DETERMINATION OF SEASONAL MACROSYNOPTIC TYPES USING CLUSTER ANALYSIS AND ROTATED EOF ANALYSIS

by

Judith Bartholy

CENTRAL METEOROLOGICAL INSTITUTE, BUDAPEST, HUNGARY

Évszakos makroszinoptikus típusrendszerek meghatározása rotált empirikus ortogonális függvényana-analízis és clusteranalízis felhasználásával. A tanulmány rövid összefoglalását adja egy sokéves vizsgálatsorozatnak, melyben kísérlet történt évszakos makrocirkulációs osztályozások létrehozására az atlanti-európai szektorban, illetve az északi hemiszféra térségére. A különböző matematikai statisztikai módszer módszerekkel létrehozott típusrendszerek összehasonlító elemzésére is kitér a dolgozat.

A short summary is presented of a research carried out in the last few years. In this work an attempt was made to establish seasonal macrocirculational classifications for the Atlantic-European region and for the Northern Hemisphere. Comparative examinations of macrosynoptic systems produced by different mathematical statistical methods are discussed too.

1. Introduction

As early as in the 1950s and 1960s climatological and macrosynoptical classifications were made by several researchers for example *DZERDZEJEVSKIJ (1946)*, *VANGENGEJM (BOLOTINSKAJA - 1964)*, *HESS* and *BREZOWSKY (1969)*, *LAMB (1972)* etc., with different aims and for different geographical regions. In each case different criteria were taken as bases of classification (the geographical position of cyclones and anticyclones, the direction of ridges, etc.). A common feature of these early classifications was that both creating the classes and arranging the phenomena were subjective processes and were carried out with the help of the human eye and synoptical practice.

With the widerspread use of high-capacity computers more objective methods have become possible. A new branch of mathematics, cluster analysis deals with the problems of classification algorithms. Some good summaries on this topic are given by *ANDERBERG (1973)*, *SPATH (1980)*. Several attempts have been made in the recent past at meteorological application of clustering procedures: *SULOCHANA (1980)* made a precipitation classification for the region of India, *HARTMAN (1984)* classified the tropical cloud configurations, and *PANAGIOTIS (1984)* gave a classification of weather situations in Greece. *KRUIZINGA (1979)* and *MARYON (1985)* made typisation on the basis of 500 mb height fields.

In the last years we have also applied several kinds of clustering technique in our researches, and created objective macrocirculation systems for large regions. The aim of this work was to eliminate the subjectively coded *HESS-BREZOWSKY* macrosynoptic system from the analogous forecasting model used by the *Central Forecasting Institute of the Meteorological Service of Hungary*. In our first work *[AMBROZY-BARTHOLY-GULYAS (1983)]* we made a system for the Atlantic-European region. With the increased validity period of the forecasts, however, it become necessary to make a hemispheric scale system. So, later on, in our experiments we made attempts at hemispherical clustering of the 500 and 700 mb height level by means of different classification algorithms. Probably due to the high number of dimensions, these classes were not separeted from each other sufficiently. Therefore, prior to the use of the classification algorithms a procedure for feature extraction and for data reduction was carried out on the entire data base: the rotated empirical orthogonal function analysis.

2. Attempts at clustering

Since the meteorological application of clustering is not yet much used, I'm going to give a brief summary of the principles of the procedures we used and their place among the methods of clustering. Cluster analysis, as a rule, has a dual aim: first, to explore the structure of the set of objects; second, to select the separate objects in such a way that the similar ones get into the same group, while the differing ones are placed in different classes.

The cluster analysis method can be divided into two main types: the hierarchical and the non-hierarchical methods. The methods included in the first group constitute a hierarchical system of clusters in such way that any two clusters are either disjoint or one of them implies the other. The hierarchical methods belong either to the agglomerative or the divisive type, depending on whether they reach the hierarchical system of clusters by unification or by division. Non-hierarchical methods can also be divided into two parts, namely, the overlapping and the disjoint classifications, depending on whether the grouping allows overlappings between the cluster or not.

Another important component of cluster analysis, the connection function, is also of two types: similarity function or non-uniformity function, depending on whether they take their maximum values in case of similarity or non-uniformity of the objects. In the past years in our researches hierarchical agglomerative and non-hierarchical disjoint methods are used, and in every case the classification was performed with the use of a connection function of the similarity type.

3. Atlantic-European region

In our first clustering attempt based on the daily 500 mb height fields, we constructed a seasonal system for the Atlantic-European region. Here the separate types represented the macrocirculational situations most characteristic of the season. For determining the classes the dynamical *"k-means"* method worked out by *McQUEEN (1967)* was used which belongs to the non-hierarchical disjoint methods. In the algorithm the computation was carried out with similarity connection functions according to *EUCLIDEAN* metrics. In the iteratively approximating version of method, the stability of the system during the separate iterations was measured by the number of objects regrouped in a new type (non-identical with that of the previous iteration). *Figure 1* illustrates the pressure maps of AT_{500}, $RT_{500/1000}$ and of surface level of the 9^{th} spring type. The first one was obtained as a result of clustering, and the latter two were computed on the basis of the archive of the full time-series. The seasonal cluster system contains 19 spring, 8 summer, 15 autumn and 17 winter types. To measure the effectiveness of the types we considered the external and internal distances of the systems which, by definition, represent the distances between the class centers and the internal radius of the various classes respectively. Internal radius: the average distance of all the fields included in the given type, from the cluster center.

With this classification we succeeded in reducing the internal distances by 42 per cent, and in increasing the external distances on average by 60-70 per cent in comparison with the *HESS-BREZOWSKY* macrosynoptic system. This means a better filling and spanning of the physically given 80 dimensional space (the Atlantic-European region is represented by 80 grid point values). The good results of this systematization were confirmed by the comparative verification results obtained after its insertion in to the long-range forecasting model.

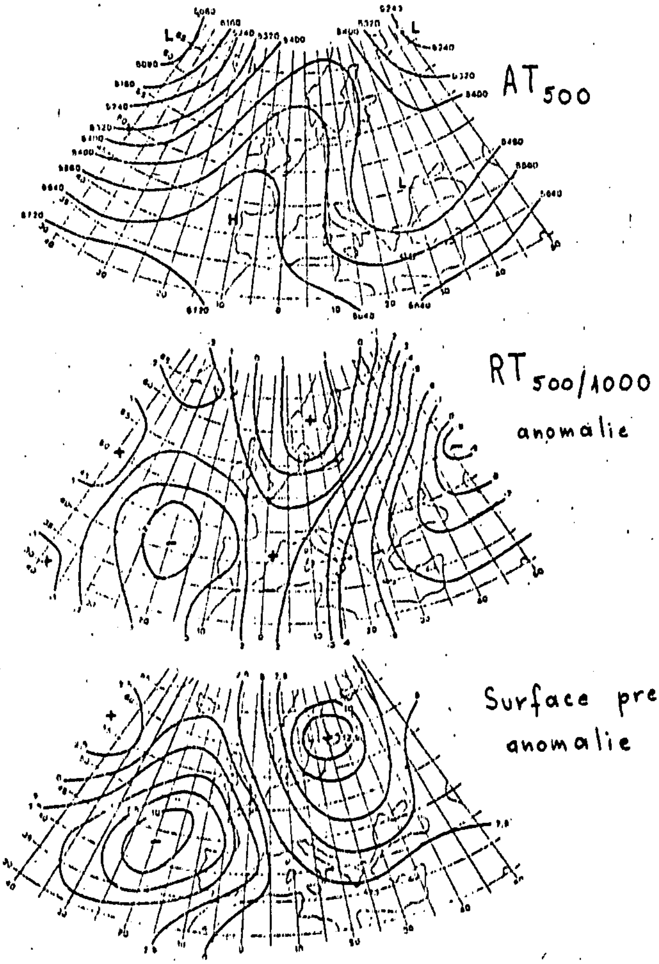

Fig. 1 9th class for spring season of the cluster system for the Atlantic-European region

4. The Northern Hemisphere

In our further experiments the investigated geographic region was expanded to the Northern Hemisphere; for technical reasons we switched over to the 700 mb level (on 500 mb there was a significant lack of data for this region in the data set at our disposal); and we chose a larger time scale: the fields of decade averages instead of daily fields. Like *FOLLAND (1985)* and *MAKICH (1985)* in their classifications for the Atlantic-European region, we also made our hemispherical cluster system for 6 natural seasons (Jan.-Febr., March-Apr., etc.) The low sample size does not cause a problem: as running means of decade fields were used in our researches, there was no reductions of the data set compared to the daily field set; only a smoothing was applied. The analysis was carried out on the data series of a 35 year period (1950-1984) of the *NMC (National Meteorological Center), Washington, D. C.*

Two attempts were made at clustering the running decade fields. In one case the hierarchical agglomerative, in the other case the non-hierarchical disjoint, non dynamical *"k-means"* method was used, similarly to the investigation of the Atlantic-European region. In both cases, for each season there were determined 10 types. In the investigations each hemispheric running means of 700 mb decade field was characterized by 358 grid point values, each representing approximately equal territories. The results satisfied us moderately: the types of the class systems did not separate well, presumably due to the high dimension numbers.

We were confident that carrying out an empirical orthogonal function analysis (EOF) prior to the implementation of the clustering procedures, the reduction in dimension number allowed by the concentration of the information will give better results. A comparative evaluation of the cluster systems will be discussed later.

5. Joint application of the rotated empirical orthogonal function analysis and the clustering

For reasons described above we applied a combined method, in which clustering was done after having carried out the EOF analysis of the fields and varimax rotation of the amplitudes. The EOF analysis was first used by *LORENZ (1956)* and *GILMAN (1957)*. In *CRADDOCK*'s investigations *(1969)* the hemispherical 500 mb fields are represented with the expansion coefficients of the fields expanded by eigenvectors of the correlation matrix, and the coefficients of the field are used as the data base of analog forecasting procedure. So the procedure is applied as a feature extraction and data reduction method. The EOF analysis - besides its several advantages described by *WALLACE (1972)*, *BARNSTON* and *LIVEZEY (1985)*, etc. - was used by us mainly in the *CRADDOCKIAN* sense. In the procedure the eigenvalue equation of the correlation matrix of the standardized data set is solved. The obtained eigenvectors are orthogonal, but their physical interpretation is difficult. Carrying out the varimax rotation analysed in detail among others by *HORAL (1981)* and *HARMAN (1960)*, we obtain results that can be interpreted better physically, at a price of a little deterioration of orthogonality. Inside the separate modes rotation maximalizes variance and eliminates smoothing caused by averaging, and it represents another feature extraction procedure.

After the EOF analysis and the rotation every field of every season is characterized by 10 coefficients and, after carrying out our iterative clustering procedure on these data as the data base, for each season there are obtained 10 types. In the clustering procedure there were made several attempts at selecting the initial cluster centers in order to accelerate the convergency of the method. Naturally, in order to obtain the final

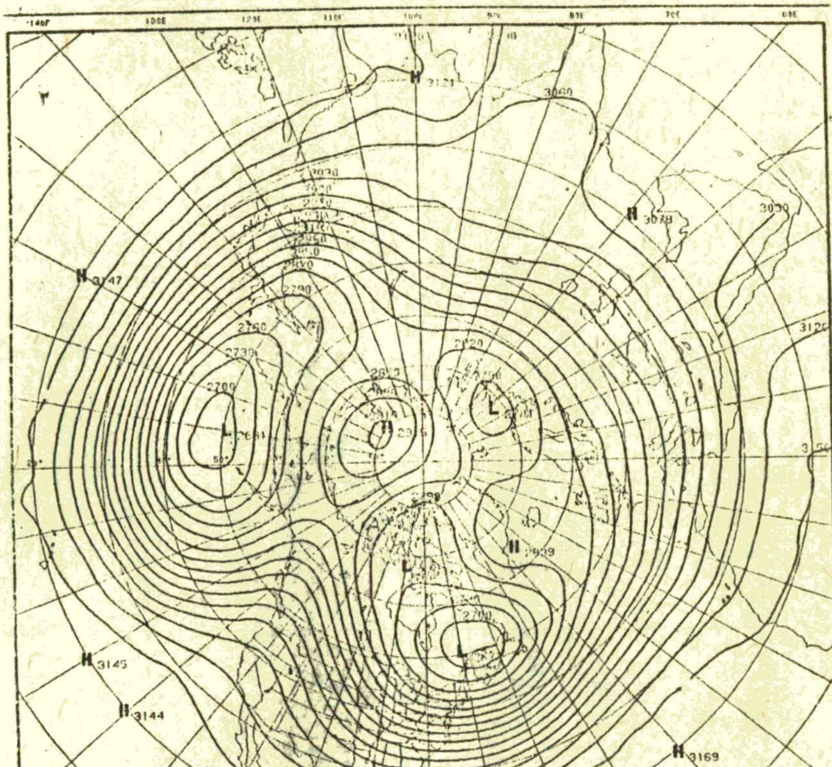

Fig. 2a 7^{th} class for winter season of the northern hemispherical cluster system (the method use rotated EOF analysis and dynamical cluster technics), 700 mb

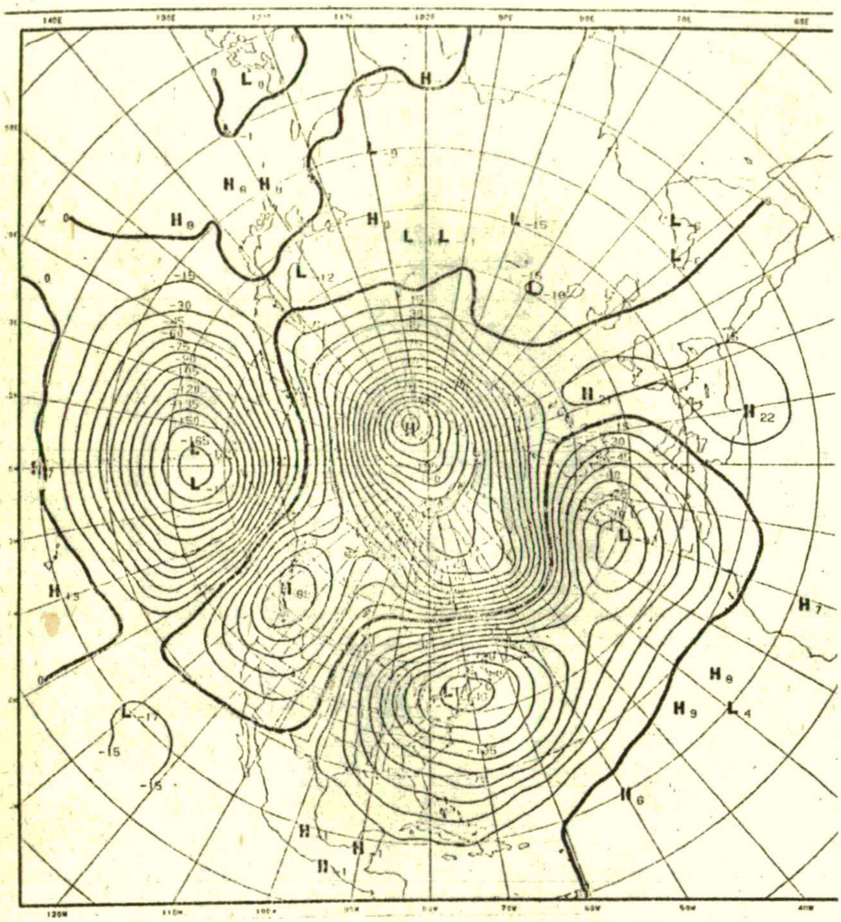

Fig. 2b 7th class for winter season of the northern hemispherical cluster system (the method used rotared EOF analysis and dynamical cluster technics), 700 mb, anomalies class centers, the inverse transformation of the rotation and the restoration of the 700 mb cluster centers have to be accomplished on the basis of the EOF coefficients and the eigenvectors. As an example, *Figure 2* illustrates the 700 mb height and anomaly fields of type *No.*7 of the winter season (January-February).

Comparative evaluation

During the 3 clustering experiments for the Northern Hemisphere and for the Atlantic-European region the stability of classifications was checked separately for each method. The mathematical background of the methods ensures the convergence of the algorithms in the classification procedures used. The invariance of the procedures on the randomly chosen initial cluster centers was checked.

In these methods each iteration sweeps through the whole data set once and every field is placed in a class. So, counting the fields placed in a new class (differing from the previous inclusion) a good index can be obtained for the stability of the cluster system. In *Table I* the final stability indices are demonstrated (for each method uniformly 12 iterations were made). It can be seen that, while in the case of the clustering after the hemispherical rotated EOF analysis, in the last iteration only 0.3 per cent of the fields were regrouped into a new class (altogether 7 fields), in the other two case their number was almost 5-10 times higher.

The external and internal distances of the three hemispherical cluster systems (using hierarchical, dynamical, EOF + dynamical methods) were compared. These results are shown in *Table II*. It can be seen that the procedure applying the EOF analysis compared to the hierarchical method; 1/ ensures a separation between the clusters about 30 per cent better (increase of the external distances); 2/ the different clusters are concentrated around the type centers 22 per cent better (reduction of the internal radius of the classes). These results are mean values, after averaging for 6 seasons.

The relatively little mean internal and great mean external distances can be evaluated positively if it is not accompanied by extreme frequency distribution for each type. But this condition is also fulfilled, because the amount of data included in the separate types is not less than 6 and not more than 17 per cent of the whole data set. The life-time of the separate types is on average two weeks, but naturally, it varies greatly varies depending on the seasons and types.

It was demonstrated that out of the 3 classification methods for the Northern Hemisphere, the system with clustering after a rotated EOF analysis is the best one in all aspects, therefore we are going to use this in our long-range forecasting analog method. Insertation into the model and the verification work is being done.

Table I

Instability indices for the two dynamical and the combined methods (the recorded field numbers by the last iteration given in percentage)

	Atl. - Eur. region	Northern Hemisphere	
	methods		
	Dynamical	Dynamical	EOF + Dynamical
Instability indices	2.5 %	1.3 %	0.3 %

Table II

Separation indices between the classes and concentration indices inside the classes (changes of external and internal distances in percentage), calculated for each pars of the 3 methods

	methods		
	Hierarch. - Dynamical	Dynamical - EOF + Dynamical	Hierarch. - EOF + Dynamical
Increasing of the external distances	11 %	18 %	29 %
Decreasing of the internal distances	8 %	14 %	22 %

References

AMBROZY, P.-BARTHOLY, J.-GULYAS, O., (1984): A system of seasonal macrocirculation pattern for the Atlantic- European region, Idöjárás, Vol. 88. No.3., pp.121-133.

ANDERBERG, M.R., (1973): Clusteranalysis for applications, Academic Press, New York.

BARNSTON, A.G.-LIVEZEY, R.E., (1985): High resolution rotated EOF analysis of northern hemisphere 700 mb heights for predictive purposes, "9. Conf. on Prob. and Stat. in Atm. Sci." Virginia Beach, Am.Met.Soc., pp.290-298.

BARTHOLY, J.-KABA, M., (1985): Further development of seasonal climate forecasts for the territory of Hungary, Idöjárás, Vol. 89. No.4., p.185.

CRADDOCK, J.M.-FLOOD, C.R., (1969): Eigenvectors for representing the 500 mb geopot. surface over the northern hemisphere, Quart.J.R. Met.Soc. No.405., p.576.

DZERDZEJEVSKIJ, B.L., (1946): Typification of circulation mechanism in northern hem. and characteristics of synoptic seasons, Moscow, Gidrometeoizdat.

FOLLAND, C.K.-COLMAN, A., (1985): A multivariate tech. for use in long-range forecasting, Proc.1.WMO Conf. Diag. Pred. Monthly-Seasonal Atm. Variations over Globe, Maryland.

GILMAN, D.L., (1957): EOF-s applied to thity-day forecasting, Sci.Rep. No.*1*. Contract AF19 (604)-1283, Dept. of Met. M.I.T. Cambridge, Mass.

HARMAN, H.H., (1960): Modern factor analysis, The University of Chicago Press, London.

HARTMAN, D.L., (1980): Some implication of the mesoscale circ. in trop. cloud clusters for large-scale dyn. and clim. - J. of the Atm. Sci., Vol.*41*. No.1., pp.113-121.

HESS, P.-BREZOWSKY, H., (1969): Katalog der Grosswetterlagen Europas, Ber. Wetterd. Offenbach, No.*113*.

HOREL, J.D., (1981): A rot. principal comp. anal. of the interannual variability of the Northern Hem. 500 mb height field, Mon.W.Rev. Vol.*109*., pp.2080-2092.

KRUIZINGA, S., (1979): Objective classification of daily 500 mb patterns, Bull. of six Conf. on Prob. and Stat. in Atmospheric Science of Amer. Met. Soc.

LAMB, H.H., (1972): British Isles weather types and a register of the daily sequence of circulation patterns, Geophys. Mem., London 16. No.*116*.

LORENZ, E.N., (1956): EOF-s statistical weather prediction, Sci. Rep. No.*1*. Stat. Forecasting Project, Dept. of Met. M.I.T. Cambridge, Mass.

MACQUEEN,J., (1967): Some methods for classification and analysis of multivariate observations, Proc. of Fifth Berkeley Symp. on Prob. and Stat.

MARYON, R.H.-STOREY, A.M., (1985): A multivariate statistical model, for forecasting anomalies on half-monthly mean surf. pressure, Jour. of Clim. Vol.*5*., pp.561-578.

PANAGIOTIS, M., (1984): Weather-type classification by factor anal., Jour. of Clim. Vol.4., p.437.

SPATH, H., (1980): Cluster analysis algorithms for data reduction and classification of objects, John Wiley and Sons, New York.

SULOCHANA, G., (1980): Cluster analysis of rainfall stations of the Indian peninsula, Quart.J.R.Met.Soc. *106*., pp.873-886.

WALLACE, J.M., (1972): Empirical orthogonal representation of time series in the frequency domain, Jour.Appl.Met. Vol.*11*. No.6., pp.887-900.