ARIDE

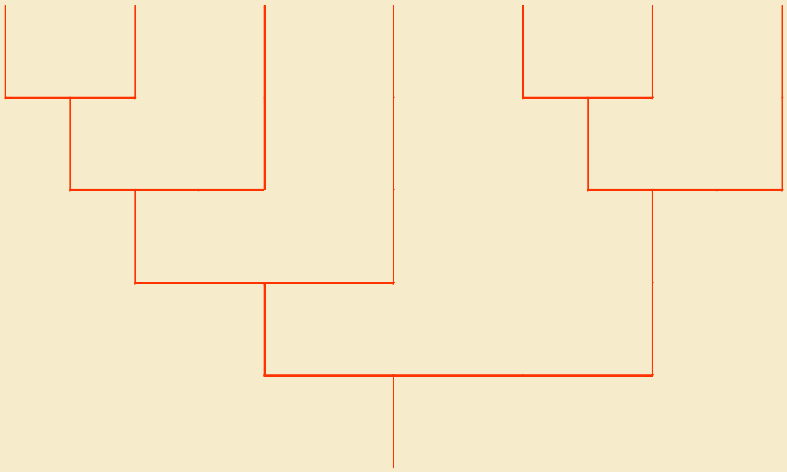
Assessment of the Regional Impact of Droughts in Europe

Technical Report no. 1

November 1999

Methods for Regional Classification of Streamflow Drought Series:

Cluster Analysis



Workpackage 2 Hydro-meteorological Droughts

Activity 2.5 Regional Drought Characteristics

Technical Report to the ARIDE project:

Work Package 2 Hydro-meteorological Drought
Activity 2.5 Atmospheric Circulation and Drought

Methods for Regional Classification of Streamflow Drought Series:

Cluster Analysis

by
Kerstin Stahl & Siegfried Demuth
Institute of Hydrology, University of Freiburg
Germany

Table of Contents

This report is available at: Institute of Hydrology, University of Freiburg,
Fahnenbergplatz 1, 79098 Freiburg, Germany.
demuths@uni-freiburg.de.

1 Objectives

There are two objectives for the application of a statistical classification method to the historic streamflow drought time-series of the EWA stations:

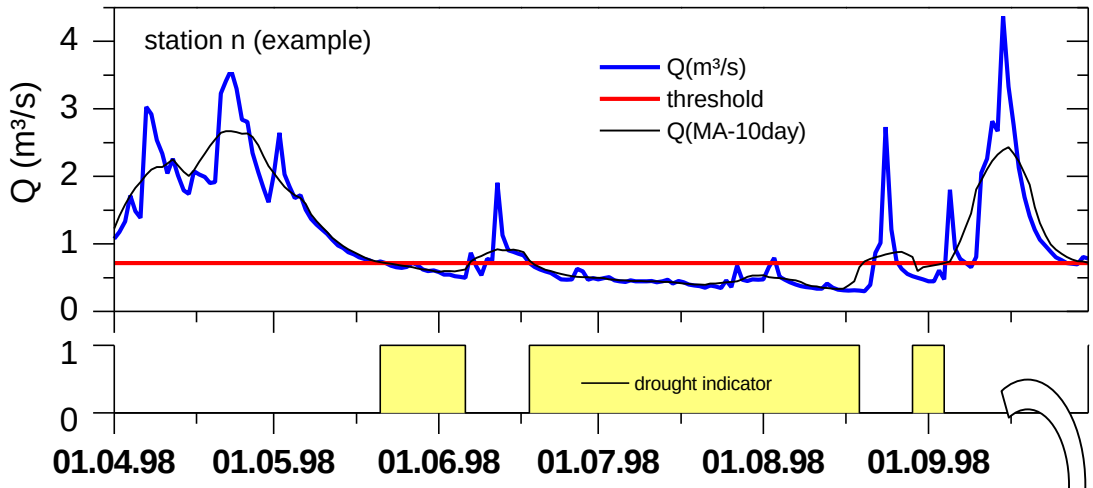
- The first is to investigate the regional patterns of droughts in terms of simultaneous occurrence of streamflow drought at the gauging stations of the EWA.
- The second objective, specific for activity 2.5 'Atmospheric circulation and drought', is to test the performance of the varying threshold, which is assumed to reveal large scale spatial patterns due to the 'anomaly character' of the investigated streamflow parameter (see Annex 3.2, First Annual Report, 1998). Here, anomaly describes a departure from the 'normal' behaviour and consequently a signal, which is relatively independent from the typical annual cycle and partly from typical catchment characteristics. The parameter is therefore expected to describe the streamflow response to synoptic meteorology.

2 Method

2.1 Cluster Analysis - Theory

The goal of a cluster analysis is to group the *variables* of a data matrix in a way that the characteristics of the variables within a group are as homogeneous as possible, but the characteristics of the variables between groups are as contrasting as possible. Therefore, the first and most important step is to define a *measure* for the similarity or dissimilarity (called 'distance') of the characteristics of two variables. Depending on the scaling of the variables being quantitative, binary, or count data, various distance measures are available (most common are e.g. Euclidean Distance, Pearson Correlation, Chi^2, etc.). The choice for a distance measure strongly depends on the purpose of the classification. In the next step, the distance measures for each pair of variables are calculated (distance matrix). Searching this distance matrix, *Hierarchical Clustering* procedures identify relatively homogeneous groups of variables using an algorithm which starts with each variable in a separate cluster and combines clusters until only one is left. Since hierarchical cluster analysis is an exploratory method, results should be treated as tentative until they are confirmed with an independent method (Norisus, 1995).

A) Drought Indicator series (0,1) determination



B) Data Matrix

date	station 1	station 2	station 3	.....	station n
1.1.1961	0	1	0	.....	0
2.1.1961	0	1	1	.....	0
3.1.1961	0	1	1	.....	0
...	...	...	...	.....	...
30.12.90	1	0	0	.....	1
31.12.90	1	0	0	.....	1
	variable 1	variable 2	variable 3	.....	variable n

distance calculation

C) Distance Matrix

	variable 1	variable 2	variable 3	.....	variable n
variable1	0				
variable 2	D(V1,V2)	0			
variable 3	D(V1,V3)	D(V2,V3)	0		
...	D(V1,...)	D(V2,...)	D(V3,...)	0	
variable n	D(V1,Vn)	D(V2,Vn)	D(V3,Vn)	D(...,Vn)	0

D) Hierarchical Clustering

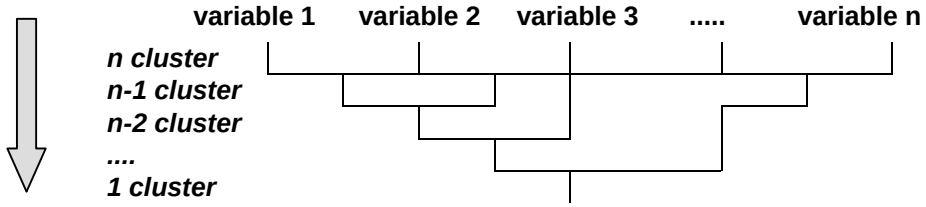


Figure 1 Scheme for the procedure of the cluster analysis application to drought data.

2.2 Cluster Analysis with drought data

The *variables* to be grouped in this case are the gauging stations from the EWA. Those variables contain the daily drought indicator series data (0 = no drought, 1 = drought) over a common time period, derived from the application of the threshold level approach to daily discharge data (Fig.1 A,B). This type of binary variable is chosen to meet the objective of studying the simultaneous drought occurrence.

Binary Euclidean distance was used as *distance measure* and therefore calculated for all pairs of variables (Fig.1C). The Euclidean distance is computed from a fourfold (frequency-) table as SQRT(b+c), where b and c represent the diagonal cells corresponding to the number of cases present on one variable but absent on the other:

variable 1 variable 2	drought	no drought
drought	*a*	*b*
no drought	*c*	*d*

$$D_{bin.Euklid}(Var1, Var2) = \sqrt{b+c}$$

The commonly used *Ward-Method*, which minimises the distances within a cluster, was then used for the hierarchical clustering, schematically shown in Fig.1D. This method produces small, clearly separated clusters of hyperspheric shape (Norius, 1995).

The decision for a certain number of clusters that represent 'homogeneous' groups is not fixed and strongly depends on the objective of the study. Since there is a lack of a consistent definition of 'cluster' and its structure and content, it is difficult to define hypothesis tests (Aldenderfer and Blashfield, 1984). The Statistics software package SPSS, which is used for the calculations, computes several statistics, which assist in reaching a decision:

- The agglomeration schedule contains the variables or clusters combined at each stage, the distances between the variables or clusters being combined, and the last cluster level at which a variable joined the cluster.
  - The distances between the combined clusters (example see Figure 2) can be used to determine the optimum number of clusters, meaning the least possible number with the highest possible homogeneity. This solution is usually determined by a jump in the plot of the distances at each step.
- The Proximity Matrix, which gives the distances or similarities between variables.
  - During validation, the values can be used to check for outliers or can indicate misclassification.

- The Cluster Membership, which displays the cluster to which each variable is assigned at one or more stages in the combination of clusters.
  - The membership table can be imported into a GIS or visualization software together with the geographic co-ordinates of the gauging stations to plot the spatial distribution of several cluster number solutions (example of one solution see Figure 3).

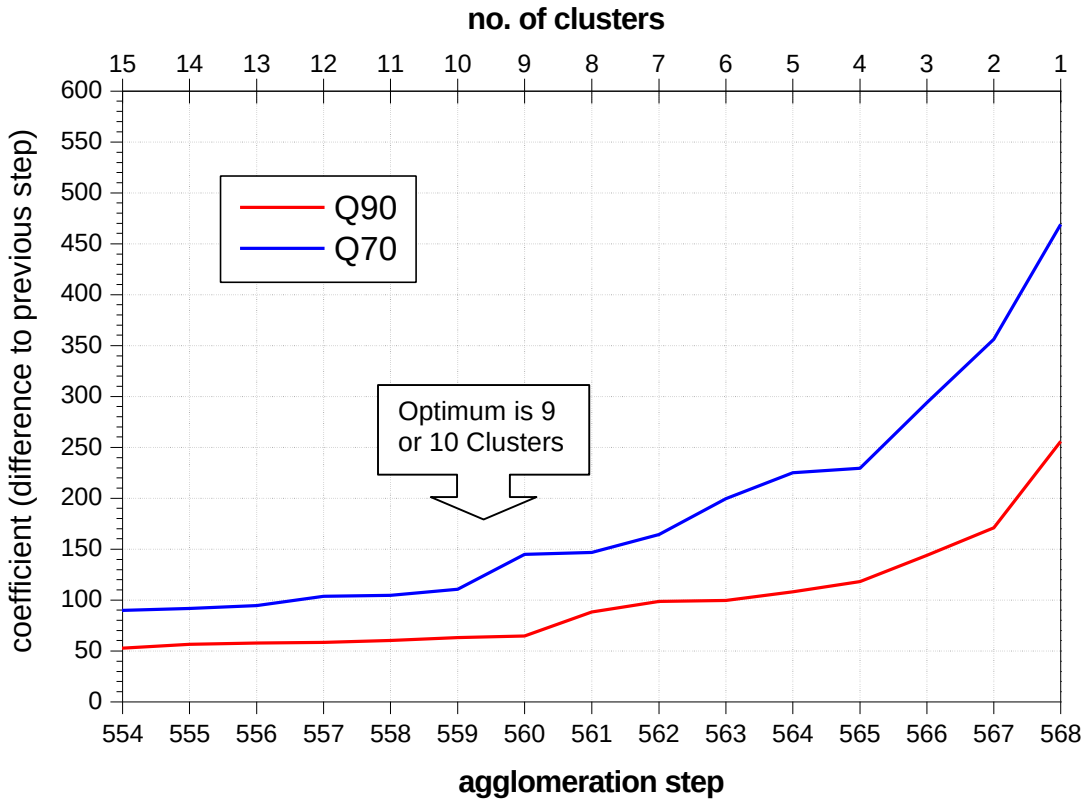


Figure 2 Agglomeration Schedule Example from the Cluster Analysis of the drought indicator series of 569 stations.

3 Verification/Validation

If the objective of the classification exceeds the exploratory stage, the homogeneity of the groups should be tested independently, and each station should be checked for misclassification.

Discriminant analysis is often used after a cluster analysis. It is useful for situations where one wants to build a predictive model of group membership based on observed characteristics of each case. A set of discriminant functions is generated based on linear combinations of the predictor variables that provide the best discrimination between the groups. The functions are generated from a sample of cases for which group membership is known (determined by the cluster analysis) the functions can then be applied to new cases with measurements for the predictor variables but unknown group membership (Norius, 1995).

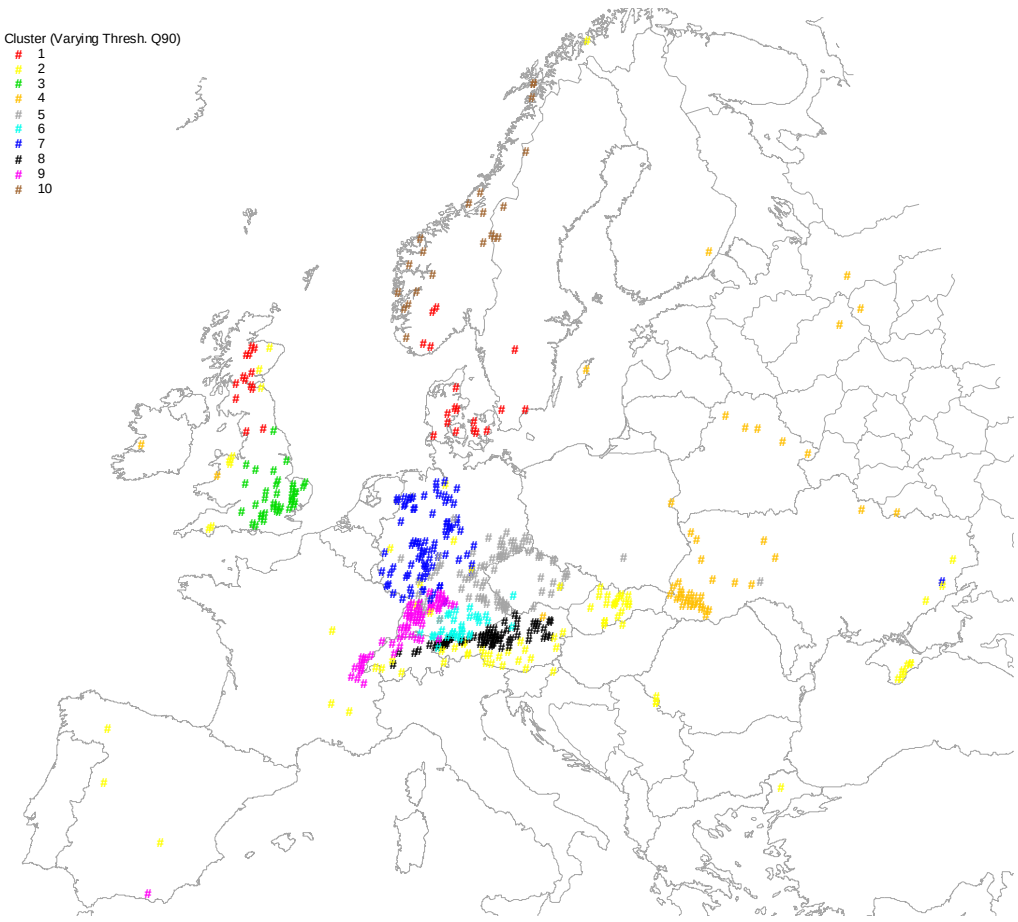


Figure 3: Example for the cluster membership of the stations (10 cluster solution, variable Q90)

4 Discussion

The type of analysis described above has been applied to a preliminary dataset with the time series 1961-90. The results - at the exploratory stage - have been presented at the last meeting in Lisbon. The final results including validation of the application to a revised dataset will be presented at the next meeting and in the annual report '99.

A few characteristics of this type of regional classification method should be emphasised. First of all is the method strictly statistical, based only on the observed discharge data. Relative or absolute spatial location of the stations or basin properties are not taken into account. The method, the type of data and the distance measure are to be considered when interpreting the results: the obtained clusters do not necessarily have to be spatially coherent, as their homogeneity consists of a relatively simultaneous drought occurrence of their members during the entire time period. A single event, however, may show a completely different spatial pattern.

Cluster Analysis in general can be applied to other variable sets to obtain regions, which are homogenous in another respect. In principle, two different types of variables would be interesting to classify:

- time series of different drought parameters (e.g. yearly maxima, like AMS)
  - depending on the distance measure, again the simultaneous behaviour (e.g. *Pearson Correlation Coefficient* measure – for relative fluctuations) would be the objective, or groups with similar durations/volumes of their stations (e.g. *Block Distance* measure – for absolute difference of the values) would be the result.
- variables containing several different characteristics of the station/catchment important to streamflow drought (e.g. parameters describing the frequency distribution of AMS/PDS, longest/severest drought, month of highest drought risk, existence of multi-year droughts, basin properties and climate parameters, geographic location, regime type, etc..)
  - such a classification would then define regions more in the classical way of regionalisation and regional hydrology – geographical regions with a certain type of drought behaviour, which would then have to be described for each cluster.

5 References

ALDENDERFER M.S. & BLASHFIELD, R. K. (1984): Cluster analysis. Quantitative applications in the social sciences: 87 S. Sage Publications. Newbury Park.

NORIUS, M.J.(1995): SPSS for Windows. Professional Statistics, Release 6.1. SPSS Inc. Chicago.