Global NEST Journal, Vol 20, No 1, pp 122-127

A GIS methodology to support sea water quality assessment in coastal areas

Kitsiou D.*, Patera A. and Kostopoulou M.
Lab. of Environmental Quality and Geospatial Applications, Department of Marine Sciences, School of the Environment, University of the Aegean, University Hill, GR 81100, Mytilene, Greece
Received: 13/06/2017, Accepted: 14/12/2017, Available online: 25/01/2018
*to whom all correspondence should be addressed: e-mail: dkit@aegean.gr

Abstract

The development of methodologies for assessing water quality in coastal areas including mapping of eutrophication levels is a research area of high interest. A wide range of methodological approaches can be found in the literature, including multivariate techniques, since marine eutrophication is a multi-parametric phenomenon. In this context, statistical analysis and in particular Principal Component Analysis (PCA) have been widely applied. However, no attempt has been presented so far for mapping eutrophication levels based on information acquired from PCA results in integration with spatial analysis methods. The rapid development of Geographical Information Systems provides the appropriate framework for the development and application of methodologies integrating statistical analysis, spatial analysis methods and mapping techniques. This paper proposes such a methodological approach for assessing sea water quality in coastal areas. The methodology is clearly described and the Strait of Mytilene at the east of the Island of Lesvos in the NE Aegean Sea, Greece is used as a case study.

Keywords: coastal areas, spatial analysis, interpolation, mapping, PCA, eutrophication, Aegean Sea

1. Introduction

Marine eutrophication is a natural process where excessive algal growth is observed due to nutrient supply to marine ecosystems. Physical factors such as geomorphology, bathymetry and existing currents in the study area affect this procedure (de Jonge *et al.*, 2002). Much research has been carried out so far on marine eutrophication leading to important conclusions about the biogeochemical processes related to this phenomenon (Kitsiou and Karydis, 2011). Due to its multi-dimensional nature, most assessment methods integrate physico-chemical and biological indicators based on a variety of variables such as chlorophyll-a, dissolved oxygen, nutrient concentrations or ecological indices (Washington, 1984; Ferreira *et al.*, 2011).

Multivariate techniques such as Principal Component Analysis (PCA), Cluster Analysis, Multivariate Dimensional Scaling, Discriminant Analysis and Factor Analysis have been widely used in sea water quality studies and in particular, in eutrophication assessment (Kitsiou and Karydis, 2001; Caruso *et al.*, 2010).

PCA has been applied in exploratory data analysis for assessing water quality of rivers and coastal areas and quantifying anthropogenic impacts (Vega *et al.*, 1998; Akbal *et al.*, 2011) by using a number of variables such as nutrients, oxygen concentration, BOD, and COD, as well as to predict sources of Dissolved Organic Nitrogen to estuaries (Osburn *et al.*, 2016). Development of multi-metric indices based on PCA for assessing eutrophication has been also performed (Primpas *et al.*, 2010). PCA has been proved useful as well for the detection of regime shifts in marine ecosystems by identifying coherent patterns of variability among a number of timeseries (Andersen *et al.*, 2008). Nevertheless, in the bibliography, there is not any attempt so far of integrating PCA with spatial analysis methods for mapping eutrophication levels in coastal areas.

Spatial analysis methods are the means for describing and analyzing the spatial structure and heterogeneity in datasets and the Geographical Information Systems (GIS) provide the appropriate framework for their implementation (Huang *et al.*, 2001) since they can store, manage, process and analyze spatially referenced data (Burrough and McDonnell, 2000). GIS has been widely used for the development of accurate spatial databases and maps as well as the processing and analysis of environmental data sets, while their rapid development during the last decades, has established them as the most relevant tool for supporting decision-making in coastal areas (Kitsiou *et al.*, 2002).

In this paper, a methodological approach for assessing and mapping water quality in coastal areas is presented based on PCA and spatial analysis methods implemented in the framework of a GIS. The dataset used included data of Dissolved Nitrogen ($N\text{-}NO_3$ and $N\text{-}NH_4$) and chlorophyll-a (chl-a) concentrations measured in seawater samples collected during sampling surveys in the coastal area of the Strait of Mytilene (Lesvos Island, NE Aegean Sea) in December 2007. The advantages of and the need for such methodological approaches are presented and discussed.

2. Materials and methods

2.1. Study area and dataset

The study area known as the Strait of Mytilene is located in the southeastern part of Lesvos Island (NE Aegean). Urban, suburban areas and small settlements are deployed along the coastline accounting approximately for the 45% of the total population of the island. Mytilene with 36,000 inhabitants is the capital of the island and the main center for any economic, business and administrative activities. In the early 1990's, hydrographic surveys carried out to study dissolved oxygen and nutrient distributions in the study area revealed impacts by discharged untreated sewage. Since 2001, a wastewater treatment plant (WWTP) with a secondary treatment stage discharges treated effluents into the sea northern of the capital. However, a part of untreated effluents of urban and suburban areas are still discharged into the sea mostly due to the incomplete wastewater network.

In the framework of the Interreg III (Greece-Cyprus 2007-2009) project, data for assessing the quality of Mytilene Strait marine environment were collected. In December 2007, seawater samples were collected at 1 m depth from twenty-two sampling stations (Fig. 1). The measured values of chlorophyll-a (chl-a, μg/L), dissolved nitrates ($N\text{-}NO_3$, μmol N/L) and ammonium ($N\text{-}NH_4$, μmol N/L) formed the dataset for the present study. It should be noticed, however, that dissolved inorganic phosphorous ($P\text{-}PO_4$, μmol P/L) values were close to the detection limit of the method used indicating no significant variance. Therefore, this variable has not been included in the dataset.

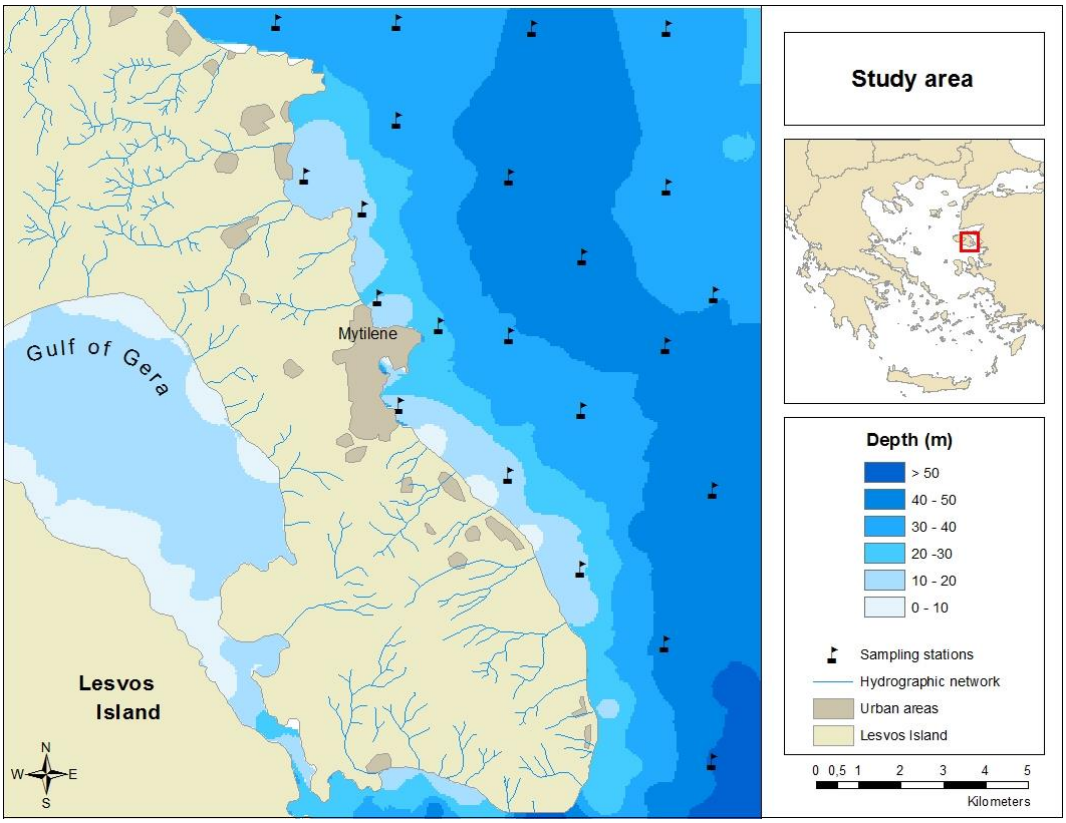


Figure 1. The coastal area of the Strait of Lesvos (NE Aegean Sea) and the location of the sampling sites

2.2. Methodology

2.2.1 Inverse Distance Weighted Interpolation (IDW) method

In two-dimensional space, spatial interpolation methods provide the means for converting the collected fragmented information from sampling surveys into a smooth gradient of data values and the creation of surfaces. There is a wide spectrum of interpolation methods. Among them, the IDW method is a spatial interpolator for estimating the value of a parameter *z* at a point *(x,y)* where no sample was available, based on values from the nearby area (Weber and Englund, 1994). During the calculation of a specific value in space, higher weight is assigned to the values measured at neighboring points. The general formula of the method is the following:

$$f(x,y) = \frac{\sum_{i=1}^{n} w(d_i) \cdot z_i}{\sum_{i=1}^{n} w(d_i)} \quad (1)$$

where *f(x,y)* is the simulated value of the parameter at point *(x,y)*, $w(d_i)$ the weighted function, z_i the measured value at point *i*, d_i the distance of point *i* from point *(x,y)* and *n* the number of the neighboring measured values considered. In this paper the weighted function $1/d^r$, where $r = 1, 2, 3, \ldots$ was used.

2.2.2 Cross-validation

The evaluation of the deviations between the interpolated surface and the data values is the most straightforward method to assess interpolation accuracy. For this purpose,

an evaluation dataset not used in the interpolation procedure should be available. However, in many applications, the availability of an independent evaluation dataset is impossible due to the limited number of input points. In the present study, the cross-validation procedure (Vehtari *et al.*, 2017) was applied. Accuracy assessment by cross-validation is based on the removal of one sampling station at a time, application of the interpolation for the location of the removed station using the remaining samples and calculation of the residual between the measured value of the removed data point and its estimate. The procedure is repeated until every sample has been, in turn, removed. The cross-validation is especially suitable for relatively no dense datasets since in under-sampled areas the removal of points to create the evaluation dataset can lead to misrepresentation of the surface to be interpolated. In this paper, the cross-validation was applied in order to assess the accuracy and select the better IDW interpolator (the better combination of *n* and *r* values) among five IDW interpolators tested. The selection of the best IDW interpolator was based on the results of the calculation of the spatial average interpolation errors and in particular of the RMSE (Root Mean Square Error), MAE (Mean Average Error), MBE (Mean Bias Error) and NMSE (Normalized Mean Square Error).

2.2.3 Principal Component Analysis (PCA)

PCA is a multivariate statistical method that creates new variables, called principal components, which are uncorrelated linear transformations of the original ones representing a high percentage of their variance (Sharma, 1996). PCA converts the initial set of data into a lower dimensional space; therefore, a better understanding of the complex and voluminous information of the initial data matrix is achieved (Dunteman, 1989). The creation of a lower dimensional space is known as "dimensional reduction", thus PCA is also characterized as a "dimensional reduction technique" (Sharma, 1996).

3. Results

The IDW interpolation method was applied in order to transform the 22 points dataset of chl-a, N-NO_3 and N-NH_4 in the study area into three (3) rasters ($R_{chl\text{-}a}$, R_{NO3}, R_{NH4}) of 800 x 800 m spatial resolution. The optimal IDW interpolator was selected for each variable after the testing of four different interpolators by application of cross-validation and calculation of the representative errors. The results are shown in Table 1. The best IDW interpolator for chl-a is that with $n = 4$ and $w(d_i) = 1/d^2$, for N-NO_3 that with $n = 4$ and $w(d_i) = 1/d$ and for N-NH_4 that with $n = 3$ and $w(d_i) = 1/d$, since the majority of the calculated errors were minimum for these interpolators. The application of the selected interpolators to the point dataset resulted in the three thematic maps illustrating the spatial distribution of each variable shown in Fig. 2. These thematic maps were the result of the classification of the pixel values of each raster based on the eutrophication scale of each variable shown in Table 4. All pixel values from the three rasters were then used for the application of PCA.

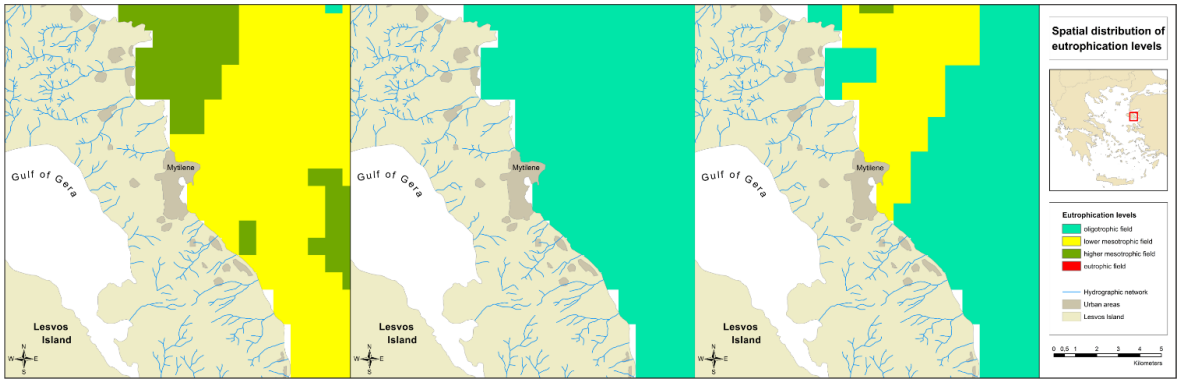
Table 1. Values of MBE, MAE, RMSE, and NMSE and final ranking of the tested IDW interpolators after application of cross-validation for each variable

chl-a								
ranking	**MBE**	**IDW**	**MAE**	**IDW**	**RMSE**	**IDW**	**NMSE**	**IDW**
1	0,018287	*r*=1 *n*=3	***0,099562***	***r=2 n=4***	***0,121885***	***r=2 n=4***	***0,048454***	***r=2 n=4***
2	0,018645	*r*=2 *n*=3	0,101016	*r*=2 *n*=3	0,122629	*r*=2 *n*=3	0,049115	*r*=2 *n*=3
3	***0,019414***	***r=2 n=4***	0,102066	*r*=1 *n*=4	0,125492	*r*=1 *n*=4	0,051348	*r*=1 *n*=4
4	0,019590	*r*=1 *n*=4	0,104253	*r*=1 *n*=3	0,127121	*r*=1 *n*=3	0,052812	*r*=1 *n*=3
N-NO_3								
ranking	**MBE**	**IDW**	**MAE**	**IDW**	**RMSE**	**IDW**	**NMSE**	**IDW**
1	-0,00019	*r*=2 *n*=3	***0,071068***	***r=1 n=4***	***0,105120***	***r=1 n=4***	***0,456937***	***r=1 n=4***
2	-0,00180	*r*=2 *n*=4	0,078024	*r*=2 *n*=4	0,110985	*r*=1 *n*=3	0,501377	*r*=2 *n*=4
3	-0,00491	*r*=1 *n*=3	0,080438	*r*=1 *n*=3	0,112010	*r*=2 *n*=4	0,502171	*r*=1 *n*=3
4	***-0,00708***	***r=1 n=4***	0,086886	*r*=2 n=3	0,118697	*r*=2 *n*=3	0,557309	*r*=2 *n*=3
N-NH_4								
ranking	**MBE**	**IDW**	**MAE**	**IDW**	**RMSE**	**IDW**	**NMSE**	**IDW**
1	0,020435	*r*=2 *n*=4	***0,235205***	***r=1 n=3***	***0,322551***	***r=1 n=3***	***0,510318***	***r=1 n=3***
2	0,023398	*r*=1 *n*=4	0,236938	*r*=2 *n*=3	0,328182	*r*=1 *n*=4	0,542336	*r*=1 *n*=4
3	0,029587	*r*=2 *n*=3	0,248922	*r*=2 *n*=4	0,333335	*r*=2 *n*=3	0,552033	*r*=2 *n*=3
4	***0,035560***	***r=1 n=3***	0,250276	*r*=1 *n*=4	0,336171	*r*=2 *n*=4	0,572773	*r*=2 *n*=4

RMSE: Root Mean Square Error, MAE: Mean Average Error, MBE: Mean Bias Error, NMSE: Normalized Mean Square Error

The higher concentrations of chl-a were recorded close to land sources and especially in the northern part of the study area. The area was characterized in general as lower mesotrophic, while in the northern part near the coast as higher mesotrophic. The values of N-NO_3 varied from 0.02 to 0.49 μmol/L with the highest ones being recorded in the north of the study area; however, the total area is characterized as oligotrophic according to the eutrophication scale. The values of N-NH_4 were significantly higher (0.10 to 1.34 μmol/L) than those of N-NO_3 and showed a similar pattern with higher values in the northern part of the study area. The higher concentrations of N-NH_4 could be attributed to the anthropogenic activities in the study area i.e. sewage effluents.

Further processing of the dataset derived from the application of the IDW interpolators, included a test of normality for each variable and calculation of their correlation matrix. The analysis showed that the variables were not normally distributed, therefore, the Spearman correlation matrix (Hauke and Kossowski, 2011) was calculated (Table 2). The higher correlation was detected between chl-a and N-NH_4 while the lower between chl-a and N-NO_3.



(a) (b) (c)

Figure 2. Spatial distributions of (a) chl-a, (b) N-NO_3 and (c) N-NH_4 in the study area

Table 2. Spearman correlation matrix for the variables chl-a, N-NO_3, and N-NH_4

	chl-a	**N-NO_3**	**N-NH_4**
chl-a	1.000		
N-NO_3	-0.036	1.000	
N-NH_4	0.433	-0.121	1.000

The next step was the application of PCA and the results are summarized in Table 3, including the loadings and eigenvalues of each principal component. It is shown that the first component (PC1) accounted for 48.7% of the total variance, the second (PC2) for 32.7%, and the third (PC3) only for 18.6%, while both PC1 and PC2 accounted for 81.4% of the total variance.

Table 3. Loadings and eigenvalues of the three variables for the principal components PC1, PC2 and PC3

	PC1	**PC2**	**PC3**
chl-a	-0.676	0.266	0.688
N-NO_3	0.237	0.962	-0.139
N-NH_4	-0.698	0.069	-0.713
eigenvalue	1.460	0.981	0.559
variance (%)	48.7	32.7	18.6
cumulative variance (%)	48.7	81.4	100

At this point, it is important to select the number of PCs that should be used for assessing and mapping eutrophication levels in the study area. In the present study both PC1 and PC2 were retained since they accounted for 81.4% of the total variance (Solanas *et al.*, 2011). The next step was the calculation and visualization of the spatial distribution of PC1 and PC2 by application of the equations (2) and (3) derived from Table 3.

$$R_{PC1} = -0.676 \cdot R_{chl\text{-}a} + 0.237 \cdot R_{N\text{-}NO3} - 0.698 \cdot R_{N\text{-}NH4} \quad (2)$$

$$R_{PC2} = 0.266 \cdot R_{chl\text{-}a} + 0.962 \cdot R_{N\text{-}NO3} + 0.069 \cdot R_{N\text{-}NH4} \quad (3)$$

where $R_{chl\text{-}a}$, $R_{N\text{-}NO3}$, $R_{N\text{-}NH4}$ the rasters produced after application of the IDW interpolation method for each variable and R_{PC1} and R_{PC2} the rasters of PC1 and PC2 respectively with a spatial resolution of 800 x 800 m. The equations were applied on the datasets on a pixel-by-pixel basis.

The values of R_{PC1} and R_{PC2} were subsequently classified into four categories based on the eutrophication scales for chl-a, N-NO_3, and N-NH_4 given in Table 4. The boundary values of the eutrophication levels were calculated using equations (2) and (3) and the result is illustrated in Fig 3.

In Fig. 3(a) it is observed that regarding PC1 most of the study area is characterized as lower mesotrophic and only two limited areas at north and south are characterized as higher mesotrophic and oligotrophic respectively. Regarding PC2, shown in Fig. 3(b), the whole study area is characterized as oligotrophic.

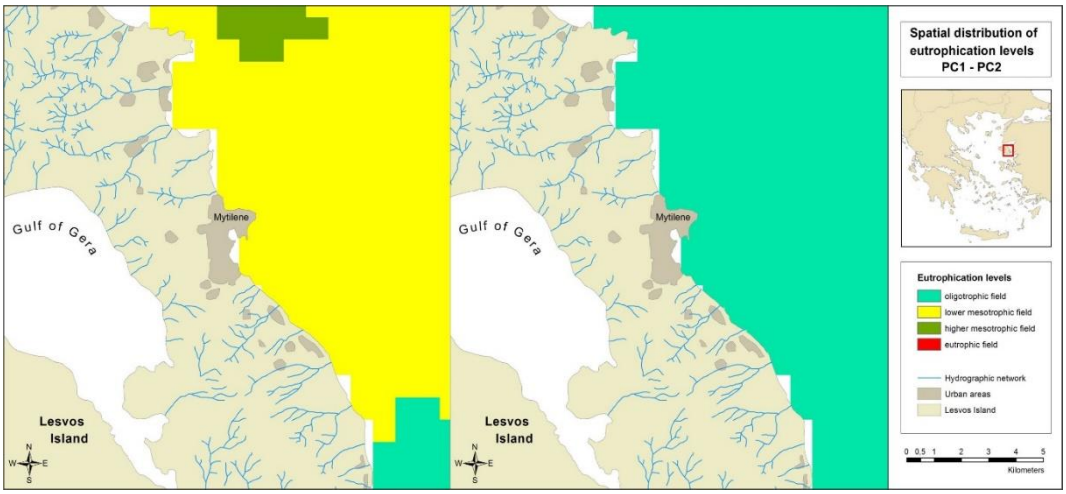
The final step of the methodological procedure involved the integration of R_{PC1} and R_{PC2} in order to produce one final thematic map illustrating the eutrophication levels in the study area. The final raster was produced by overlay of R_{PC1} and R_{PC2} (equation 4):

$$R_f = R_{PC1} + R_{PC2} \quad (4)$$

The classification of the values of R_f to four eutrophication levels was performed based on the scale shown in Table 4 which was developed by calculation of the boundary values of the eutrophication levels based on equation (4). The result is shown in Fig. 4. It is observed that (i) the northern part, (ii) the central part of the study area near the coast extending approximately 3 km offshore and (iii) a limited area at southeast are characterized as lower mesotrophic; the rest of the study area is characterized as oligotrophic.

Table 4. Eutrophication scales for chl-a, N-NO_3 and N-NH_4 (Simboura *et al.*, 2005; Kitsiou *et al.*, 2002) and the developed scales for the classification of the values of the rasters R_{PC1}, R_{PC2}, and R_f.

		class name		class name		class name		class name
chl-a (μg/L)	0.0		0.100		0.600		2.210	
N-NO_3 (μmol N/L)	0.0		0.620		0.650		1.190	
N-NH_4 (μmol N/L)	0.0	*oligotrophic field*	0.550	*lower mesotrophic field*	1.050	*higher mesotrophic field*	2.200	*eutrophic field*
R_{PC1}	**0.0**		**0.305**		**0.985**		**2.748**	
R_{PC2}	**0.0**		**0.661**		**0.857**		**1.883**	
R_f	**0.0**		**0.966**		**1.842**		**4.631**	



(a) (b)

Figure 3. Spatial distribution of eutrophication levels for (a) PC1 and (b) PC2

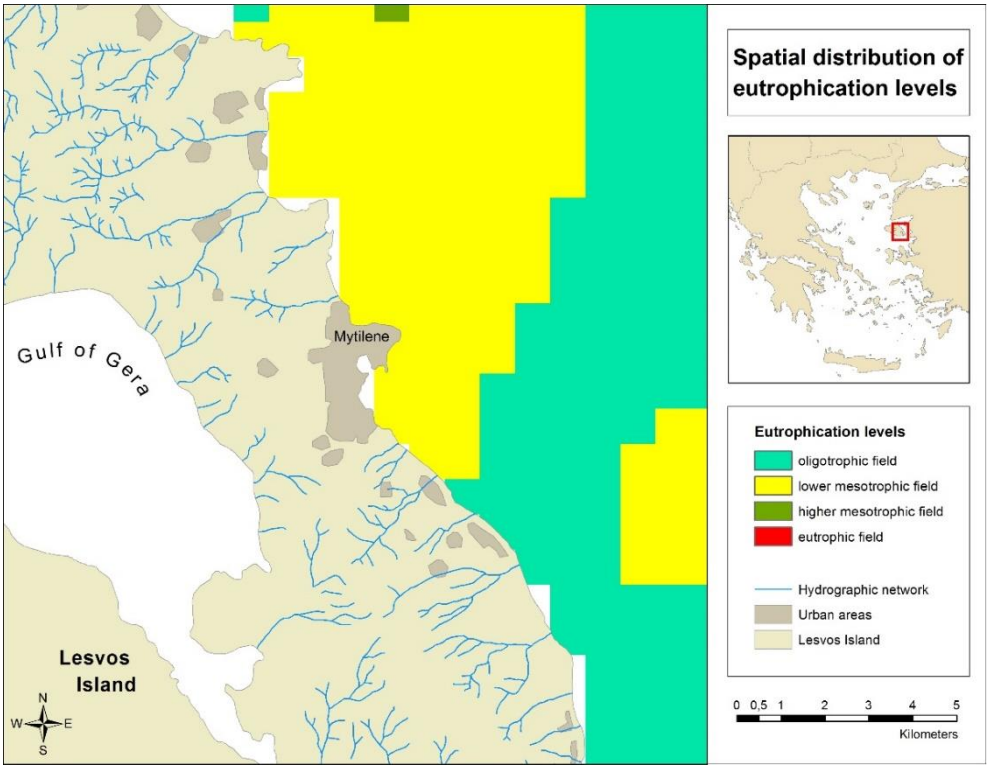


Figure 4. Spatial distribution of eutrophication levels in the study area after integration of PC1 and PC2

4. Conclusions

Since the assessment of eutrophication levels in coastal areas is of multi-parametric nature, the described methodological approach provides the means for integrating different variables in order to detect the contribution of each one to the principal components and produce a lower dimensional dataset which could be easier handled for mapping sea water eutrophication levels. The methodology described could be applied in any coastal area using a number of sea water quality variables and be useful in the field of coastal management where knowledge of sea water conditions is of great importance to stakeholders. The implementation of such methodologies in the framework of GIS is a prerequisite since GIS are both specialized Database Management Systems and a dynamic platform for the analysis of the stored spatial information; they provide, therefore, a flexible environment for the assessment of coastal water quality issues at spatial scale. Furthermore, mapping eutrophication levels by incorporating information acquired from different variables could be a useful tool to policy makers for performing effective decision-making since they require clear-cut information, clearly delineated zones relevant of water quality conditions and easily understood illustrations.

References

Akbal F., Gürel L., Bahadır T., Güler I., Bakan G. and Büyükgüngör H. (2011), Multivariate Statistical Techniques for the Assessment of Surface Water Quality at the Mid-Black Sea Coast of Turkey, *Water, Air, & Soil Pollution*, **216**, 21–37.

Andersen T., Carstensen J., Hernández-García E. and Duarte C.M. (2008), Ecological thresholds and regime shifts: approaches to identification, *Trends in Ecology and Evolution*, **24**, 49–57.

Burrough P.A., McDonnell R.A. (2000), Principles of Geographic Information Systems: Spatial Information Systems and Geostatistics, Oxford University Press, Oxford.

Caruso G., Leonardi M., Monticelli L.S., Decembrini F., Azzaro F., Crisafi E., Zappalà G., Bergamasco A. and Vizzini S. (2010), Assessment of the ecological status of transitional waters in Sicily (Italy): First characterisation and classification according to a multiparametric approach, *Marine Pollution Bulletin*, **60**, 1682-1690.

Dunteman G.H. (1989), Principal Component Analysis, Sage Publications, London.

De Jonge V.N., Elliott M. and Orive E. (2002), Causes, historical development, effects and future challenges of a common environmental problem: eutrophication, *Hydrobiologia*, **475**, 1–19.

Ferreira J.G., Andersen J.H., Borja A., Bricker S.B., Camp J., Cardoso da Silva M., Garcés E., Heiskanen A.S., Humborg C., Ignatiades L., Lancelot C., Menesguen A., Tett P., Hoepffner N. and Claussen U. (2011), Overview of eutrophication indicators to assess environmental status within the European Marine Strategy Framework Directive, *Estuarine, Coastal and Shelf Science*, **93**, 117-131.

Hauke J. and Kossowski T. (2011), Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data, *Quaestiones Geographicae*, **30**, 87–93.

Huang B., Jiang B. and Li H. (2001), An integration of GIS, virtual reality and the Internet for visualization, analysis and exploration of spatial data, *Int J Geogr Inf Sci*, **15**, 439-456.

Kitsiou D. and Karydis M. (2001), Marine eutrophication: a proposed data analysis procedure for assessing spatial trends, *Environmental Monitoring and Assessment*, **68**, 297-312.

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

Kitsiou D. and Karydis M. (2011), Coastal marine eutrophication assessment: A review on data analysis, *Environment International*, **37**, 778-801.

Osburn C.L., Handsel L.T., Peierls B.L. and Paerl H.W. (2016), Predicting Sources of Dissolved Organic Nitrogen to an Estuary from an Agro-Urban Coastal Watershed, *Environ. Sci. Technol*., **50**(16), 8473-8484.

Primpas I., Tsirtsis G., Karydis M. and Kokkoris G.D. (2010), Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive, *Ecological Indicators*, **10**, 178-183.

Sharma S. (1996), Applied multivariate techniques, John Wiley & Sons, Inc, New York.

Simboura N., Panayotidis P. and Papathanassiou E. (2005), A synthesis of the biological quality elements for the implementation of the European Water Framework Directive in the Mediterranean ecoregion: The case of Saronikos Gulf, *Ecological Indicators*, **5**, 253-266.

Solanas A., Manolov R. and Leiva D. (2011), Retaining principal components for discrete variables, *Anuario de Psicología*, **41**(1-3), 33-50.

Vega M., Pardo R., Barrado E. and Debán L. (1998), Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, *Water Research*, **32**, 3581-3592.

Vehtari A., Gelman A. and Gabry J. (2017), Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, **27**(5), 1413-1432.

Washington H.G. (1984), Diversity, biotic and similarity indices, Water Research, 18, 653–694.

Weber D.D. and Englund E.J. (1994), Evaluation and comparison of spatial interpolators 2, *Mathematical Geology*, **26**, 589-603.