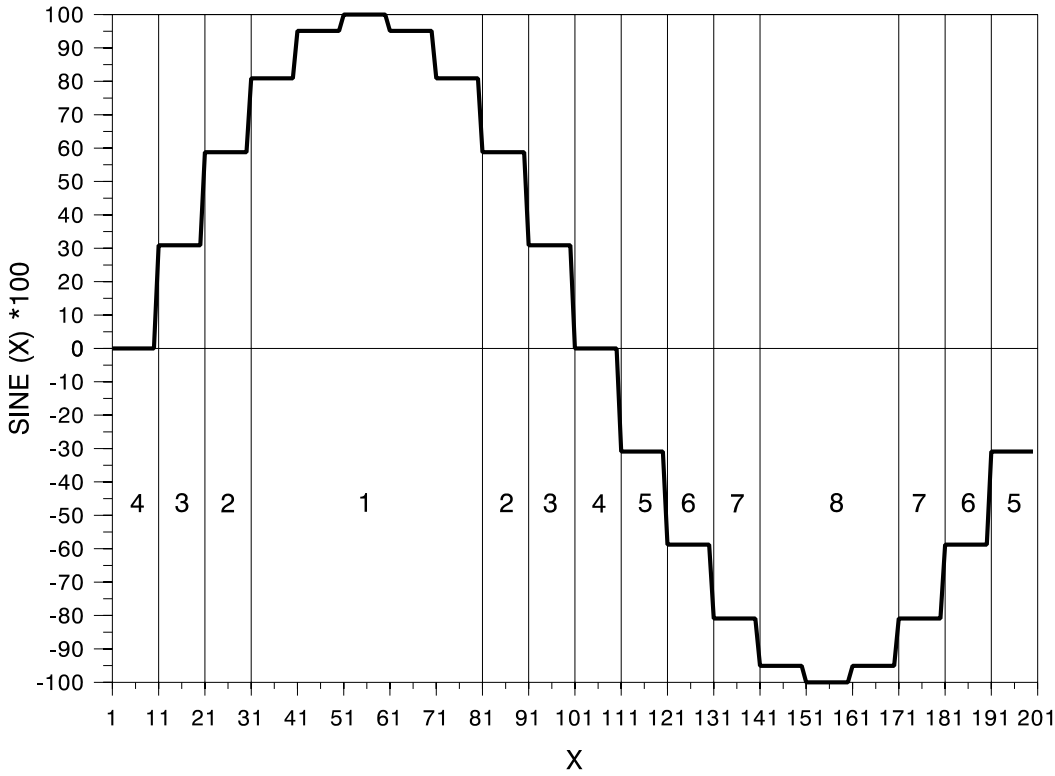

Fig. 2a Theoretical test calculation a1) - a defined initial number of clusters k_0 = 8; random ranked values of the initial partition

The results of version b1) are presented in figures 2b and 3a. Figure 3a gives an overview of the course of the target function values with respect to the number of clusters. Also included is the result of the Pettitt-test with an optimal initial cluster number of 5. This number agrees coincidentally with the optimal separated number of clusters (Fig. 2b). The symmetry in the positive and negative parts is fulfilled, but a "neutral" cluster does not exist. Thus an asymmetry exists between both parts: the positive one includes 3 clusters (1-3), while the negative one only 2 (4-5). This is why the ranges of the clusters are different.

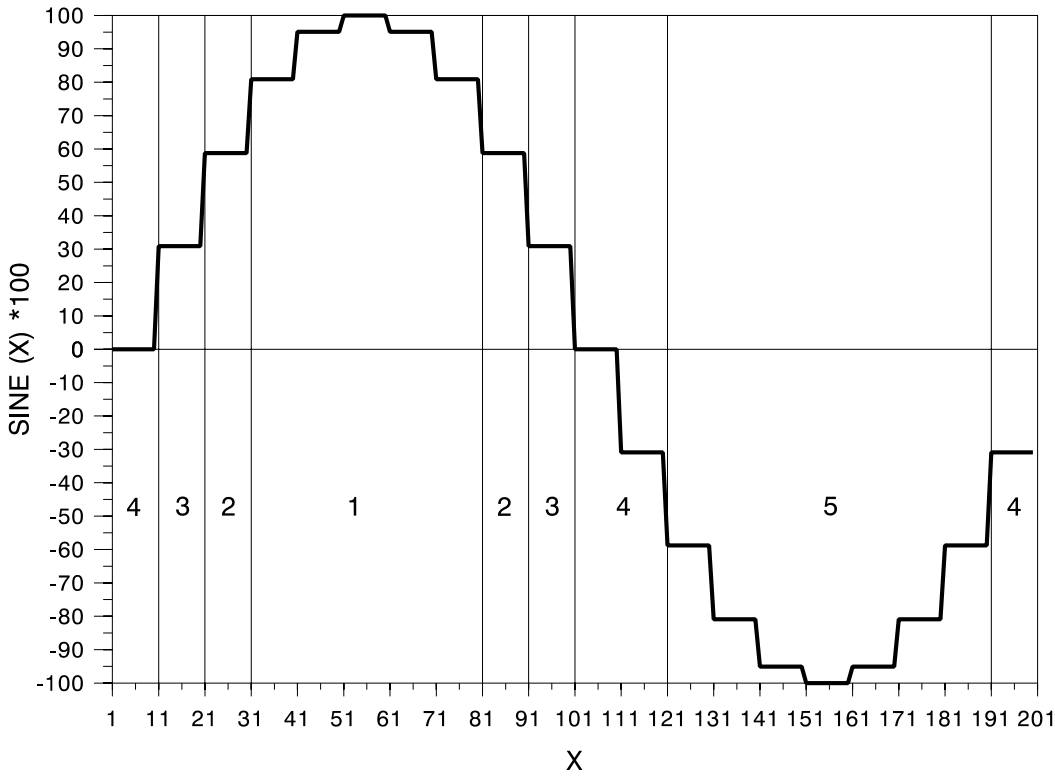

Fig. 2b Theoretical test calculation b1) - optimal initial number of clusters; the values of the initial partition are ordered from 1 - 200 in the same course like the sinus oscillation

For variant c1) we start with the same initial number of clusters as calculated for b1). The number of statistically separated clusters is also 5. In this case all conditions of symmetry are fulfilled (Fig. 2c).

This example shows that a correct solution exists for the clustering, if data of the initial partition are ranked randomly and the optimal initial number of clusters is used.

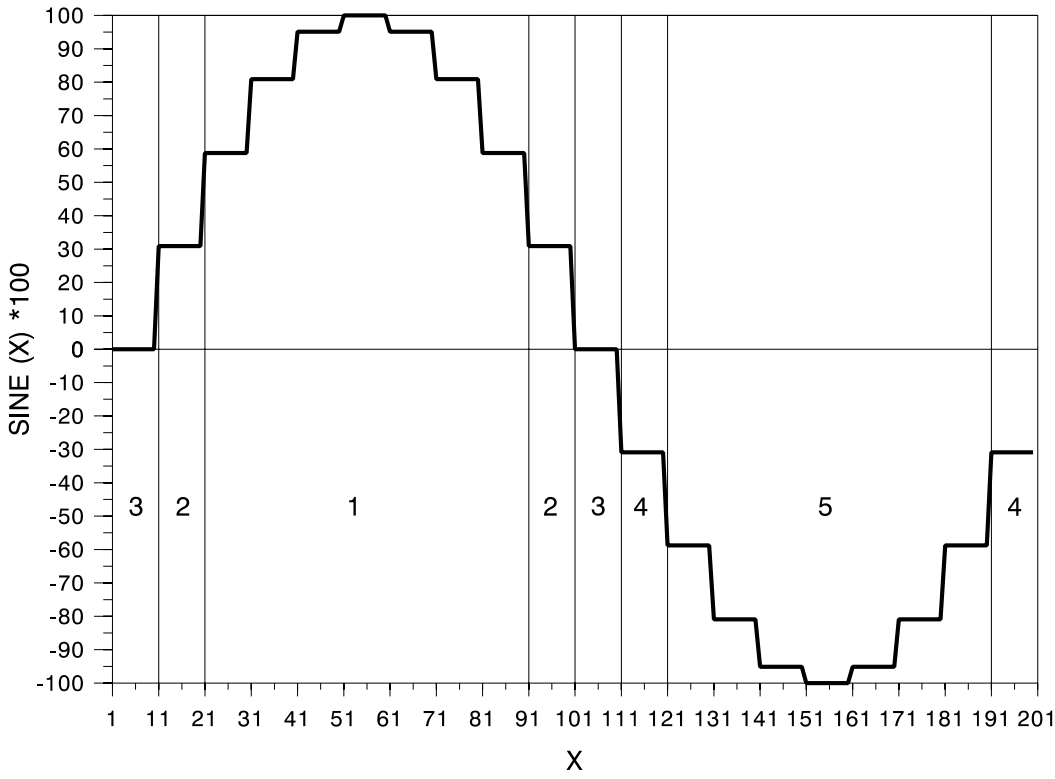

Fig. 2c Theoretical test calculation c1) - optimal initial number of clusters; random ranked values of the initial partition

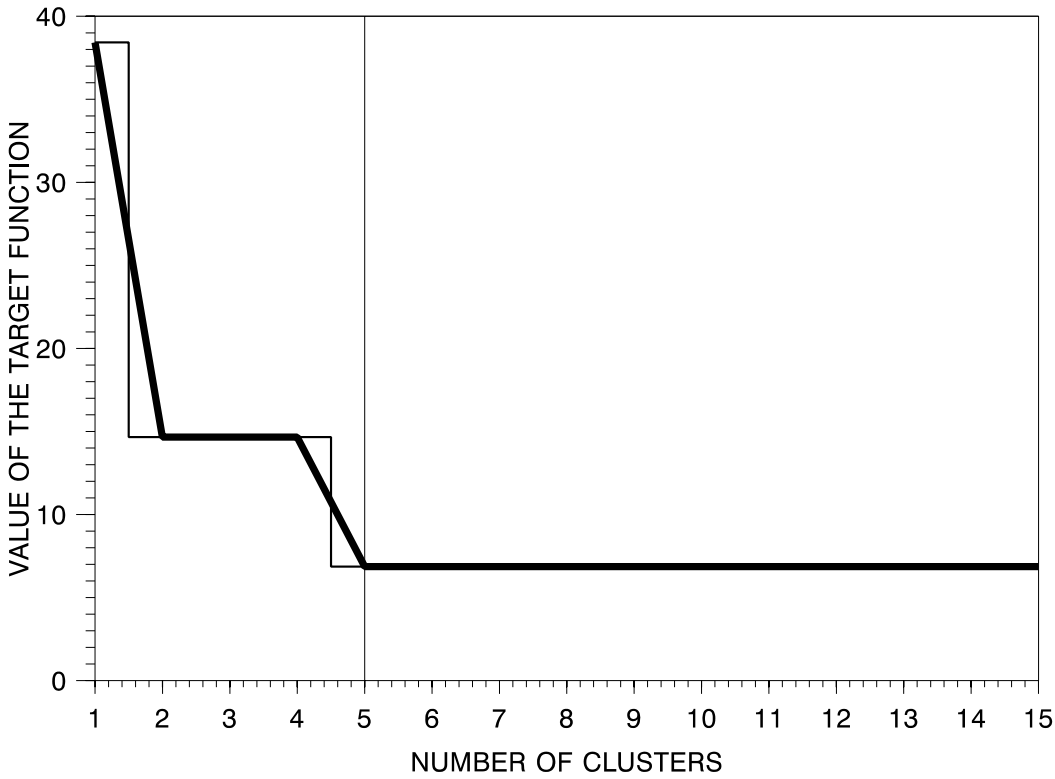

Fig. 3a Result of the Pettitt-test for the estimation of the initial number of clusters (sinus oscillation)

5.2 The climate of Europe

The aim is to classify Europe (between 45° and 70° N, 12° W and 45° E) into regional climate types using monthly and annual means of the air temperature and the sums of precipitation of 228 meteorological stations. Additionally, the monthly means of the daily range of the air temperature are also included. From the results obtained in section 5.1 we start the climate classification of Europe with the optimal variant c1) which is then compared with less satisfying versions. The following variants are discussed:

a2) Calculation of the optimal initial number of clusters; random ranked values of the initial partition (This corresponds to the correct variant c1) of section 5.1)

b2) Calculation of the optimal initial number of clusters; the data of the initial partition are ranked by countries

c2) Two initial numbers of clusters we used k_0 = 5 and k_0 = 15; with randomly ranked values of the initial partition

d2) Clustering using the standard non-hierarchical minimal distance method (without statistical significant cluster separation); number of clusters k = 11; variant d21): random ranked data; variant d22): ranking by countries. (The example d2) is also used to show the results for standard clusterings.)

Variant a2)

The calculated optimal initial number of clusters is k_0 = 7. The figure 3b shows that the course of the target function values can be divided into two different parts. First, the values decrease continuously with an increasing number of clusters; second, one observes only random oscillations of the target function values. Note the plateau at k_0 = 4 which will be relevant for the variant c2). Here in variant a2) we obtain 11 climate types shown in figure 4a. Eight stations of the used 228 are marked as outliers. Seven of them can be related to other climate types calculated before. From a climatological point of view the various results are noted:

- With 11 climate types the whole region is classified 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 3 mountain stations (Saentis, Sonnblick, Zugspitze) of the Alps fall into one cluster (cluster 2).
- Generally, the stations of one cluster are neighbouring stations.

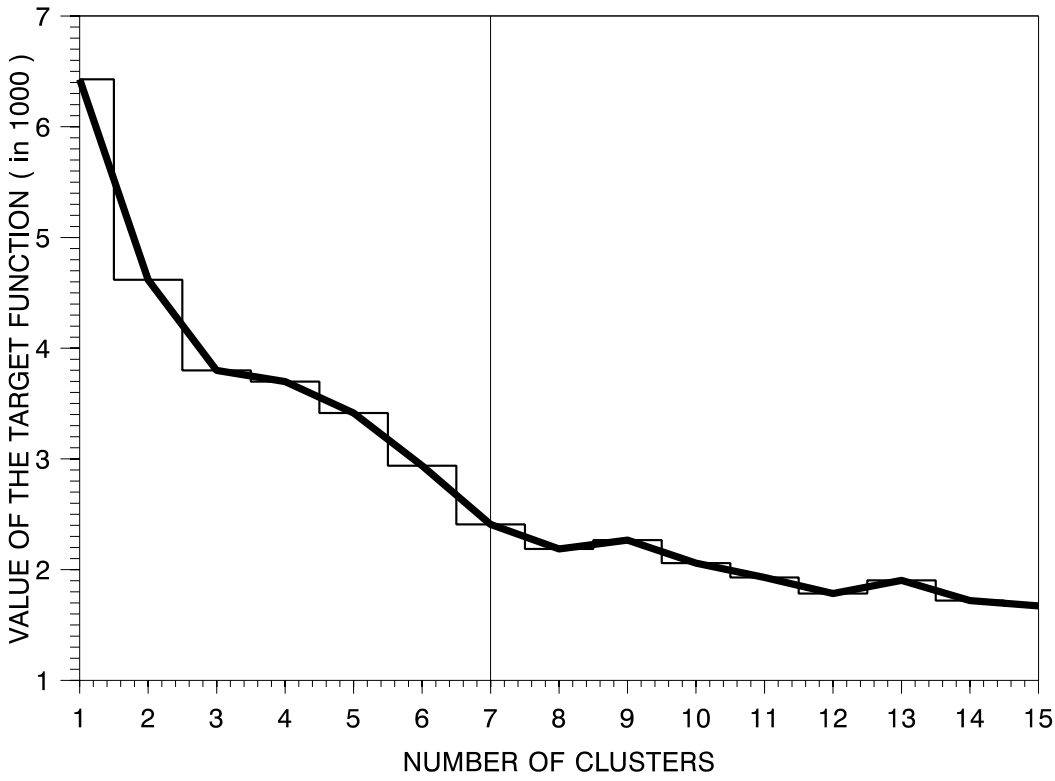

Fig. 3b Result of the Pettitt-test for the estimation of the initial number of clusters (climate classification)

Variant b2)
This variant provides 16 clusters, partially with a small number of stations (2 minimum). A comparison of these results (Fig. 4b) with variant a2) shows significant differences in several regions. Of importance is the fact that two of the Alps stations (Zugspitze, Sonnblick) appear in a cluster with only Norwegian stations while the third Alps station (Saentis) is classified as an outlier which cannot be put into another cluster. A further example of the inaccuracy of this variant is the fact that the two Milan stations (Italy) which differ only subtly, appear in different climate types.

Variant c2)
The initial number of clusters k_0 = 5 leads to a situation where a statistically significant cluster separation is impossible. With a reduction to k_0 = 4 the algorithm works and separates 4 climate types. The reason for this can be found in fig. 3b. We can see that in the case of an additional reduction of the cluster number to 3, the change of the target function is negligible. This means that a statistical solution exists for 4 clusters exists. It is obvious that only 4 climate types for the European climate represent an insufficient classification (s. Fig. 4c). If we increase the initial number of clusters to k_0 = 15, we get an optimum cluster separation for k = 33. For this number of clusters the changes of the target function values are in the noise region. This partition represents a random product with a statistically significant separation. Thus these results makes no sense as a climate classification.

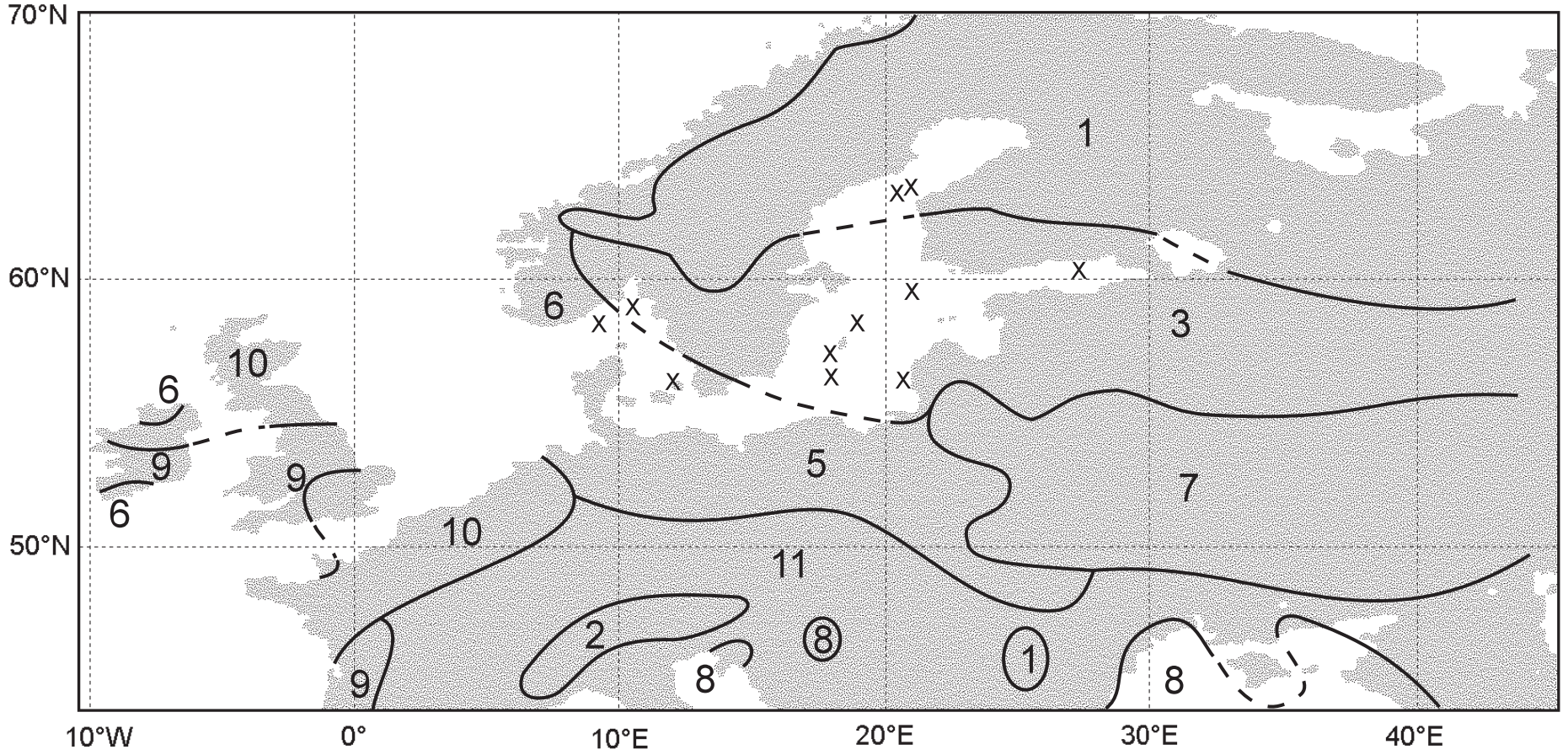

Fig. 4a Climate classification a2) - optimal initial number of clusters; random ranked values of the initial partition (**x** - climate type 4)

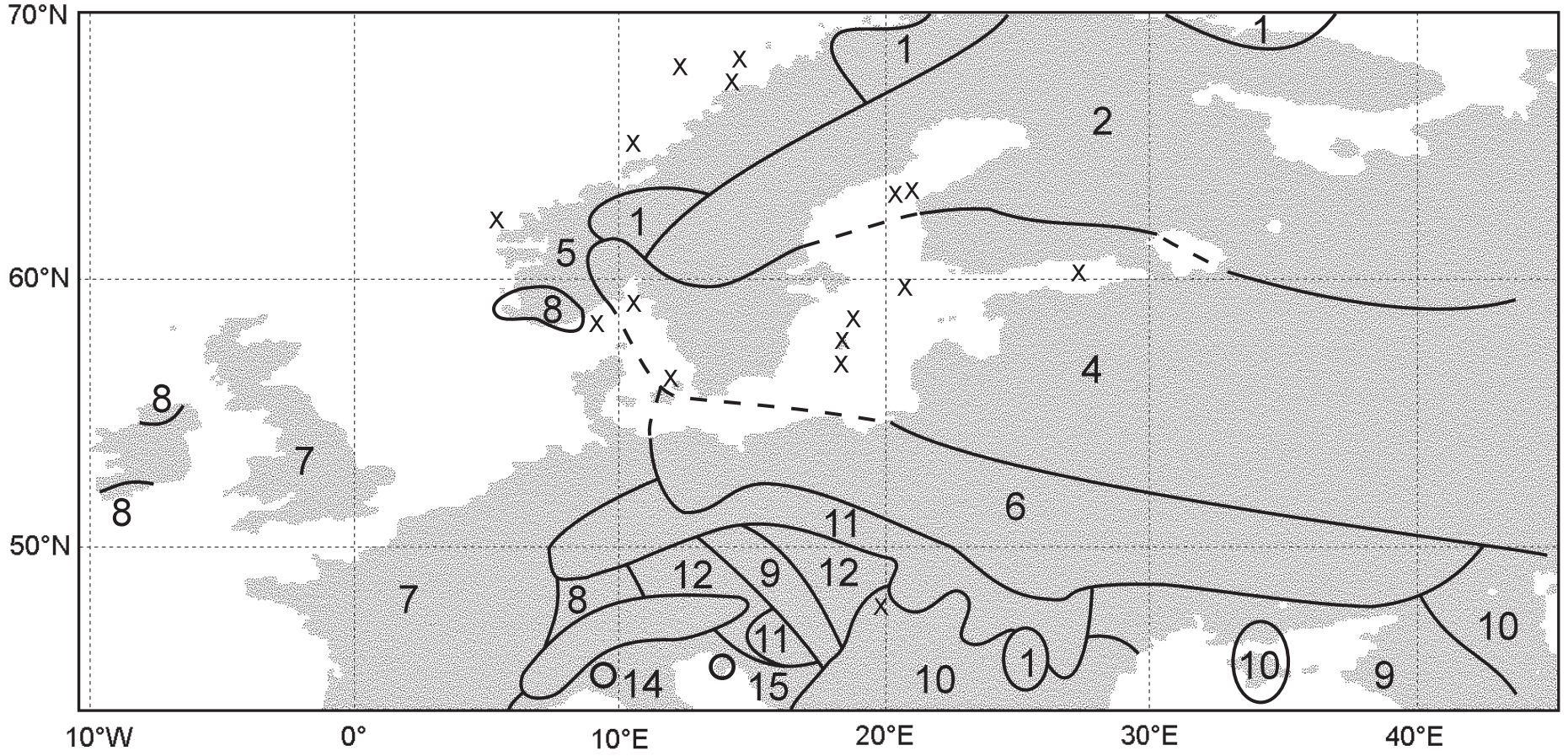

Fig. 4b Climate classification b2) - optimal initial number of clusters; the data of the initial partition are ranked by the countries (**x** - climate type 3; ○ - climate type 16)

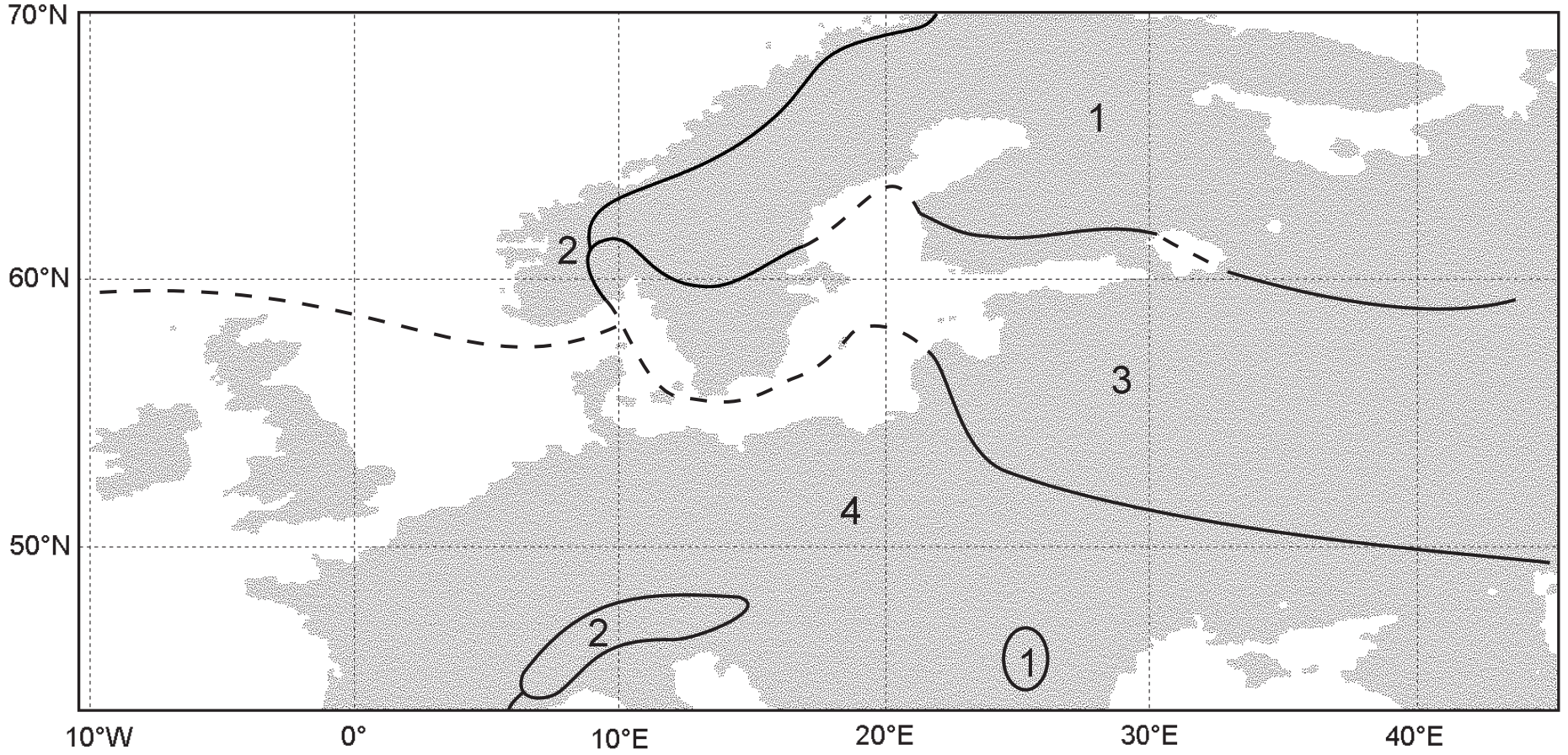

Fig. 4c Climate classification c2) - initial number of clusters k_0 = 5; random ranked values of the initial partition

Variant d21) and d22)

Finally, the standard cluster algorithm is applied: A number of clusters k = 11 is used as calculated in variant a2). For the d21) variant, one gets both reasonable and false climate classifications. For example, the Irish stations Valentia, Cork, Belmullet, and Malin Head are in the same climate cluster as the Alpine stations Zugspitze and Sonnblick. Sorting the initial partition randomly (d22), this misclassification of d21) disappears; and the Alps stations merge into one cluster. This means that data independence is required also in the standard cluster algorithms. Furthermore without the use of the statistical cluster separation the given 11 clusters are not separated significantly. This leads to differences with the variant a2), but they are not as large with the other variants. An example (see table 1) shows that the differences are not negligible. Table 1 contains the stations of cluster 3 of the variant d22) and those of clusters 9 and 10 of the variant a2). We can see that the stations in cluster 3 are the same as in clusters 9 and 10 (except for 2 stations). The question arises wether there are significant climatological distinctions between clusters 9 and 10. To answer this question, the annual course of the parameters of the two clusters are compared (figures 5a up to 5c): A large differences between the parameter air temperature exist only during the winter period. Whereas differences are evident for the daily temperature range, and the monthly sums of precipitation. That is, the standard cluster algorithm does not lead to an optimal climate classification, despite the optimal number of clusters and a randomly ranked initial partition.

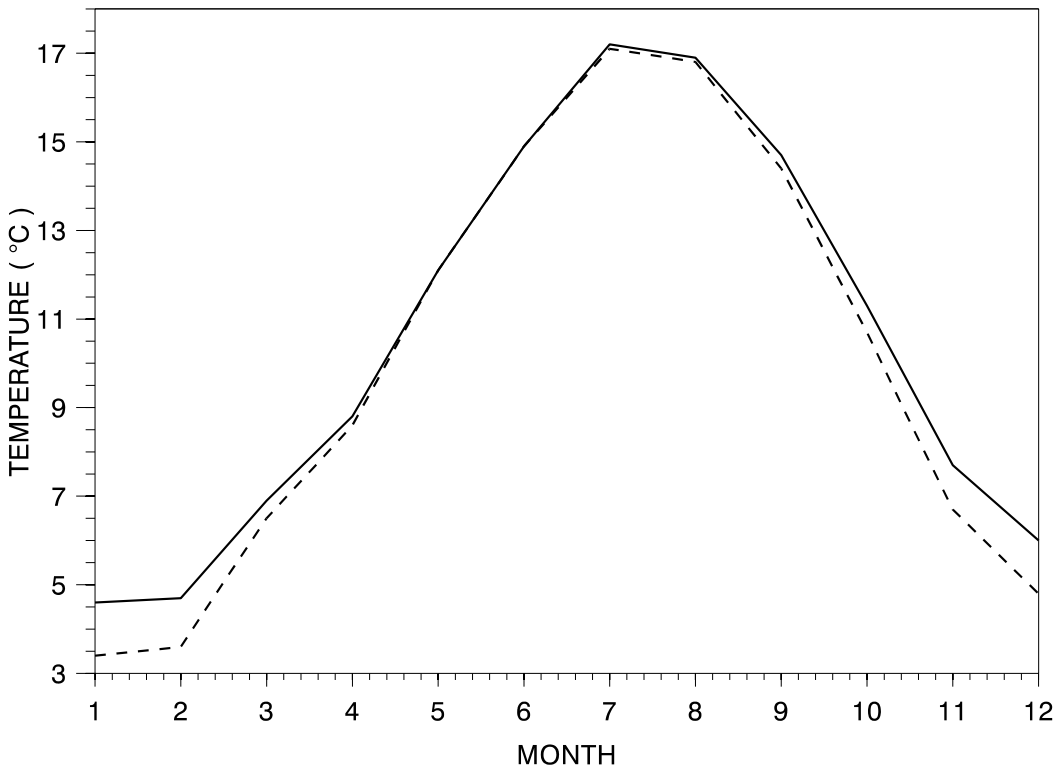

Fig. 5a Monthly mean of the air temperature - variant a2): climate type 9 (full); climate type 10 (dashed)

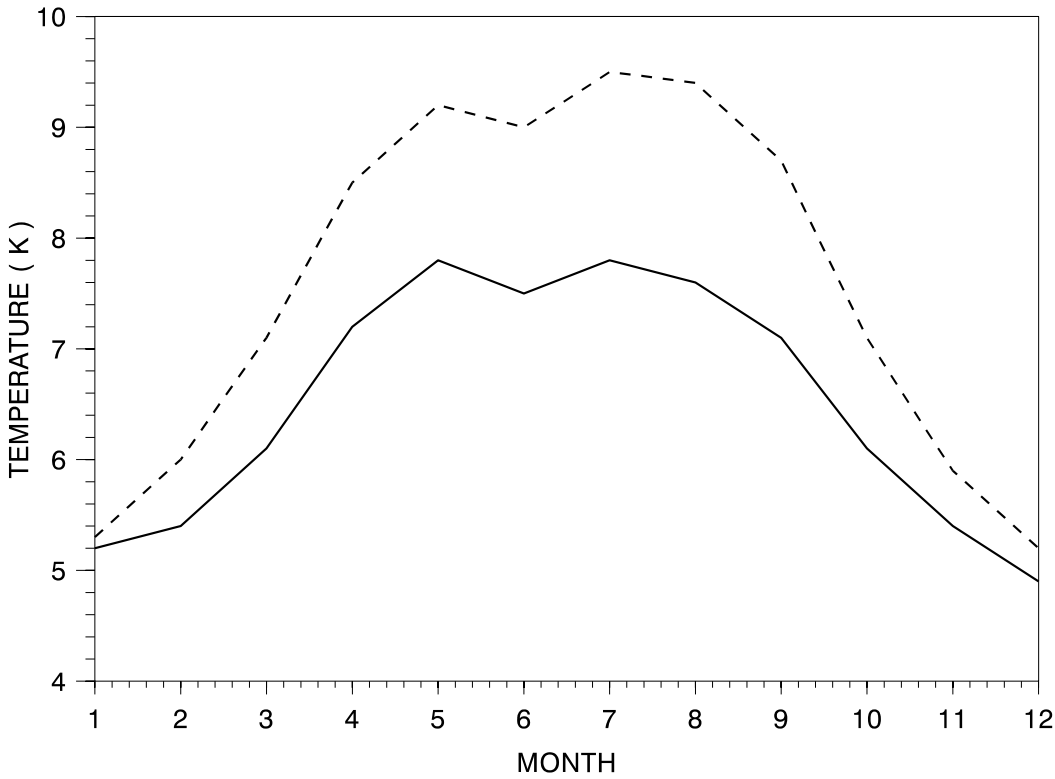

Fig. 5b Monthly mean of the daily range of air temperature - variant a2): climate type 9 (full); climate type 10 (dashed)

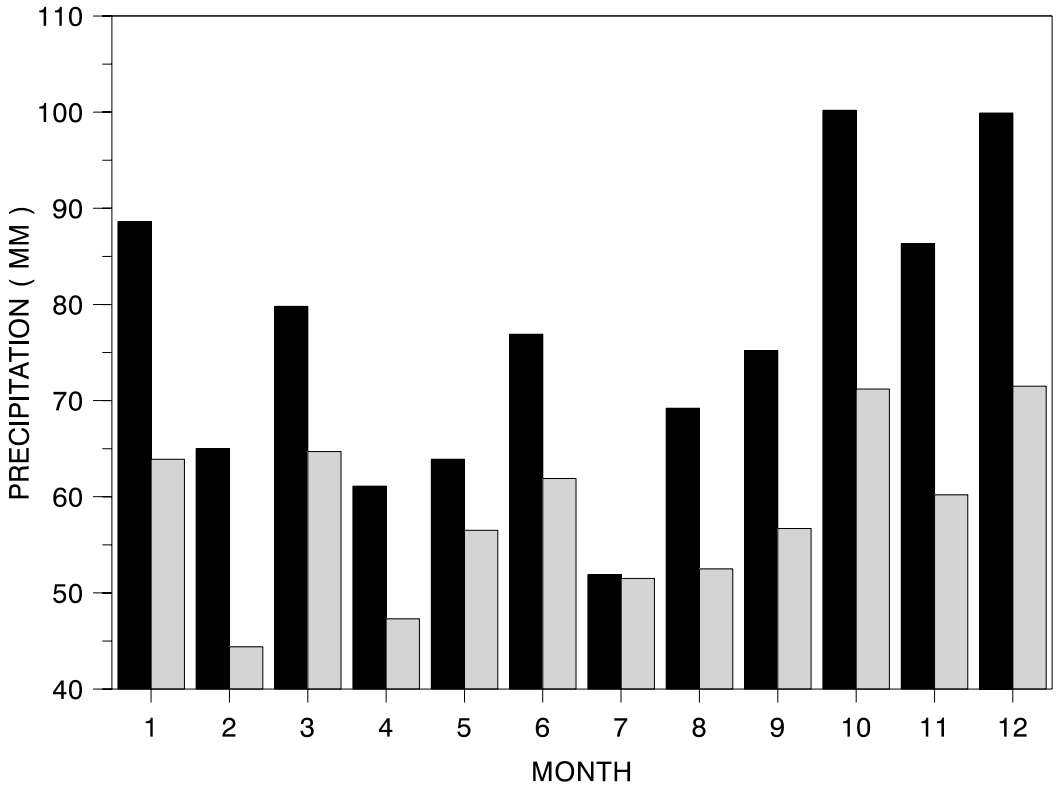

Fig. 5c Monthly sum of precipitation - variant a2): climate type 9 (black); climate type 10 (grey)

Cluster 3		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		

Tab. 1 Selected clusters of variant d22) - Cluster 3 and of variant a2) - Cluster 9 and 10 (D - Germany; F - France; GB - Great Britain; I - Italy; IR - Ireland; CR - Croatia)

6. Conclusions

The presented results show that the suggested procedure is the first which allows the quality of the separation of clusters to be calculated in a statistically well-founded way; it replaces the often adverse effects of a given number of clusters when employing the non-hierarchical cluster analysis by the application of the optimum number of clusters guaranteeing a statistically reliable separation of all clusters from each other. Additionally for all cluster analysis methods the following conclusions can also be drawn:

(1) Each method has to guarantee the statistically significant separation of the clusters.
(2) The ranking of the data within the initial partition must be random.
(3) A computer programme for a cluster analysis has to be built up in such a way that the access to the elements is random.
(4) For the optimum cluster separation the initial number of clusters is of great importance. It can be calculated using the target function values.
(5) It is recommendable that existing outliers sort into the cluster with the smallest distance between the outlier's parameters and the respective group centroid.

Considering these aspects yields a cluster analysis method which fulfills all the demands of an optimum multivariate classification.

References

Forgy, E. W., 1965: Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications (abstract), *Biometrics*, **21**, 768.

Fovell, R. G., Fovell, M. C., 1993: Climate Zones of the Conterminous United States Defined Using Cluster Analysis. *Journal of Climate*, **6**, 2103-2135.

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.

Milligan, G. W., Cooper, M. C., 1985: An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159-179.

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.

Steinhausen, D., Langer, K., 1977: Clusteranalyse - Einführung in Methoden und Verfahren der automatischen Klassifikation. Walter de Gruyter, Berlin, 411pp.

Taubenheim, J., 1969: Statistische Auswertung geophysikalischer und meteorologischer Daten. Akad. Verlagsges, Geest & Portig, Leipzig, 386pp.

PIK Report-Reference:

No. 1 3. Deutsche Klimatagung, Potsdam 11.-14. April 1994,
Tagungsband der Vorträge und Poster (April 1994)

No. 2 Extremer Nordsommer '92
Meteorologische Ausprägung, Wirkungen auf naturnahe und vom Menschen beeinflußte Ökosysteme, gesellschaftliche Perzeption und situationsbezogene politisch-administrative bzw. individuelle Maßnahmen (Vol. 1 - Vol. 4)
H.-J. Schellnhuber, W. Enke, M. Flechsig (Mai 1994)

No. 3 Using Plant Functional Types in a Global Vegetation Model
W. Cramer (September 1994)

No. 4 Interannual variability of Central European climate parameters and their relation to the large-scale circulation
P. C. Werner (Oktober 1994)

No. 5 Coupling Global Models of Vegetation Structure and Ecosystem Processes - An Example from Arctic and Boreal Ecosystems
M. Plöchl, W. Cramer (Oktober 1994)

No. 6 The use of a European forest model in North America: A study of ecosystem response to climate gradients
H. Bugmann, A. Solomon (Mai 1995)

No. 7 A comparison of forest gap models: Model structure and behaviour
H. Bugmann, Y. Xiaodong, M. T. Sykes, Ph. Martin, M. Lindner, P. V. Desanker, S. G. Cumming (Mai 1995)

No. 8 Simulating forest dynamics in complex topography using gridded climatic data
H. Bugmann, A. Fischlin (Mai 1995)

No. 9 Application of two forest succession models at sites in Northeast Germany
P. Lasch, M. Lindner (Juni 1995)

No. 10 Application of a forest succession model to a continentality gradient through Central Europe
M. Lindner, P. Lasch, W. Cramer (Juni 1995)

No. 11 Possible Impacts of global warming on tundra and boreal forest ecosystems - Comparison of some biogeochemical models
M. Plöchl, W. Cramer (Juni 1995)

No. 12 Wirkung von Klimaveränderungen auf Waldökosysteme
P. Lasch, M. Lindner (August 1995)

No. 13 MOSES - Modellierung und Simulation ökologischer Systeme - Eine Sprachbeschreibung mit Anwendungsbeispielen
V. Wenzel, M. Kücken, M. Flechsig (Dezember 1995)

No. 14 TOYS - Materials to the Brandenburg biosphere model / GAIA
Part 1 - Simple models of the "Climate + Biosphere" system
Yu. Svirezhev (ed.), A. Block, W. v. Bloh, V. Brovkin, A. Ganopolski, V. Petoukhov, V. Razzhevaikin (Januar 1996)

No. 15 Änderung von Hochwassercharakteristiken im Zusammenhang mit Klimaänderungen - Stand der Forschung
A. Bronstert (April 1996)

No. 16 Entwicklung eines Instruments zur Unterstützung der klimapolitischen Entscheidungsfindung
M. Leimbach (Mai 1996)

No. 17 Hochwasser in Deutschland unter Aspekten globaler Veränderungen - Bericht über das DFG-Rundgespräch am 9. Oktober 1995 in Potsdam
A. Bronstert (ed.) (Juni 1996)

No. 18 Integrated modelling of hydrology and water quality in mesoscale watersheds
V. Krysanova, D.-I. Müller-Wohlfeil, A. Becker (Juli 1996)

No. 19 Identification of vulnerable subregions in the Elbe drainage basin under global change impact
V. Krysanova, D.-I. Müller-Wohlfeil, W. Cramer, A. Becker (Juli 1996)

No. 20 Simulation of soil moisture patterns using a topography-based model at different scales
D.-I. Müller-Wohlfeil, W. Lahmer, W. Cramer, V. Krysanova (Juli 1996)

No. 21 International relations and global climate change
D. Sprinz, U. Luterbacher (1st ed. July, 2n ed. December 1996)

No. 22 Modelling the possible impact of climate change on broad-scale vegetation structure - examples from Northern Europe
W. Cramer (August 1996)

No. 23 A methode to estimate the statistical security for cluster separation
F.-W. Gerstengarbe, P.C. Werner (Oktober 1996)

No. 24 Improving the behaviour of forest gap models along drought gradients
H. Bugmann, W. Cramer (Januar 1997)

No. 25 The development of climate scenarios
P.C. Werner, F.-W. Gerstengarbe (Januar 1997)

No. 26 On the Influence of Southern Hemisphere Winds on North Atlantic Deep Water Flow
S. Rahmstorf, M. H. England (Januar 1977)

No. 27 Integrated systems analysis at PIK: A brief epistemology
A. Bronstert, V. Brovkin, M. Krol, M. Lüdeke, G. Petschel-Held, Yu. Svirezhev, V. Wenzel (März 1997)

No. 28 Implementing carbon mitigation measures in the forestry sector - A review
M. Lindner (Mai 1997)

No. 29 Implementation of a Parallel Version of a Regional Climate Model
M. Kücken, U. Schättler (Oktober 1997)

No. 30 Comparing global models of terrestrial net primary productivity (NPP): Overview and key results
W. Cramer, D. W. Kicklighter, A. Bondeau, B. Moore III, G. Churkina, A. Ruimy, A. Schloss, participants of "Potsdam '95" (Oktober 1997)

No. 31 Comparing global models of terrestrial net primary productivity (NPP): Analysis of the seasonal behaviour of NPP, LAI, FPAR along climatic gradients across ecotones
A. Bondeau, J. Kaduk, D. W. Kicklighter, participants of "Potsdam '95" (Oktober 1997)

No. 32 Evaluation of the physiologically-based forest growth model FORSANA
R. Grote, M. Erhard, F. Suckow (November 1997)

No. 33 Modelling the Global Carbon Cycle for the Past and Future Evolution of the Earth System
S. Franck, K. Kossacki, Ch. Bounama (Dezember 1997)

No. 34 Simulation of the global bio-geophysical interactions during the Last Glacial Maximum
C. Kubatzki, M. Claussen (Januar 1998)

No. 35 CLIMBER-2: A climate system model of intermediate complexity. Part I: Model description and performance for present climate
V. Petoukhov, A. Ganopolski, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, S. Rahmstorf (Februar 1998)

No. 36 Geocybernetics: Controlling a rather complex dynamical system under uncertainty
H.-J. Schellnhuber, J. Kropp (Februar 1998)

No. 37 Untersuchung der Auswirkungen erhöhter atmosphärischer CO_2-Konzentrationen auf Weizenbestände des Free-Air Carbondioxid Enrichment (FACE) - Experimentes Maricopa (USA)
Th. Kartschall, S. Grossman, P. Michaelis, F. Wechsung, J. Gräfe, K. Waloszczyk, G. Wechsung, E. Blum, M. Blum (Februar 1998)

No. 38 Die Berücksichtigung natürlicher Störungen in der Vegetationsdynamik verschiedener Klimagebiete
K. Thonicke (Februar 1998)

No. 39 Decadal Variability of the Thermohaline Ocean Circulation
S. Rahmstorf (März 1998)

No. 40 SANA-Project results and PIK contributions
K. Bellmann, M. Erhard, M. Flechsig, R. Grote, F. Suckow (März 1998)

No. 41 Umwelt und Sicherheit: Die Rolle von Umweltschwellenwerten in der empirisch-quantitativen Modellierung
D. F. Sprinz (März 1998)

No. 42 Reversing Course: Germany's Response to the Challenge of Transboundary Air Pollution
D. F. Sprinz, A. Wahl (März 1998)

No. 43 Modellierung des Wasser- und Stofftransportes in großen Einzugsgebieten. Zusammenstellung der Beiträge des Workshops am 15. Dezember 1997 in Potsdam
A. Bronstert, V. Krysanova, A. Schröder, A. Becker, H.-R. Bork (eds.) (April 1998)

No. 44 Capabilities and Limitations of Physically Based Hydrological Modelling on the Hillslope Scale
A. Bronstert (April 1998)

No. 45 Sensitivity Analysis of a Forest Gap Model Concerning Current and Future Climate Variability
P. Lasch, F. Suckow, G. Bürger, M. Lindner (Juli 1998)

No. 46 Wirkung von Klimaveränderungen in mitteleuropäischen Wirtschaftswäldern
M. Lindner (Juli 1998)

No. 47 SPRINT-S: A Parallelization Tool for Experiments with Simulation Models
M. Flechsig (Juli 1998)

No. 48 The Odra/Oder Flood in Summer 1997: Proceedings of the European Expert Meeting in Potsdam, 18 May 1998
A. Bronstert, A. Ghazi, J. Hladny, Z. Kundzewicz, L. Menzel (eds.) (September 1998)

No. 49 Struktur, Aufbau und statistische Programmbibliothek der meteorologischen Datenbank am Potsdam-Institut für Klimafolgenforschung
H. Österle, J. Glauer, M. Denhard (Januar 1999)

No. 50 The complete non-hierarchical cluster analysis
F.-W. Gerstengarbe, P. C. Werner (Januar 1999)