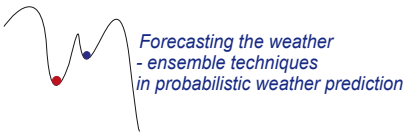
EMS Annual Meeting Abstracts
Vol. 8, EMS2011-738, 2011
11th EMS / 10th ECAM



Random medoid classification of circulation patterns - a benchmark method for common classification algorithms

A. Philipp and C. Beck

University of Augsburg, Geography, Physical Geography and Quantitative Methods, Augsburg, Germany
(andreas.philipp@geo.uni-augsburg.de)

Classification of circulations patterns is a basic tool for synoptic climatology. It allows a simplified view on the complex co- and interaction of processes linking circulation to surface climate variability. However patterns classification may be accomplished by many different methods. Within COST action 733 "Harmonisation and Applications of Weather Type Classifications for European regions" more than 20 classification algorithms have been compared recently. Those methods comprise threshold based algorithms, those based on principal component analysis, based on the leader algorithm and optimisation methods like non-hierarchical cluster analysis or self organising feature maps. It turned out that none of the methods under examination is superior in all aspects concerning explanation (or stratification) of surface climate states (e.g. air temperature or precipitation). Further on it has been found that, even though the mean type patterns of many classifications show some similarity, the resulting partitioning is quite different between the methods.

These results raise the question whether the choice for one of those commonly used classification methods matters at all or in other words: Are the methods able to detect natural groups of circulation patterns which are related to distinct states of surface climate parameters?
If those groups or clusters exist, methods which are designed to find them, in particular cluster analysis, should be be superior to classification schemes which define the types arbitrarily.

In order to prove this assumption, a method called "random medoid classification" has been designed, for each class choosing one single circulation pattern by random (i.e. using a random number generator) and assigning all remaining patterns to them according to their maximum similarity, which might be measured by the minimum Euclidean distance or other similarity metrics. Evaluation statistics like the explained cluster variance for pressure, temperature and precipitation are calculated in order to compare those pseudo random classifications to classifications provided by the cost733cat dataset including many different classification catalogs for various methods. By running the random medoid method 1000 times the empirical probability density function of the evaluation metrics can be established which in turn can be used for intercomparison to the established deliberate methods.
The results show that most of the classifications fail to succeed the 95th percentile of the empirical probability density functions indicating that the respective methods cannot find intrinsic groups of similar and distinctively meaningful - with respect to surface climate - patterns in the data. However, since some few methods do show a tendency for being significantly better - at least in some aspects - it is concluded that any structure in the data cannot be excluded and might be detected.