Water quality observation using Principal Component Analysis

Jongchul Jeong and Sinjae Yoo

Korea Ocean Research and Development Institue
Ansan Sa-Dong 1270
South Korea, 425-170

E-mail : jcjung@sari.kordi.re.kr
sjyoo@sari.kordi.re.kr

Abstract

The aim of the present study is to define and tentatively to interpret the distribution of polluted water released from Lake Sihwa into Yellow Sea using Landsat TM. Since the region is an extreme case 2 water, empirical algorithms for chlorophyll-a and suspended sediments have limitations.

This work focuses on the use of multi-temporal Landsat TM. We applied PCA to detect evolution of spatial feature of polluted water after release from the lake.

The PCA results were compared with in situ data, such as chlorophyll-a, suspended sediments, Secchi disk depth (SDD), surface temperature, radiance reflectance at six bands. The in situ remote sensing reflectance was analysed with PCA. On the basis of these in situ data we found good correlation between first Principal Component and Secchi disk depth (R^2=0.7631), although other variables did not result in such a good correlation.

The problems in applying PCA techniques to multi-spectral remote sensed data are also discussed.

Keywords ; PCA, water quality

I. Introduction

Principal component analysis (PCA) can be used to simplify data structure of satellite multi-spectral images. The PCA is a method of redundancy reduction of multi-dimensional data in which the axis variables are orthogonal (Richards, 1986). PCA is an important data transformation technique used in remote sensing work with multi-spectral data or other multi-dimensional data. There have been many studies on the application of PCA to land use and land cover classification. (Lu jiaju 1988, Coneses et al. 1988, Kramber et al. 1988, Ceballos and Bottino 1997), while PCA has not been commonly used for analysis of in-land water and coastal water quality.

For case 2 water, some authors have presented methods to obtain estimation of the concentration of SS, chlorophyll-a and organic material. (Tassan 1988, Tassan and d'Alcala 1993). The most popular method is to derive empirical regression equation between remote sensing reflectance and sea-truth data collected concurrently. Algorithms including a combination of channels have been presented by a number of authors (Hinton, 1991). Expecially, Tassan (1983) proposed band ratio and band difference algorithm for coastal water.

This work focuses on the use of multi-temporal in-situ remote sensing reflectance to which pca is applied to detect evolution of spatial feature of polluted water after release from a polluted lake. And, the problems in applying PCA techniques to multi-spectral remote sensed data are also discussed.

II. Study area

The Lake Sihwa was artificially constructed in 1994. The area showed a typical coastal water environment before the construction. The lake has been polluted and changed in its ecological properties recently. Due to the polluted water inflow from Sihwa and Banwul, municipal and industrial region, Lake Sihwa has been highly eutrophicated since 1994. To mitigate the problem, the government decided to release the lake water into the adjacent sea since 1996.

In this study dispersion area of polluted water released from Lake Sihwa were estimated using Landsat TM analysed with PCA technique. The in situ sea-truth data were collected in Lake Sihwa and near coastal area, Kyunggi Bay.

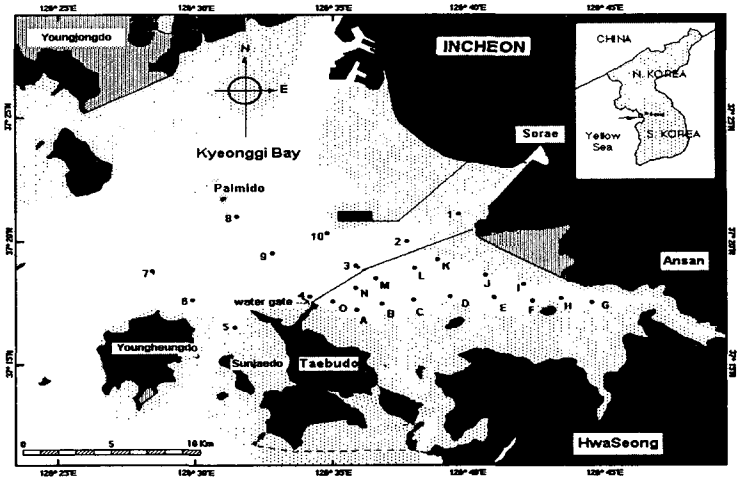

Fig. 1 Study area and sampling points

III. Dataset and processing

The in-situ remote sensing reflectance which was analysed with PCA, were compared with in situ sea-truth data, such as chlorophyll-a, suspended sediments, Secchi disk depth and surface temperature. In situ sea-truth data were collected on Jun 1997, Aug 1997, Oct 1997 and Apr 1998. The in situ downwelling irradiance and upwelling radiance were measured with PRR600 (Biospherical Inc.) at 412, 443, 490, 510, 555 and 665 with 10 nm band width.

In order to assess dispersion area, multi-spectral analysis was carried out with PCA method applied to two datasets of Landsat TM images (band 1,2,3,4,5,7) in two different seasons in the year 1997.

Table 1. Dataset specifications

Date	Sun elevation (deg)	Tide height (hh:mm)	(m)	Tide stage	Tide measured
28 Mar 1997	46	07:36	0.82	flood tide	Inchon
		19:48	7.9		
16 Jun 1997	63	07:11	2.94	flood tide	Inchon
		13:21	6.80		

Table 2. Dataset specifications using in-situ remote sensing reflectance

	Rrs412	Rrs443	Rrs490	Rrs510	Rrs555	Rrs665
average	0.00381	0.00431	0.00740	0.00959	0.02086	0.00814
std.dev	0.00585	0.00667	0.01063	0.01270	0.0229	0.00977
variance	0.00003	0.00004	0.00011	0.00016	0.00052	0.00009

Table 3. Covariance matrix of in-situ remote sensing reflectance

	Rrs412	Rrs443	Rrs490	Rrs510	Rrs555	Rrs665
Rrs412	0.00003					
Rrs443	0.00003	0.00004				
Rrs490	0.00006	0.00006	0.00011			
Rrs510	0.00007	0.00008	0.00013	0.00015		
Rrs555	0.00011	0.00012	0.00020	0.00026	0.00051	
Rrs665	0.00004	0.00004	0.00006	0.00009	0.00020	0.00009

Significant correlations among three channels were found for Rrs412, Rrs443, and Rrs 490 (maximum r=0.99077) in Table 4. In the covariance matrix of in-situ remote sensing reflectance, Rrs510 and Rrs555 are of high value.

For all datasets, at least approximately 89 per cent of the total variance was explained by the first eigenvector. Eigenvalues and percentage variance explained by each vector are presented in table 5. It is interesting to note that the first eigenvector takes the greatest portion of the total variance.

Table 4. Correlation matrix of in-situ remote sensing reflectance

	Rrs412	Rrs443	Rrs490	Rrs510	Rrs555	Rrs665
Rrs412	1.00000					
Rrs443	0.98905	1.00000				
Rrs490	0.97999	0.98844	1.00000			
Rrs510	0.96791	0.96384	0.99077	1.00000		
Rrs555	0.85531	0.80300	0.85771	0.91430	1.00000	
Rrs665	0.71143	0.62826	0.68185	0.74826	0.91376	1.00000

Table 5. Eigenvalue and variance of the PCA computed on the in-situ remote sensing reflectance.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
eigenvalue	5.3497	0.56502	0.06472	0.01857	0.00158	0.00034
variance	0.89162	0.09417	0.01078	0.00309	0.00026	0.00005

An eigenvector analysis has been done on the reflectance curves in figure 2. In this figure we could find that Lake Sihwa and near coastal belong to case 2 water which is dominated by CDOM. It has a low reflectance value in 400 - 500 nm. The variance of eigenvector loadings for each principal component (PC) in table 5 and figure 2 suggested that in-situ remote sensing reflectance was predominantly affected by CDOM in this study area.

PC 1 and PC 2 have been loaded at 412 and 443 bands which was decreased in reflectance in figure 2. The greatest factor loadings occur at the shorter wavelength. Since 412 band has been more effected by CDOM and 443 band by chlorophyll, it is not difficult to distinguish the effect of chlorophyll-a and CDOM at these wavelengths. When chlorophyll concentration are low, the effect of CDOM dominated in these bands.

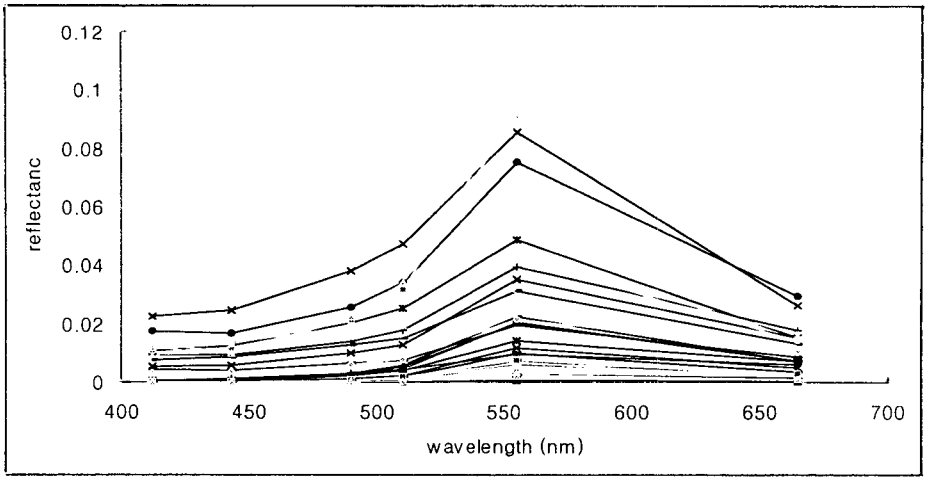

Fig.2 Reflectance curves in study area

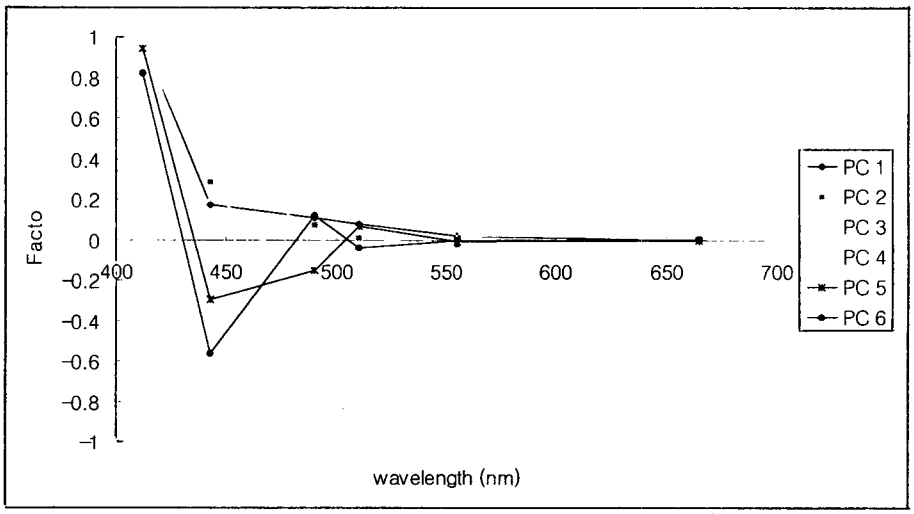

Fig.3 Factor loading of the six eigenvector with wavelength

IV. Results

Because the factor scores indicate the relative dominance of each variable at each station, the aim of PCA is to summarise the optical characteristics of the coastal water and Lake Sihwa water. When the station are ordinated in the PC1 and PC2, space has ordination which can be interpreted to be related to the water quality. In the figure 4, Lake Sihwa station in Jun 97, Aug 97 and Oct 97 were relatively more polluted, Lake Sihwa station in Apr 98 were less polluted after continuous release.

PC1 is highly correlated with Secchi disk depth in Lake Sihwa and coastal area. (R^2=0.7631) To be sure, SDD is a visual measure of the clarity of the water. In this study area CDOM is the most important water quality parameter for assessment of dispersion area, however, CDOM was not measured. Therefore, SDD was used as an indicator.

In order to assess the dispersion area of polluted water released from Lake Sihwa, Landsat TM image was analysed with PCA. The in-situ remote sensing reflectance was highly correlated with SDD, while remotely sensed data did not (R^2=0.6476).

Each sampling point was distributed in the plot of PC1 and PC2 with a linear trend from station 9 to station E (Figure 6).

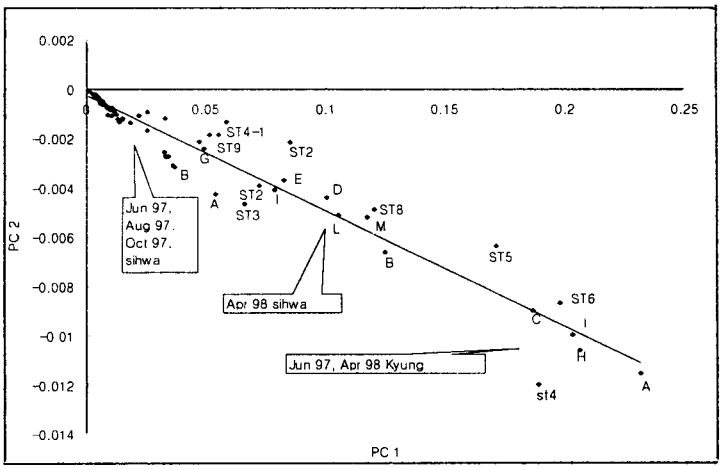

Fig. 4 Plot of the relationship bewteen PC 1 and PC 2

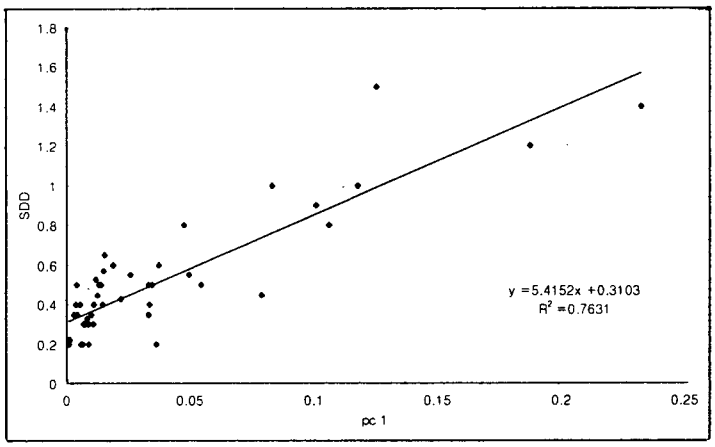

Fig. 5 Plot of the relationship bewteen PC 1 and Secchi Depth

They are ordinated from high transparency to low transparency. This sequence roughly corresponds to decreasing water quality.

The scores of PC 1 were plotted with Secchi depth collected in 25 sampling points (Figure 7). They were analyzed with Pearson product moment correlation (r=-0.805) and Spearman rank order correlation (r= -0.845). A similar relationship between SDD and PC 1 scores from remotely sensed data obtained in Jun 1997. However, Landsat TM has six wavelength bands in difference band position and width: blue, green, red and IR bands. Especially, band 1 and band 2 are blue-green and yellow-green bands.

The variance explained by PC1 of remotely sensed data using TM is 76 per cent. The band 1 of TM is 450 - 520 nm. Therefore, this band is less effected by CDOM than chlorophyll.

Although in-situ based studies have produced some primary results in applying PCA for water quality, there are difficulties in applying the technique to the remotely sensed data. Landsat TM data are contaminated with atmospheric noises such as Rayleigh and Mie scattering.

Table 6. Eigenvalue and variance of the PCA computed on the TM image in Jun 1997.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
eigenvalue	4.6062	0.7548	0.3393	0.1400	0.1200	0.0388
variance	0.7677	0.1258	0.0565	0.0233	0.0201	0.0064

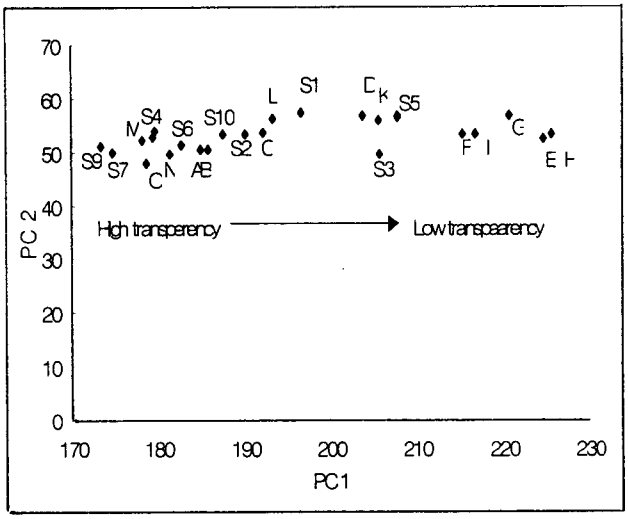

Fig. 6 Plot of the relationship bewteen PC 1 and PC 2

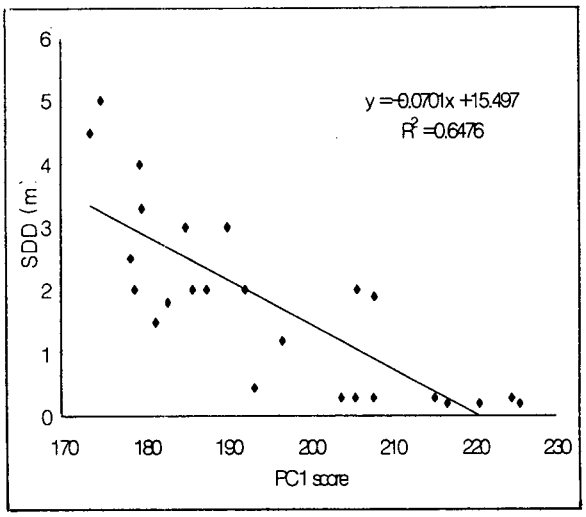

Fig. 7 Relationship of the PC1 and Secchi depth

V. Summary

PCA was suggested as a method for detecting dispersion area of polluted water. When the PC1 was compared with SDD, they are highly correlated, although other variables did not result in such a good correlation. The PC1 obtained by in-situ remote sensing reflectance was useful for detection of water quality from the water body where CDOM dominated.

On the other hand, TM has the different band width from PRR, and band 1 of TM does not include 412 and 443 nm bands. Atmospheric noises which were not corrected in this study might add further errors. Therefore, it is maybe less effective to detect water quality where CDOM dominates. In spite of such limitations, PCA results of Landsat TM data provided an Pearson product moment correlation (r=-0.805) and Spearman rank order correlation (r=-0.845) between PC1 and SDD.

Reference

Ceballos, J. C., and Bottino, M. J., 1997, The discrimination of scenes by principal components analysis of multi-spectral imagery,*Int. J. Remote Sensing,* vol. 18, No. 11 :2437-2449.

Hinton, J. C., 1991, Application of eigenvector analysis to remote sensing of coastal water quality, *Int. J. Remote Sensing,* vol. 12, No. 7 :1441-1460.

Richards, J. A., 1990, Thematic Mapping from Multitemporal Image data using the principal components transformation, *Remote Sens. Environ.* 16:35-46.

Tassan, S., 1988, The effect of dissolved 'yellow substance' on the quantitative retrieval of chlorophyll and total suspended sediment concentration from remote measurements of water colour, *Int. J. Remote Sensing,* 4:787-797.

Tassan, S., and d'Alcala., 1993, Water quality monitoring by thematic mapper in coastal environments. A performance analysis of local biooptical algorithm and atmospheric correction procedures, *Remote Sens. Environ.* 45:177-191.