

**The use of principle component analysis in data reduction
for GIS Analysis of water quality data**

M. McAdams, A. Demirci

Geography Department

Fatih University

Istanbul, Turkey

Abstract

Water quality is measured by several key indicators, but its overall composition and spatial distribution is often difficult to discern. In this paper, the authors will explore the usefulness of the application of Principle Component Analysis (PCA)—a statistical database reduction technique- in examining the spatial attributes of water quality indicators (Nitrogen, Phosphorus, COD, BOD, DO, Ph, turbidity, conductivity, Chlorophyll-A) from samples from Küçücekmece Lake in Istanbul, Turkey. The research showed that the use of PCA may be a useful tool in reducing the complexity of the data and revealing spatial distribution of the water quality indicators.

Introduction

Environmental analysis is innately multi-dimensional. When measuring environmental variables such as air pollution, or water quality, one uses multiple measurements. In both the cases of water and air pollution, one is particularly concerned with their spatial distribution related to the sources, mixing and intensity. These measurements are not mutually exclusive and are often intercorrelated. When you map one measurement, you must also be concerned about the relationship of another.

In the case of water quality and pollution, there are multiple measurements that are taken. However, it may not be clear how these measurements are spatially connected. When, taking multiple samples over different time periods, the discernment of trends is difficult. The use of statistical analysis such as database reduction techniques such as Principle Component Analysis is indicated in such conditions.

As part of a project to inspect the water quality and land use relationships of Küçükçekmece Lake in Istanbul, the research team is in the process of taking multiple measurements at different locations in the lake (see figure 1). Samples were scheduled to be taken three times periods: early spring, early summer and late summer to measure the intensity of key indicators. The research team has already taken samples in April, and June in 2006 and will take another in August of 2006. (The analysis from the April samples have been completed, but those from June are still in process of being analyzed.) It became clear that mapping the spatial distribution of the measurements was helpful but did not truly reflect their interrelation.. As a result, the team explored the use of PCA and believes that it may hold some promise for examining the interrelationships of the measurements.

GIS analysis and Principle Component Analysis (PCA)

When visualizing data in a GIS, one is able to map several variables and their distribution, but it is often difficult to not determine the relationship of these variables clearly. One could show a bar chart for multi-variable analysis, but this only implies interrelationships. To show the interrelationship between data and variables there is a need to integrate statistical analysis into the GIS. PCA is a robust statistical technique to reduce data and develop composite variables. However, it is not linked spatially. The GIS can take the data and display the spatial tendencies.

Principle Component Analysis or factor analysis is a statistical technique to analysis individual measurements that are inter-correlated. Through the use of statistical groupings, the measurements are organized into components or variables.

The use of PCA in GIS analysis has been used in several cases in a variety of wildlife studies. (1),(2) (3) and water quality analysis (4). It has been widely used in classification of Remote Sensing images because of the multi-dimensional aspect of spectral analysis.

Methodology of PCA use in analysis of water quality samples of Küçücekmece Lake

The sample data from April 2006 (N, Phosphorus, DO, COD, BOD, Chlorophyll-A, conductivity, Ph, temperature, depth, and turbidity) was inserted into a GIS system (ArcGIS). The data was first interpolated using the Spatial Analysis extension for ArcGIS. to show the spatial distribution of the data samples. The data was then imported in SPSS and PCA was performed. It yielded three principle components and the relationship to each component by sampling point. The PCA analysis was then joined to the GIS and an addition spatial interpolation performed using Spatial Analysis

Results

When inspecting the results of the April 2006 samples in Figure 2 there are discernable spatial relationships for the different measurements. However, these measurements are not mutually exclusive. After performing Principle Component Analysis, three principle components were extracted. (see Table 1, Table 2 and Figure 3). These are indicated as P-1, P-2, and P-3 in the Figure 2. The most significant relationships are found in P-1. This component shows that there are strong positive relationships between P, N, turbidity and chlorophyll-A, but a negative relationship for DO, Ph and Conductivity. This could be considered the Pollution variable. The leading measurement in this component is Chlorophyll-A. P-2 loads positively on other measurements, but it is lead by a strong negative relationship with DO. This could be considered the Good Water Quality variable. P-3 is lead by Phosphorus, but has moderately strong relationships with the polluting elements. This could be considered the Moderately Polluted Variable.

When the relationships with the different components are mapped (see Figure 2), it summarizes what the team knows about the lake and the relationships between the different measurement elements. The PCA analysis seems to distinguish between different areas of the composition of the lake. The areas correlating strongly with the Polluted variable (P-1) are areas know for discharge of pollutants such a Phosphorus and Nitrogen due to the entry of pollutants from streams that were carrying industrial waste at the time the samples were collected. The other section of the lake which is buffered by a wildlife protection area and minimal urbanization are where the highest loadings on the Good Water Quality variable are found (P-2). The third component the Moderate Water quality variable are found in a area which is receiving mixing of saline water from the Marmara Sea and area of diffusion of pollutants

Conclusions

It is clearly shown that the mapping of the different measurements can be useful for indicating the levels of the different pollutants. However, it is difficult to discern the relationships between the variables. After PCA was performed, the different component appear to be measuring composite variables related to water quality. The sample team will explore the further use of PCA analysis and perhaps alternative statistical techniques when analyzing the samples from June and August of 2006.

References

1. N. Pettorelli, S. Dray, D. Malliard, 2005, Coupling Principle Component Analysis and GIS to map deer habitats, *Wildlife Biology* 11: 363-370.
<http://www.steph280.freesurf.fr/files/articles/wb2005.pdf>. Last visited 15 June 2006.
2. P/ Reunanen, M. Monkkonen, A. Nikula, 2000, Managing boreal forest landscapes for flying squirrels, *Conservation Biology*, Vol. 14, No. 1 (Feb., 2000) , pp. 218-226.
3. W. Hargrove, R. J. Luxmoore, 2006, A spatial clustering technique for the identification of customizable ecoregions, Oak Ridge National Laboratories (USA) website, <http://research.esd.ornl.gov/~hnw/esri/> . Last visited 15 June 2006/
4. L. Matejcek, 2006, Modelling of Water Pollution in Urban Areas with GIS and Multivariate Statistical Methods, Proceedings of the 3rd Biennial meeting of the International Environmental Modelling and Software Society, http://www.iemss.org/summit/papers/s2/25_Matejcek_3.pdf, Last visited 15 June 2006.

**Figure 1;
Sample Locations**

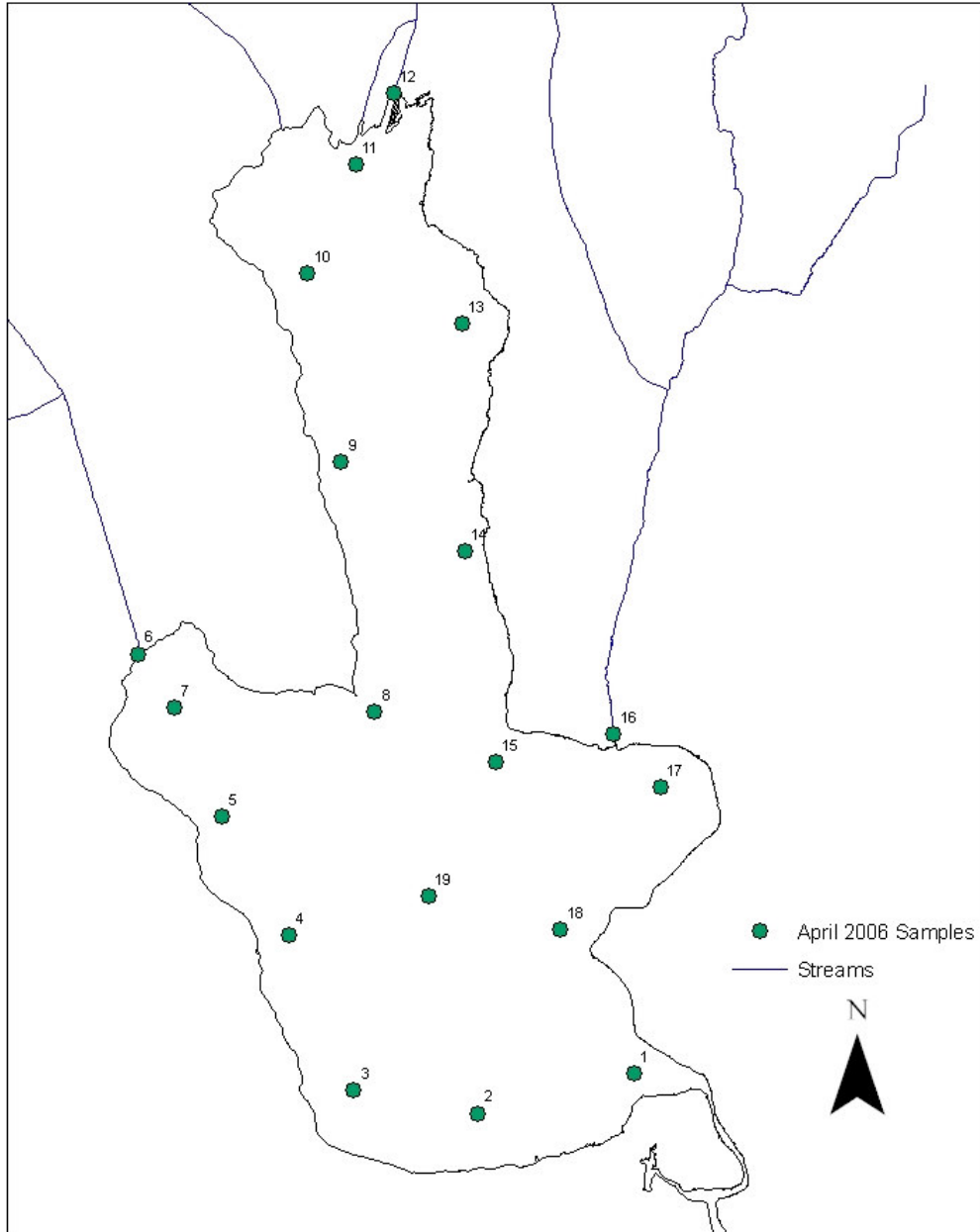


Figure 2

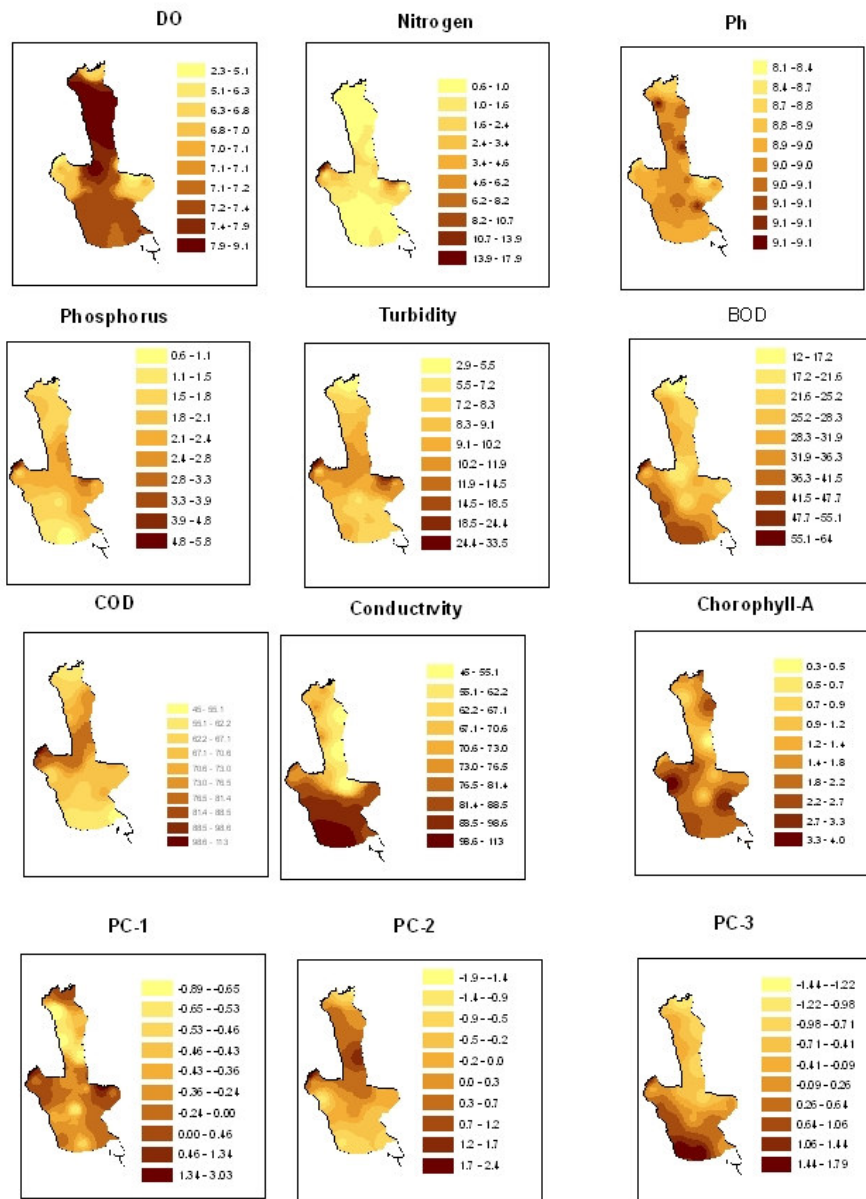
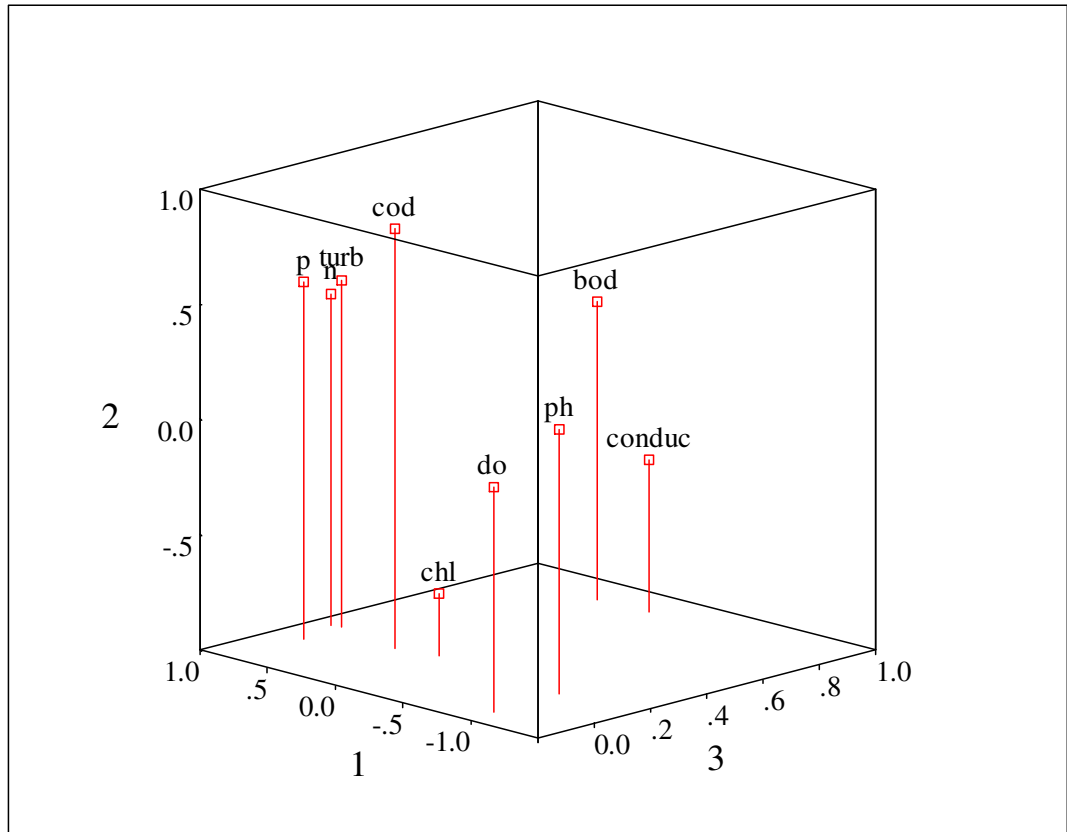


Figure 3: Principle Components Rotated in 3-D



**Table 1:
Measurements by Sampling Points and Associated Principle Components**

NO	DEPTH	PH	DO	TEMP	CONDUC	CHL	BOD	COD	N	P	TURB	FAC1_1	FAC2_1	FAC3_1
1	11.50	8.91	7.40	11.80	8.27	1.87	30.00	51.00	.56	2.01	7.54	-0.1	-0.6	0.0
2	19.00	8.97	7.15	11.70	15.46	1.63	48.00	61.00	1.12	.62	7.82	-0.5	-0.5	1.8
3	13.50	8.96	7.21	11.10	15.07	2.74	46.00	58.00	.56	1.49	9.60	-0.2	-1.0	1.8
4	3.00	8.98	7.35	11.50	11.40	2.00	46.00	69.00	.56	1.49	8.40	-0.4	-0.1	1.2
5	8.00	8.95	6.84	11.70	8.90	3.99	38.00	62.00	1.12	2.00	9.01	0.2	-1.2	0.7
6	1.00	8.53	3.07	11.70	7.36	.62	64.00	113.00	17.92	5.78	33.50	2.2	2.4	1.4
7	2.50	8.96	7.10	11.50	6.15	2.07	26.00	80.00	.56	1.57	7.02	-0.3	0.2	-0.3
8	1.00	9.06	8.46	12.00	5.88	2.42	12.00	83.00	.56	2.18	11.20	-0.4	0.4	-0.7
9	2.70	9.05	8.64	11.20	6.41	1.12	30.00	70.00	.56	1.56	8.57	-0.8	0.6	-0.1
10	2.70	9.09	9.06	11.20	6.35	.70	32.00	63.00	.56	2.02	8.20	-0.9	0.8	-0.1
11	1.00	8.73	6.52	11.50	6.26	1.39	14.00	60.00	.56	1.56	4.51	0.2	-0.6	-1.1
12	1.00	8.57	6.32	11.60	3.68	2.92	14.00	45.00	.56	1.04	2.90	0.7	-1.9	-1.4
13	3.60	9.06	8.17	12.00	2.31	2.47	22.00	76.00	.56	1.47	8.38	-0.5	0.2	-0.8
14	1.20	9.08	8.25	12.60	1.37	.34	26.00	80.00	.56	2.49	9.16	-0.7	1.6	-1.0
15	2.10	9.05	7.39	12.80	.89	.96	30.00	66.00	.56	2.16	9.07	-0.4	0.8	-1.0
16	.50	8.08	2.30	12.09	4.66	1.70	28.00	72.00	11.76	4.49	28.00	3.0	-0.4	-0.8
17	6.40	9.02	7.10	12.60	9.36	1.73	22.00	70.00	.56	1.99	9.49	-0.3	0.0	0.0
18	14.50	9.08	7.39	12.40	9.60	3.76	32.00	72.00	.56	2.48	7.05	-0.2	-0.6	0.7
19	3.10	9.06	7.36	12.50	9.42	.82	20.00	66.00	.56	1.40	6.75	-0.6	0.2	-0.2

**Table 2:
Principle Component Rotated Matrix**

	Component		
	1	2	3
Ph	-.946	.146	.140
DO	-.970	-2.668E-02	-.102
Conductivity	-.108	-.345	.862
Chlorophyll	3.635E-02	-.731	.187
BOD	.254	.290	.851
COD	.312	.813	.165
N	.866	.440	.200
P	.776	.545	6.478E-02
turbidity	.813	.502	.216

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization' a Rotation converged in 6 iterations.