Research Journal of Forestry

ISSN 1819-3439

Academic
Journals Inc.

www.academicjournals.com

OPEN ACCESS Research Journal of Forestry

ISSN 1819-3439
DOI: 10.3923/rjf.2016.23.29

Research Article
Use of Principal Component Analysis in Accuracy of Classification Maps (Case Study: North of Iran)

[1]Masoumeh Khedmatgozar Dolati and [2]Amir EslamBonyad

[1]Department of Forestry, Faculty of Natural Sciences, Guilan University, Somesara City, Iran
[2]Department of Forestry, Faculty of Natural Sciences, Guilan University, Iran

Abstract

Objective: The main objective of this study was to investigate the role of principal component analysis to improve the accuracy of classification. **Methodology:** In the present study, PCA has been successfully applied in IRS images of north of Iran (Shafaroud) showing that the first principal components contain more variance of the information in the original four bands. **Results:** Classification was performed on two sets of data. In the first one, original bands of LISS III (b2-b3-b4-b5) and in the second classification, original bands of LISS III in combination with first component (pca1) was used. Classification was performed with five classes including sea, agriculture, settlements, broad leaf forest and needle leaf forest. In (b2-b3-b4-b5) and b2-b3-b4-b5 in composition with pca1, obtained overall accuracy and kappa coefficient were 99.27-98.94 and 99.37-99.09, respectively. **Conclusion:** The obtained results indicate that overall accuracy and kappa coefficient increases when pca 1 used along with main bands.

Key words: PCA, remote sensing, Iran

Received: August 12, 2015 Accepted: October 26, 2015 Published: June 15, 2016

Citation: Masoumeh Khedmatgozar Dolati and Amir EslamBonyad, 2016. Use of principal component analysis in accuracy of classification maps (Case Study: North of Iran). Res. J. For., 10: 23-29.

Corresponding Author: Masoumeh Khedmatgozar Dolati, Department of Foresty, Faculty of Natural Sciences, Guilan University, Somesara City, Iran Tel: 00981823223023

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Principal Component Analysis (PCA) is a mathematical technique for reducing the dimensionality of a data set[1]. Satellite remote sensing digital images are numeric; therefore, their dimensionality can be reduced using PCA[2]. Dimension reduction leads to visualization of the data clearly and subsequent data analysis more manageable[3]. In multi-band remote sensing images, the bands are the original variables. Some of the original bands may be highly correlated and to save on data storage space and computing time, such bands could be combined into new, less correlated eigen images by PCA. In addition to its use in this way, PCA can be used as a change detection technique in remote sensing[4-6]. Many studies have been used PCA for various purpose including detect geomorphologic features and sediment textural classes[7] as one of the index in decision tree classifier for land use classification[8] and to distinguish between geologic features[9]. Furthermore, this technique could be used to evaluation interannual vegetation anomalies[10]. The aim of this study was to investigate the accuracy of the classification using the combination of PCA1 with original bands.

MATERIALS AND METHODS

This work was conducted from January, 2010 to October, 2010 (10 months) at Guilan University. A multispectral digital image data set was used in this study. The images underwent radiometric and geometric pre-processing before PCA as described below. All works were done using The Environment for Visualizing Images (ENVI), GIS and image processing softwares.

Study area: A part of North of Iran was selected as the study area (Fig. 1). The maximum height of this area is 550 m above sea level with main species of forest including *Alnus glutinosa*, *Pinusteada*, *Populoussp*, *Diosperus lotus*, *Parrotia persica*, *Pinus elliotti* and *Chriptomeria*. The average maximum temperature in the warmest month and average minimum temperature in the coldest month of the year are 30.8 and 1.3C, respectively.

Image data: The IRS images can register the energy reflected by the terrestrial surface at different intervals of the electromagnetic spectrum with wavelengths ranging from the green region to the infrared (2:0.52-0.59, 3:062-0.68, 4:0.77-0.86 and 5:1.50-1.70 μm). The information from these wavelength ranges is stored in independent bands. Each band is handled as a matrix structured image where their pixels contain a Digital Number (DN) which is related with the electromagnetic energy reflected or emitted from a target.

A combination of IRS (LISS III) images obtained at May, 2007 was used (Table 1). The radiometric characteristics of the raw image data were as shown in Table 2, while the band correlation per image was shown in Table 3. Highly correlated bands are for example band2 versus band3 (Table 3). High inter band correlation indicates that the bands contain nearly the same information (in terms of radiance or reflectance data depicted). Therefore, using one of such bands instead of both may reduce the volume of data and save computation space and time. On the LISS III images the visible bands that cover the green and red spectral regions (Band2 and 3, respectively) are highly correlated, because of being in the same spectral region. As expected, infrared band is slightly independent of visible bands but the high correlation (r = 0.9149) between band5 (mid infrared) and band3 (visible red) could be due to the dry land surrounding the wetland. The dry land (dry soil) is brownish in color (giving it some red reflectance detected in band3) and quite bare. Therefore, has some mid infrared reflectance both from dry soil and dry grass. High correlation between band2 (green) and band5 (mid infrared) could be due to the dry grassland because the dry yellowish-brownish grass has elements of green and mid infrared reflectance.

Geometric control: It was essential to make the images the same size and to co-register them so that a particular pixel on one image could be identified to be the same point on another image from a different date in spite of land cover change in the period between the dates. The X-Y co-ordinate system of the images was, therefore, made similar by resampling. In this study the nearest neighbor resampling algorithm used. Nearest neighbor resampling is a favorable computationally efficient procedure, because it does not alter the pixel brightness values during resampling, whereas other interpolation techniques like bilinear interpolation and cubic convolution use averages to compute the new brightness values, often removing valuable spectral information[4,11].

The amount of error in the resampling process as shown in Table 4 and the Root Mean Square Error (RMSE) were within the advisable range of 0.25-0.50 pixel[4] or 1 pixel at the most[12].

Principal component analysis: Principal component analysis[13] is a linear transformation which decorrelates

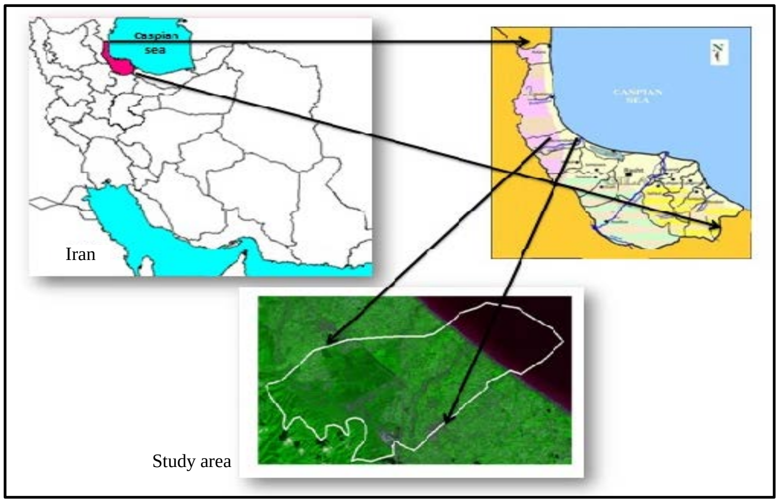

Fig. 1: Study area

Table 1: Images used for this present study

Stellite	Sensor	Path\Row	Date	Format
IRS	LISS III	67\44	2007	TIFF

IRS: Internal revenue service, LISS: Linear imaging self scanning, TIFF: Tagged image file format

Table 2: Image data univariate statistics

Image and Band	Min	Max	Mean	Standard deviation
LISS III				
2 (Green)	0	178	30.27	31.42
3 (Red)	0	169	19.42	21.2
4 (Near IR)	0	133	35.45	41.23
5 (Mid IR)	0	93	16.96	19.66

LISS: Linear imaging self scanning, IR: Infra red

Table 3: Correlation matrices of image data

	Band 2	Band 3	Band 4	Band 5
Band 2	1			
Band 3	0.977482	1		
Band 4	0.855819	0.856792	1	
Band 5	0.883851	0.914935	0.964264	1

Table 4: Results of geometric control pre-processing of LISS III images

Image co-registered	Number of ground control points	Root Mean Square Error (RMSE)
2007 LISS III	30	0.31

LISS: Linear imaging self scanning

multi variate data by translating and/or rotating the axes of the original feature space, so that the data can be represented without correlation in a new component space. Computationally, three steps are involved in the principal component transformation[14]. The first step is the calculation of a covariance or correlation matrix using the input data sets, the second step is the calculation of eigen values and eigen vectors and the third one is the calculation of principal components. The principal components calculated using the covariance matrix are referred to as unstandardized principal components and those calculated using the correlation matrix are referred to as standardized principal components[11,14]. The use of a correlation matrix in calculating principal components, implies scaling of the axes so that each feature has unit variance. This normalization process prevents certain features from dominating the analysis because of their large numerical values. A IRS image can be expressed in matrix format in the following way:

$$X_{n,b} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{4,1} & \cdots & x_{4,n} \end{pmatrix}$$

where, n represents the number of the pixels and b the number of bands. Considering each band as a vector, the above matrix can be simplified as follows:

$$X_k = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_4 \end{pmatrix}$$

where, k is the number of bands.

To reduce the dimensionality of the original bands, the eigenvalues of the covariance matrix must be calculated as follows:

$$C_{b,b} = \begin{pmatrix} \sigma_{1,1} & \cdots & \sigma_{1,4} \\ \vdots & \ddots & \vdots \\ \sigma_{4,1} & \cdots & \sigma_{4,4} \end{pmatrix}$$

where, $\sigma_{i,j}$ is the covariance of each pair of different bands.

$$\sigma_{i,j} = \frac{1}{N-1}\sum_{p=1}^{N}(DN_{p,i} - \mu_i)\left(DN_{p,i} - \mu_j\right)$$

where, $DN_{p,i}$ is a digital number of a pixel p in the band i, $DN_{p,i}$ is a digital number of a pixel P in the band j, μ_j and μ_j are the averages of the DN for the bands i and j, respectively.

From the variance-covariance matrix, the eigenvalue (λ) are calculated as the roots of the characteristic equation:

$$\det\left(C - \lambda I\right) = 0$$

where, C is the covariance matrix of the bands and I is the diagonal identity matrix.

The eigenvalues indicate the original information that they retain. From these values, the percentage of original variance explained by each principal component can be obtained calculating the ratio of each eigenvalue in relation to the sum of all those[15].

Those components which contain minimum variance and thus minimum information can be discarded.

The principal components can be expressed in matrix form:

$$Y_6 = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_4 \end{pmatrix} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,4} \\ \vdots & \ddots & \vdots \\ w_{4,1} & \cdots & w_{4,4} \end{pmatrix} \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_4 \end{pmatrix}$$

where, Y is the vector of the principal components, W is the transformation matrix and X is the vector of the original data. The coefficients of the transformation matrix W are the eigenvectors that diagonalizes the covariance matrix of the original bands. These values provide information on the relationship of the bands with each principal component. From these values it is possible to link a main component with a real variable. The eigenvectors can be calculated from the vector - matrix equation for each eigenvalue λ_k:

$$\left(C - \lambda_k I\right) w_k = 0$$

where, C is the covariance matrix, λ_k is the k eigen values (four in our example), I is the diagonal identity matrix and w_k is the k eigenvectors.

RESULTS

Principal component analysis of image data: All computation of principal components was performed using the principal component analysis facility within ENVI. Eigen values and factor loadings of the principal components from the original image data are shown in Table 5. On LISS III image, PC1 and PC2 contain 90.72% of the total variance (Table 5). Most of the data variance in this image was in near and mid infrared bands.

Principal component analysis in classification: In this study, classification was performed on two sets of data. In the first classification, just original bands of LISS III (b2-b3-b4-b5) was used. Classification was performed with five classes (sea, agriculture, settlements, broad leaf forest and needle leaf forest). Moreover, supervised classification based on the maximum likelihood algorithm classifiers was used in the classification of satellite images. Figure 2 shows classification with five classes on four original bands of LISS III.

The image classification accuracy was further assessed by calculating the kappa coefficient 'K'. The confusion matrix gave an overall accuracy of 99.27% and calculation of kappa statistics (K) gave accuracy 98.94% from bands (b2-b3-b4-b5). Other factors such asuser's accuracy and producer's accuracy were calculated by using confusion matrix. Table 6 shows overall accuracy, kappa coefficient, user's accuracy and producer's accuracy factors for original bands (b2-b3-b4-b5).

In this study, four principal components were prepared from four bands of LISS III. The first component contains more information and in other component, the amount of data decreased. Thus the first component (pca1) was used to

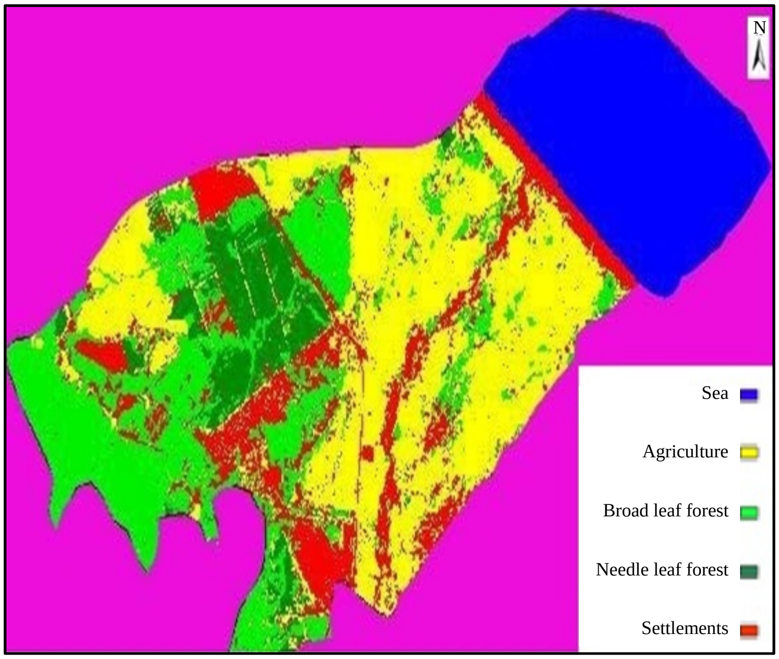

Fig. 2: Classification with five classes on four original bands of LISS III

Table 5: Principal components of image data

	Principal component			
	1	2	3	4
Band2	0.9482	0.3089	-0.0689	0.0251
Band3	0.9488	0.2838	0.1175	-0.0729
Band4	0.9736	-0.2251	-0.0337	-0.0109
Band5	0.9752	-0.0964	0.1867	0.0684
Eigen value of component	3281.30	220.73	26.36	5.04
Data SD in component	72.04%	18.68%	6.46%	2.82%

Table 6: Overall accuracy, kappa coefficient, user's accuracy and producer's accuracy factors for original bands (b2-b3-b4-b5)

Classes	Producer's accuracy (%)	User's accuracy (%)
Agriculture	99.36	99.15
Needle leaf forest	100	100
Broad leaf forest	100	100
Settlements	92.73	94.44
Sea	100	100
Kappa coefficient overall accuracy	98.94	99.27

classifying because of its high standard deviation. Figure 3 shows classification with five classes on four original bands of LISS III in combination with first component (pca1).

The confusion matrix gave an overall accuracy of 99.37% and calculation of kappa statistics (K) gave accuracy 99.09% from b2-b3-b4-b5-pca1. Other factors such asuser's accuracy and producer's accuracy were calculated by using confusion matrix. Table 7 shows overall accuracy, kappa coefficient, user's accuracy and producer's accuracy factors for b2-b3-b4-b5-pca1.

DISCUSSION

One of the important pre-processing satellite images is principal component analysis. Several Purposes of applying this technique were considered such as image enhancement and reduction of data. The findings of other previous studies have been approved the possibility of obtaining information from the land's surface using PCA on satellite images.

This is a good example of the importance of mathematics analysis to handle Information and communication

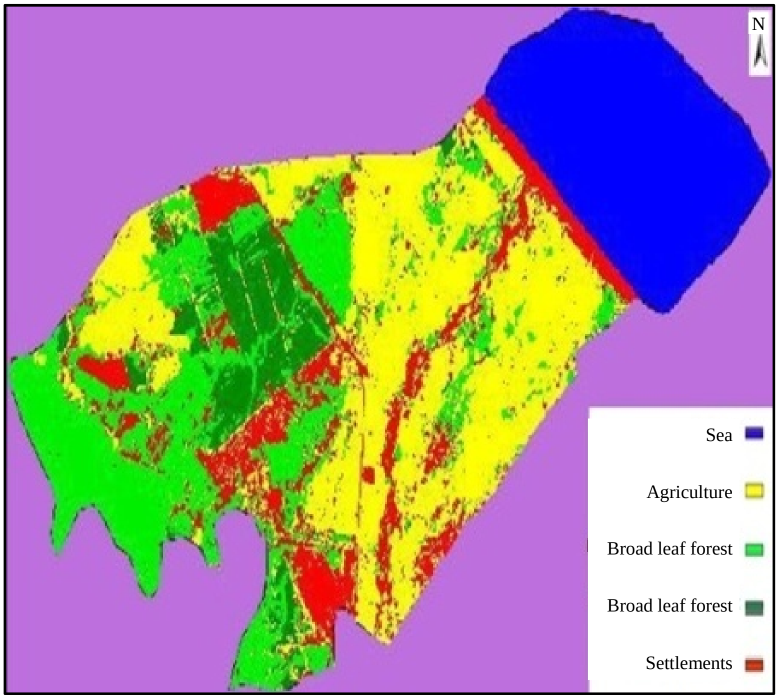

Fig. 3: Classification with five classes on four original bands of LISS III in combination with first component (pca1)

Table 7: Overall accuracy, kappa coefficient, user's accuracy and producer's accuracy factors for b2-b3-b4-b5- pca1

Classes	Producer's accuracy (%)	User's accuracy (%)
Agriculture	99.57	99.15
Needle leaf forest	100	100
Broad leaf forest	100	100
Settlements	92.73	96.23
Sea	100	100
Kappa coefficient overall accuracy	98.09	99.37

technology[16]. By employing this technique, data were placed in new space that correlation between bands vanished and created independent bands. Due to the more variance, first component among the other components of the principal components analysis selected. Some authors in the literature have found the first component has maximum variance for each class and helps to better separation phenomena[17]. In a study, Bonyad[18] used principal component analysis to reduce the correlation satellite images and overall accuracy was estimated 99.37%.

Principal components are derived from the original data in which the first principal component accounts for the maximum proportion of the variance of the original data set[19,20]. Lasaponara[10] used the Principal Component Analysis (PCA) for evaluating the vegetation interannual anomalies and shown the Principal Component1 (PC1) accounts 99.35% of the total data set variance. In another study, Estornell *et al.*[16] used PCA and found that the first three components accounted for 99.3% of the variance in the original data. But, Munyati[2] used PCA in combination with 12-band image for change detection and found first three principal components, accounting 74.04% of the total variance. Accordingly Ding *et al.*[8] pointed out the first three principal components of the variance contribution cumulative value reached for 99.78%. Other principal components contribution is little and the corresponding eigen vectors are irregular. These results are coincident with our finding and shows that the first three components particularly pca1 account highest variance.

The results of the classification of satellite image sin this study indicate the accuracy of obtained map increased when using pca1 along with the main bands. As the result of this study shows, kappa coefficient and overall accuracy in the first classification, that used just original bands of LISS III is lower than the second classification that used original bands of LISS III beside pca1. In original bands of LISS III, overall accuracy and kappa coefficient were 99.27 and 98.94, respectively, but in (b2-b3-b4-b5-pca1) composition, overall accuracy and kappa coefficient obtained 99.37 and 99.09, respectively. These high accuracy demonstrate that the combination of the spectral and textural characteristics increases the accuracy of classification. In conclusion, this study confirmed the feasibility of using PCA in remote sensing to extract land use information and increases the accuracy of classification. These calculations have been widely used in remote sensing to classify the land surface[21] and detect changes[22].

ACKNOWLEDGMENT

The authors wish to express their appreciations to Mr. Sima and Mr. Goodarzi for their valuable assistance during this project.

REFERENCES

1. Jackson, B.B. and B. Bund, 1983. Multivariate Data Analysis: An Introduction. McGraw-Hill, New York, USA., ISBN-13: 9780256028485, Pages: 244.
2. Munyati, C., 2004. Use of Principal Component Analysis (PCA) of remote sensing images in wetland change detection on the Kafue Flats, Zambia. Geocarto Int., 19: 11-22.
3. Lattin, J.M., J.D. Carroll and P.E. Green, 2004. Analyzing Multivariate Data. Thomson Brooks/Cole, Singapore, Pages: 556.
4. Jensen, J.R., 1986. Introductory Digital Image Processing: A Remote Sensing Perspective. Prentice Hall, New Jersey, USA., ISBN-13: 9780135008287, Pages: 379.
5. Fung, T. and E. LeDrew, 1987. Application of principal components analysis to change detection. Photogrammetric Eng. Remote Sens., 53: 1649-1658.
6. Muchoney, D.M. and B.N. Haack, 1994. Change detection for monitoring forest defoliation. Photogramm. Eng. Remote Sens., 60: 1243-1251.
7. Dewidar, K.H.M. and O.E. Frihy, 2003. Thematic Mapper analysis to identify geomorphologic and sediment texture of El Tineh plain, North-Western Coast of Sinai, Egypt. Int. J. Remote Sens., 24: 2377-2385.
8. Ding, J.L., M.C. Wu and T. Tiyip, 2011. Study on soil salinization information in arid region using remote sensing technique. Agric. Sci. China, 10: 401-411.
9. Sadiq, A. and F. Howari, 2009. Remote sensing and spectral characteristics of desert sand from Qatar Peninsula, Arabian/Persian Gulf. Remote Sens., 1: 915-933.
10. Lasaponara, R., 2006. On the use of Principal Component Analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series. Ecol. Modell., 194: 429-434.
11. ERDAS Inc., 1994. Erdas Imagine Field Guide. 3rd Edn., Erdas Inc., Atlanta, GA., USA.
12. Milne, A.K., 1988. Change detection analysis using Landsat imagery a review of methodology. Proceedings of the IGARSS`88 Symposium, Sept. 13-16, Edinburgh, Scotland, pp: 541-544.
13. Richards, J.A., 1986. Remote sensing Digital Image Analysis. Springer, New York, USA., ISBN-13: 9783642300622, pp: 127-138.
14. Eklundh, L. and A. Singh, 1993. A comparative analysis of standardised and unstandardised principal components analysis in remote sensing. Int. J. Remote Sens., 14: 1359-1370.
15. Chuvieco, E., 2010. Teledeteccion Ambiental: La Observacion de la Tierra Desde el Espacio. Editorial Ariel S.A., Barcelona, Spain, ISBN-13: 9788434434981, Pages: 528.
16. Estornell, J., J.M. Marti-Gavila, M.T. Sebastia and J. Mengual, 2013. Principal component analysis applied to remote sensing. Modell. Sci. Educ. Learn., 6: 83-89.
17. Connese, C., G. Maracchi, F. Miglietta and F. Maselli, 1988. Forest classification by principal component analysis of TM data. Int. J. Remote Sens., 9: 1597-1612.
18. Bonyad, A., 2005. Multitemporal satellite image database classification for land cover inventory and mapping. J. Applied Sci., 5: 834-837.
19. Holden, H. and E. LeDrew, 1998. Spectral discrimination of healthy and non-healthy corals based on cluster analysis, principal components analysis and derivative spectroscopy. Remote Sens. Environ., 65: 217-224.
20. Zhao, G. and A.L. Maclean, 2000. A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. Photogramm. Eng. Remote Sens., 66: 841-847.
21. Jia, X. and J.A. Richards, 1999. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. IEEE Trans. Geosci. Remote Sens., 37: 538-542.
22. Eastman, J.R. and M. Fulk, 1993. Long sequence time series evaluation using standardized principal components. Photogramm. Eng. Remote Sens., 59: 991-996.