Utilizing Principal Component Analysis in Unsupervised Classification Based on Remote Sensing Data

Byung-Gul Lee[1] and In-Joon Kang[2]

[1]Department of Civil and Environmental Engineering, Cheju National University, Jeju, 690-756, Korea
[2]Department of Civil Engineering, Pusan Natioanl University, Busan, 609-735, Korea

Principal component analysis (PCA) was used to improve image classification by the unsupervised classification techniques, the K-means. To do this, I selected a Landsat TM scene of Jeju Island, Korea and proposed two methods for PCA: unstandardized PCA (UPCA) and standardized PCA (SPCA). The estimated accuracy of the image classification of Jeju area was computed by error matrix. The error matrix was derived from three unsupervised classification methods. Error matrices indicated that classifications done on the first three principal components for UPCA and SPCA of the scene were more accurate than those done on the seven bands of TM data and that also the results of UPCA and SPCA were better than those of the raw Landsat TM data. The classification of TM data by the K-means algorithm was particularly poor at distinguishing different land covers on the island. From the classification results, we also found that the principal component based classifications had characteristics independent of the unsupervised techniques (numerical algorithms) while the TM data based classifications were very dependent upon the techniques. This means that PCA data has uniform characteristics for image classification that are less affected by choice of classification scheme. In the results, we also found that UPCA results are better than SPCA since UPCA has wider range of digital number of an image.

Key words: Principal component analysis (PCA), K-means, Unstandardized PCA, Standardized PCA, Error matrix, Unsupervised classification

1. Introduction

The important thing of remote sensing lies in the energy that is transmitted and received by every substance on the Earth. This energy travels in various wavelengths and frequencies and depending on those, takes on different forms, such as blue light, red light, microwaves, or radio waves[1)].

Remote sensing also refers to the process of collecting images with non ground-based systems Remote sensing mainly includes radar, photography, and spectrometry, both orbital and airborn. In other words, if we look at something from an airplane or from space, we have engaged in remote sensing[2)].

Each image from remote sensing is comprised of a series of square pixels or building blocks arranged in a regular pattern of rows and columns

Corresponds Authors: Byung-Gul Lee, Cheju National Univeristy, Jeju 690-756, Korea
Phone: 064-754-3455
e-mail: leebg@cheju.ac.kr

The intensity at which pixel is displayed is governed by the digital value representing it which in turn is a representation of the reflected light from that portion of the target which the pixel represents. For example, in gray color scale, the low digital number is low light intensity that is dark color, the high number is high intensity that is related to bright one. By its very nature, digital imagery is readily amenable to processing by computer such as image classification. Often times we can enhance images through this understanding and by stretching images and band ratio. By manipulating data that we understand very well, we can often produce images that are far more clear than the original ones we started off with[3)]

2. Principal component analysis

In a typical multi-spectral image, many of the bands are correlated. This correlation results in non-zero off-diagonal elements in the covariance matrix.

The Principal components trans-formation is a

linear transformation that is closely related to Factor Analysis. Its structure can be described as a weighted linear combination. Perhaps the easiest way to understand the result of the Principal Components transformation is to think of the process as a mathematical determination of a new set of axes in band space such that the resulting images are uncorrelated with one another and ordered in terms of their explanatory power.

It produces a new set of bands (called components) by multiplying each of the bands in the original image by a weight, and adding the results as,

$$C = a1B1 + a2B2 + a3B3 .... anBn \quad (1)$$

where a1, a2...an are eigenvectors, B1,B2,...Bn are digital number of an images.

PCA has been used for monitoring land-cover change based on mulitemporal Landsat data and to highlight regions of localized change evident in satellite multi-spectral imagery associated with bushfire damage and with vegetation regrowth[4]. PCA has also been applied to land cover characterization based on multi-temporal AVHRR data [4-5]. In that study, they found that PC1 reflects the spatial variation of NDVI in global land-cover types of Arizona. In general, since the PC1 primarily relates to the variance of single variable and the second and others component represent change elements of successively decreasing magnitude, PCA can be used for temporal variation of landscape change[6]. They successfully standardized PCA to the land cover change of Africa.

2.1 Unstandardized PCA method

To perform PCA, we need to diagonalize the covariance matrix of the remote sensing data. The covariance function Var between band i and j can be calculated as[5].

$$Var(k,l) = \frac{1}{NN}[\sum_{i=0}^{N}\sum_{j=0}^{N}(P_k(i,j) - P_{km})(P_l(i,j) - P_{lm}) \quad \text{----------(2)}$$

in which N is the total number of pixels, k and l are band numbers. $P_k(i,j)$ and $P_l(i,j)$ are the pixel values (digital number) column j of row i in bands k and l. P_{km} and P_{lm} are mean of band k and l, respectively.

2.2 Standardized PCA method

In previous section, the principal components calculated using the covariance matrix are referred to as unstandardized PCA and those calculated using the correlation matrix are referred to as standardized PCA. To calculate the rotation, we can start with either a variance-covariance matrix
or a correlation matrix. The standardized PCA is an alternative method of computing a PC rotation is to derive the transformation matrix on the eigenvectors of the correlation matrix instead of the covariance matrix. The correlation matrix is equivalent to a covariance matrix for an image where each band has been standardized to zero mean and unit variance. This method tends to equalize the influence of each band, inflating the influence of bands with relatively small variance and reducing the influence of bands with high variance.

The correlation matrix can be calculated as

$$Cor(k,l) = \frac{Var(k,l)}{\sqrt{\sigma_k}\sqrt{\sigma_l}}, \quad \sigma_k = \sum_{i=1}^{n}(IB_{ik} - \mu_k)^2, \quad \sigma_l = \sum_{i=1}^{n}(IB_{il} - \mu_l)^2 \quad \text{----------(3)}$$

in which n is the total number of pixels, k and l are band numbers. IB_{ik} and IB_{il} are the pixel values of images in bands k and l. and are means of band kth and lth images. and are variance of band kth and lth images, respectively.

Var(k,l) was derived from Eq.(2). With standardized PCA, the eigenvectors are computed from the correlation matrix. The characteristic of standardized PCA is to force each band to have equal weight in the derivation of the new component images[5].

2. Study area

For the application of PCA, the field data of Jeju Island were applied to PCA. To do this, 1/25,000 digital map, field observation data, Landsat TM data were used.

The study area, Jeju Island is a volcanic island located off the southern coast of Korea (approximately 126 05′ 10″ N to 126 58′ 37″ N and 33 06′ 31″ E to 33 35′ 55″ E) (Figure 1). In the figure, the gray lines represent contour line of elevation, the line interval is 100 m. Black line is road lines. Blue lines are county of Jeju

Island. The island is located off the southern coast of Korea and is generally flat and oval-shaped. Hill side areas of 200-300m above the sea level are gently sloped but most of them are idle land or meadows.

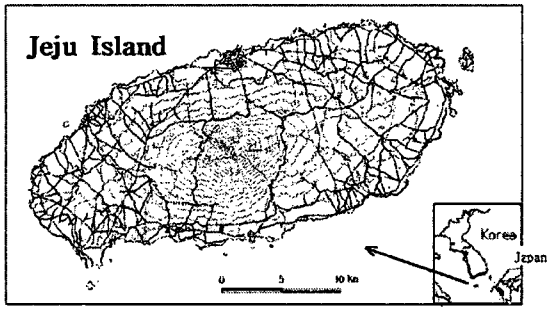

Fig. 1. The study area of Jeju Island.

The coastal area (less than 200m above sea level) is 1,013.5 ㎢. It occupies 54.9% of the whole area and is mainly used for farm land or residential areas.

And the two counties (Bukeju-gun and Nameju-gun) are total 1,335.68 ㎢ or 72.4% of the land area. The area of Jeju city is 255.13 ㎢ (13.8%), Bukeju-gun is 707.0 ㎢ (39.1%), and Nameju-gun is 614.9 ㎢ (33.3%). A high mountain, called Mt. Halla (height 1950 m) is located at the center of the island. Given this difference in elevation, the fluctuations of temperature and climate on the island are strong. Consequently, the environment of the island has a diversity of ecosystems.

The mid mountainous area (200-500m above sea level) is 496.98 ㎢. It occupies 26.9% of the whole area and is mainly meadow or idle land. The low mountainous area of 500-1000m above sea level is 253.34 ㎢. It occupies 13.7% of whole area and is mainly woods, mushroom-raising land. The area is composed of urban, grassland, forest, and agricultured and surrounded by the ocean. The island is famous for potato and tangerine production.

4. Results and discussions

Unsatnadardized PCA (UPCA) and Stand - ardized PCA (PCA) were applied to produce inputs to K-means, unsupervised classifications for Jeju Island, respectively. The two PCAs (UPCA and SPCA) data would appear to be an excellent tool for the analysis of unsupervised classification of Landsat TM data. This research results are firstly implemented in the field of unsupervised classification based on remote sensing field.

The following summarize the conclusions achieved from this study:

First, the results in the error matrices of UPCA showed that the PCA data produced classification characterized by approximately 75 % correspondence built the KRIHS reference data. Using raw TM data produced less than 61%. The K-mean algorithm applied to the TM data produced confusion among the land cover classifications, while the PCA data clearly classified the Island as three parts.

Second, SPCA also produced the same results of those of UPCA although the accuracy of SPCA is less than that of UPCA. It is also found that the PCA data had independent characteristics not affected by an classification algorithm technique, whereas the classifications of the raw data were greatly affected by the algorithm employed. Another advantage of the PC data for classification was that fewer bands (the three PC images) were used for classifying than were used to classify the original TM image(seven TM bands).

In this study, PCA was only applied the land cover classification problem of Jeju Island using Multi-spectral images. From the classification results, PCA is very useful technique to classify an area with ocean or waters that was found. Therefore, it can be expected that PCA will be also useful technique for Hyper-spectral digital images for the land classifications.

References

1) Jensen, J.R., 1996, Introductory digital image processing, Prentice-Hall, Englewood Cliffs, New Jersey, 316 p.
2) Richards, J.A.,1984, Thematic mapping from multitemporal image data using the principal components transformation, Remote Sensing of Environment, 16, 35-46.
3) Sabins, Jr, F.F., 1987, Remote Sensing: Principles and Interpretation, 2nd ed., W.H. Freeman & Co., New York City, 389 p.
4) Fung, T. and E. LeDrew, 1987, Application of principal component analysis to change

detection, Photogrammetric Engineering and Remote Sensing, 53(12), 1649-1658.
5) Hirosawa, Y, Marsh S. E. and D.H. Kliman, 1996, Application of standardized principal component analysis to land-cover characterization using mulitemporal AVHRR data, Remote Sensing of Enviroment, 58, 267-281.
6) Eastman, J.R. and M. Fulk, 1993, Long sequence time series evaluation using standardized principal components, Photogrammetric Engineering and Remote Sensing, 59(6), 991-996.