Global NEST Journal, Vol 10, No 3, pp 359-365, 2008

ASSESSMENT OF CLUSTERING ALGORITHMS IN DISCRIMINATING EUTROPHIC LEVELS IN COASTAL WATERS

I. PRIMPAS
M. KARYDIS*
G. TSIRTSIS

Department of Marine Sciences
University of the Aegean
Mytilini, GR-81100, Greece

Received: 18/10/07
Accepted: 14/12/07

*to whom all correspondence should be addressed:
e-mail: mkar@aegean.gr

ABSTRACT
Cluster analysis has been used widely as a tool for assessing eutrophic trends in coastal waters. The efficiency of clustering in discriminating between oligotrophic, mesotrophic and eutropic sites, depends on the variables used, the distance measure and the clustering algorithm applied. In the present work seven clustering algorithms were evaluated using sets of data from sampling sites of known water type. The results showed that only the Ward's algorithm had high resolution in discriminating sampling sites of different trophic status. The remaining clustering algorithms did not show remarkable resolution in classifying different water types. The use of the Ward clustering algorithm is recommended in eutrophication studies where discrete clusters of oligotrophic, mesotrophic and eutrophic water type are under investigation.

KEYWORDS: Eutrophication, Eutrophication assessment, Cluster analysis, Marine Pollution.

1. INTRODUCTION

Seawater classification into eutrophic, mesotrophic and oligotrophic water types has been approached in a number of ways: use of ecological indices (Wilm and Dorris, 1968; Mihnea, 1985; Karydis and Tsirtsis, 1996; Tsirtsis and Karydis, 1998), univariate statistical methods (Ignatiades *et al.*, 1992; Giovanardi and Tromellini, 1992), spatial analysis (Kitsiou and Karydis, 1998; Kitsiou and Karydis, 2000) and multi criteria choice methods (Moriki and Karydis, 1994; Kitsiou and Karydis, 2002) are among the most commonly used quantitative techniques of data analysis (Karydis, 2001). Multivariate techniques seem to be the most robust as they involve a number of variables related to eutrpohication: phosphate, nitrate, nitrite, ammonia and chlorophyllα concentrations, phytoplankton cell number and ecological indices, are the most important variables characterizing eutrophication (Karydis, 1996). Eutrophication assessment based on multivariate techniques is aiming at: (a) scaling levels of eutrophication and (b) discriminating between sites characterized by different trophic conditions. Cluster analysis is among the multivariate techniques used to reveal discrete trophic levels. The discriminant efficiency of this method is depended on three factors: (a) the variables used (b) the choice of the distance (similarity) measure and (c) the clustering algorithm applied. Although there is published work on both, eutrophication variables (Vollenweider, 1992) and distance measures (Sneath and Sokal, 1973; Everitt, 1981), there is no work published on the choice of the optimal algorithm so as to maximize the resolution among sampling sites. In the present work a number of clustering algorithms is evaluated and their efficiency in forming discrete clusters characteristic of different eutrophic status is assessed.

2. METHODS

Data sets: Three sets of data were used in this research to evaluate the clustering algorithm: two data sets from Saronicos Gulf, Greece and one data set from Rhodes, Greece. The Inshore Gulf water of Saronicos Gulf (data set Type A, 143obs) has already been characterized as eutrophic (Ignatiades, 1992) whereas, the second data set from Saronicos Gulf (Type B, 393 obs) as mesotrophic. The data collected from Rhodes (Type C, 112 obs) characterized oligotrophic conditions (Ignatiades, 1992, Vounatsou and Karydis, 1991).

Selection of variables: Phosphate, nitrate, nitrite, ammonia and Chla concentrations were the variables used in the present work. These are the most characteristic variables in eutrophication studies (Karydis 1996).

Matrix formation: Extreme values were removed using the box-and-whisker plot (Karydis 1994) and the mean value of each variable was calculated in every station. The data were standardized (Pielou 1984) by centering each data point and dividing by the standard deviation:

$$z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_i}$$

Selection of the distance measure: As a distance measure the Absolute Distance (AD) was used:

$$D_{ij} = \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|$$

The AD index was selected because it places less emphasis on larger differences and therefore, was expected to show a better resolution in mesotrohic conditions which are the most important from the management point of view (Karydis, 1996)

Clustering algorithms: Seven clustering algorithms, the most commonly used in pollution and eutrophication studies, were evaluated, concerning their efficiency in producing discrete clusters:

(a) **Centroid clustering**: In the centroid method each group is replaced by an average subject which is the centroid of that group (Sharma, 1996)

(b) **Farthest neighbor clustering**: In the farthest neighbor clustering (also known as complete linkage clustering) the distance between two clusters is defined as the maximum distance between a point in the cluster and a point in the others (Pielou, 1984).

(c) **Group average clustering**: In the group average distance (also known as between groups clustering) the distance between two clusters is obtained by taking the average distance between all pairs of subjects in the two clusters (Sharma, 1996)

(d) **Median clustering**: This is a variation of group average clustering which uses the median distance and appears to be more outlier-proof than the average distance (D'Andrande, 1978)

(e) **Nearest-neighbor clustering**: In the nearest-neighbor clustering also known as single-linkage the distance between two clusters is taken to be the distance separating the closest pair of points such that one is in the cluster and the other in the other (Pielou, 1984)

(f) **Ward clustering**: The Ward method does not compute distances between clusters but rather forms clusters by maximizing the within clusters homogeneity. In the Ward's clustering the within group (i.e. within cluster) sum of squares is used as a measure of homogeneity. This way the method minimizes the total within groups (or within clusters) sum of squares. Clusters are formed at each step such that the resulting cluster solution has the fewest within sums of squares (Sharma, 1996).

(g) **Within groups clustering**: In this method known also as "*average linkage clustering*" there are three ways of measuring intercluster distance: unweighed average distance, weighted and centroid distance. Irrespective of the particular procedure there is a common clustering rule: at every step of such a process, the pair of clusters which is separated by the smallest distance is united.

3. RESULTS

The mean nutrient and Chlα concentration values are given in Table 1. It is observed that the highest values appear in Stations S1 & S2 and the lowest concentrations in stations R1, R2, R3, R4 & R5. The tree-diagrams of the sampling sites based on the nearest neighbor, farthest neighbor and median clustering algorithms are given in Figure 1. It is observed that the three clustering algorithms form three discrete clusters for oligotrophic, mesotrophic and eutrophic stations respectively. However, in all three cases cluster merging takes place at a low distance, a fact that indicates poor resolving power of the algorithms used for discriminating among the three trophic conditions. Figure 2 shows the tree diagrams using the group average and the centroid clustering algorithms. A similar trend shown in Figure 2 is also observed: classification of the stations into eutrophic, mesotrophic and oligotrophic groups but at low resolving power. The within group average likage and the Ward's clustering is shown in Figure 3. Only the Ward's algorithm showed a good discriminanat efficiency since the merging of the eutrophic and mesotrophic groups takes place approximately at 40% of the tolal distance. It is also observed that the resulting eutrophic / mesotrophic group merges with the oligotrophic group at 85% of the total distance.

Table 1. Means of nutrients and chlorophyll *α* concentrations of the eutrophic, mesotrophic and oligotrophic sampling sites (stations).

a) Eutrophic System (E)

	S1	S2
PO_4	0,19	0,17
NO_3	0,33	0,33
NO_2	0,12	0,13
NH_3	1,49	1,12
Chl-a	0,32	0,38
Mean	**0,49**	**0,43**

b) Mesotrophic System (M)

	S3	S4	S5	S6	S7	S8	S9
PO_4	0,10	0,09	0,07	0,10	0,07	0,09	0,09
NO_3	0,31	0,30	0,34	0,28	0,26	0,28	0,30
NO_2	0,09	0,08	0,11	0,05	0,10	0,08	0,07
NH_3	1,04	0,96	0,93	0,87	1,16	1,06	0,80
Chl-a	0,27	0,42	0,24	0,26	0,36	0,30	0,28
Mean	**0,36**	**0,37**	**0,34**	**0,31**	**0,39**	**0,36**	**0,31**

c) Oligotrophic System (O)

	R1	R2	R3	R4	R5
PO_4	0,03	0,02	0,03	0,03	0,03
NO_3	0,26	0,19	0,22	0,25	0,23
NO_2	0,03	0,03	0,03	0,02	0,03
NH_3	0,42	0,35	0,45	0,53	0,47
Chl-a	0,09	0,09	0,09	0,09	0,09
Mean	**0,16**	**0,14**	**0,16**	**0,18**	**0,17**

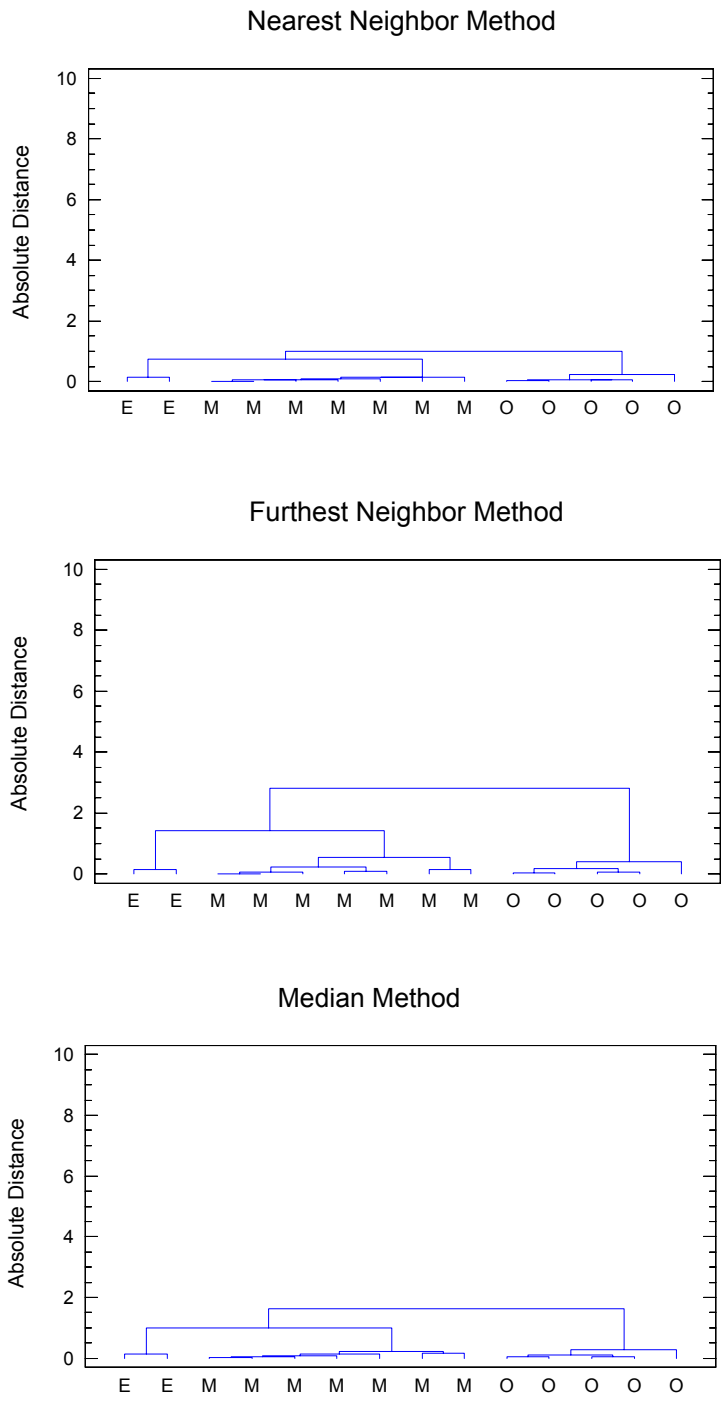

Figure 1. Classification of Eutrohic (E), Mesotrophic (M) and Oligotrophic (O) sampling sites (stations) based on the nearest neighbor, fartherst neighbor and median clustering algorithms

Table 2 shows the ranked distances of group merging for the different clustering algorithms. Maximum resolution is achieved by the Ward's algorithm whereas, the farthest-neighbor and group average algorithms account for about 50% of the resolving power, in discriminating eutrophic levels.

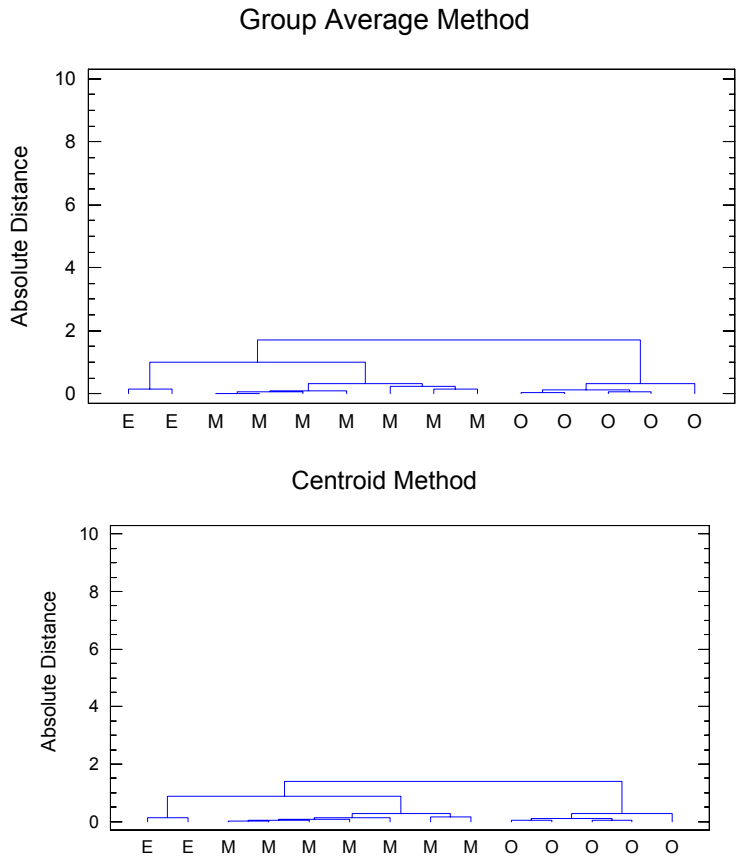

Figure 2. Classification of Eutrohic (E), Mesotrophic (M) and Oligotrophic (O) sampling sites (stations) based on the group average and the centroid clustering algorithms

Table 2. Ranked distances of group merging for the different clustering algorithms

Clustering Algorithm	Eutrophic – Mesotr. (E-M) Group	(E-M) - Oligotrophic Group
Ward	2,49	6,95
Furthest Neighbor	1,41	2,81
Group Average	1,00	1,70
Median	0,98	1,62
Centroid	0,87	1,39
Nearest Neighbor	0,74	1,00
Within Groups	0,53	1,07

4. DISCUSSION

Cluster analysis is a powerful and widely used technique in marine pollution studies (Sneath and Sokal, 1973; Whittaker, 1978; Everitt, 1981) and the assessment of eutrophication (Karydis, 1996); it accomplishes the sorting of sampling sites into clusters based on the similarity to one another. The more similar the sampling unit of each cluster, the more distinct the differences from other clusters are. In addition, using this technique the results are summarized in the form of a tree-diagram which is an illustrative way to understand the classification scheme of the sampling sites. Apart from the seven clustering algorithms evaluated in the present work there are many other, less

known methods, have been used for special purposes; details about them can be found in literature (Sokal and Sneath, 1973, Whitaker, 1978, Everitt, 1981). However, the methods tested in the present work are the most popular and widely used; they have been applied to meet the needs of marine scientists, as tools to investigate pollution studies.

The evaluation of classification techniques, proceeds from mathematical reasoning and from empirical tests with data sets. As no classification technique is perfect, the efficiency of a given method is judged empirically in relation to others, taking into account the "a priori" status of the areas (Stefanou *et al.*, 2000), the robustness and the effectiveness of the results as well as the appropriateness of the method: (a) robustness is an important feature in classification and assumes that the form of the tree diagram remains stable, despite minor changes in the sets of data (b) the effectiveness of the results means that a classification technique aids to understand the problem (Gauch, 1989) and (c) the appropriateness of the results means that the methods reveal the desirable features posed by the particular approach: in the present work discrete clusters are required when water types are classified according to their nutrient levels.

Among the clustering algorithms tested, the nearest-neighbor and farthest-neighbor clustering methods, are rarely used nowadays (Pielou, 1984). As the criterion of the two clusters to be united is the distance between individual points, the cluster derived form merging of the former clusters is represented only by one of its points. The representative point is a rather extreme one than a typical point of the cluster it represents. In addition, the nearest-neighbor clustering is prone to chaining. Chaining is the tendency for early formed clusters to grow by accretion to them on single points and therefore, it does not seem to be the right algorithm when discrete clusters are required. Chaining was also observed in the present work in the tree-diagram formed by the nearest-neighbor algorithm and the overall performance was the poorest.

The main drawback in the group average clustering is that intercluster distances are "fuzzy". It has been reported (Pielou, 1984) that "the device of using the average of all the interpoint distances between two clusters as a measure of inter-cluster distance" is just that of a device". In any case this clustering algorithm was not found efficient in the present study, the algorithm's rank being the sixth out of seven clustering algorithms assessed. The centroid clustering although it showed a strong point which is that the distance between two clusters is the distance between their centroids that is exactly specifiable points, it can happen that this clustering procedure is not always "monotonic". That means "reversals" during cluster merging can occur (Pielou, 1984). In the present work, although clustering with the centroid algorithm was monotonic, the overall efficiency was not satisfactory.

Ward's method tends to give clusters of fairly equal size. Small clusters acquire new members faster than large one's and therefore "chaining" is rather unlikely to happen. This is a great advantage in the present case where discrete clusters are required. In addition, with the Ward's algorithm, it is possible to carry out a statistical test for each margin to find whether the cluster being united is homogeneous (Vassiliou *et al.*, 1989) This is equivalent to judging objectively the information value of each node (Pielou, 1984). Concluding, the Ward's clustering algorithm is recommended as the most suitable in pollution marine studies and the assessment of eutrophication when classification of coastal waters in eutrophic, mesotrophic and oligotrophic water types is required.

ACKNOWLEDGMENT

This work was co-funded by the European Social Fund and National Resources (EPEAEK II) PYTHAGORAS

REFERENCES

Clarke K.R., (1993) Non-parametric multivariate analyses of changes in community structure, *Australian Journal of Ecology*, **18**, 117-143

Everitt B., (1981) Cluster analysis, 2nd Ed. Heineman Educational Books, London.

Field J.G., Clarke K.R. and Warwick R.M. (1982) A practical strategy for analyzing multispecies distribution patterns, *Marine Ecology – Progress Series*, **8**, 37-52

Giovanardi F. and Tromellini, E. (1992) Statistical assessment of trophic conditions. Application of the OECD methodology to the marine environment, in *Marine Coastal Eutrophication*, by Vollenweider R.A., Marchetti R. and Viviani R. (eds), Elsevier, London, pp. 211-233

Ignatiades L., Karydis M. and Vounatsou P. (1992) A possible method for evaluating Oligotrophy and Eutrophication based on nutrient concentrations, *Marine Pollution Bulletin*, **24**(5), 238-243

Karydis M. and Coccossis H. (1990) Use of multiple criteria for eutrohication assessment of coastal waters, *Environmental Monitoring and Assessment*, **14**, 89-100

Karydis M., (2001) Assessing levels of eutrophication: a short review on quantitative methodology, *Biologia Gallo Hellenica*, **27**, 135-144

Karydis M. (1992) Scaling methods in assessing environmental quality: a methodological approach to eutrophication, *Environmental Monitoring and Assessment*, **22**, 123-136

Karydis M., (1994) Environmental quality assessment based on the analysis of extreme values: a practical approach for evaluating eutrophication. *Journal of Environmental Science and Health. Part A. Environmental Science and engineering & toxic and hazardous substance control*, **29**, 775-791

Karydis M. (1996) Quantitative assessment of eutrophication : a scoring system for characterising water quality in coastal marine ecosystems, *Environmental Monitoring and Assessment*, **41**, 233-246

Karydis M. and Tsirtsis G. (1996) Ecological indices: a biometric approach for assessing eutrophication levels in the marine environment, *The Science of the Total Environment*, **186**, 209-219

Kitsiou D. and Karydis M. (1998) Development of categorical mapping for quantitative assessment of eutrophication, *Journal of Coastal Concervation*, **4**, 33-44

Kitsiou D. and Karydis M. (2000) Categorical mapping of marine eutrophication based on ecological indices, *The Science of Total Environment*, **255**, 113-127

Kitsiou D. and Karydis M. (2002) Multi- dimensional evaluation and ranking of coastal areas using GIS and multiple criteria choice methods, *The Science of Total Environment*, **284**, 1-17

Mihnea P.E. (1985) Phytoplankton diversity indices as eutrophication indicators of the Romanian inshore waters, Cercerati Marine, *I.R.C.M.*, **18**, 139-155

Moriki, A. and Karydis, M., (1994) Application of multi criteria choice methods in assessing eutrophication, *Environmental Monitoring and Assessment*, **33**, 1-18

Pielou E.C. (1984) The Interpretation of Ecological Data. J. Wiley & Sons, 263pp.

Sharma, S. (1996) Applied multivariate techniques. J. Wiley & Sons, 493pp

Sneath P.H.A. and Sokal R.R. (1973) Numerical taxonomy. W.H. Freeman & Co., San Francisco.

Stefanou P., Tsirtsis G. and Karydis M., (2000) Nutrient scaling for assessing eutrohication: the development of a simulated normal distribution, *Ecological Applications*, **10**, 303-309.

Tsirtsis G. and Karydis M. (1998) Evaluation of phytoplankton community indices for detecting eutrophic trends in the marine environment, *Environmental Monitoring and Assessment*, **50**, 255-269

Vassiliou A., Ignatiades L. and Karydis M. (1989) Clustering of transect phytoplankton collections with a quick randomization algorithm, *Journal of Experimental Marine Biology and Ecology*, **130**, 135-145.

Vollenweider R.A., Marchetti R. and Viviani R., eds. (1992) Marine Coastal Eutrophication, Elsevier, London.

Vounatsou P. and Karydis M., (1991) Environmental characteristics in oligotrophic waters. Data evaluation and statistical limitations in water quality studies, *Environmental Monitoring and Assessment*, **18**, 211-220.

Whittaker R.H. (1978) Classification of Plant Communities. W. Junk, The Hague.

Wilm L. and Dorris G.T. (1968) Biological parameters for water quality criteria, *BioScience*, **18**, 477-481.