FEATURE EXTRACTION AND CLASSIFICATION OF HYPERSPECTRAL DATA OF MONGOLIA USING MACHINE LEARNING METHODS

Damdinsuren Amarsaikhan[1], Amarsaikhan Enkhmanlai[2*], Tsedev Bat-Erdene[3]
Enkhtuya Jargaldalai[1], Chogsom Bolorchuluun[4]
[1]Institute of Geography and Geoecology, Mongolian Academy of Sciences
[2]Oyuny tsomorlig impex LLC, Ulaanbaatar, Mongolia
[3]Department of Geography, Mongolian National University of Education
[4]Department of Geography, National University Mongolia
Email(s): *Corresponding author: manlai.lp@gmail.com, amarsaikhan@mas.ac.mn

KEY WORDS: *Feature extraction, hyperspectral image, machine learning classification*

ABSTRACT: Generally, hyperspectral datasets have a number of advantages compared to multichannel datasets for the identification of the Earth's surface features. As, generating accurate land cover maps using digital images is one of the most important thematic applications in remote sensing (RS), various types of feature extraction and hyperspectral image classification techniques have been developed for the differentiation of land cover classes. Although, there are many different techniques for the feature extraction, principal component analysis (PCA) and minimum noise fraction (MNF) transformation still have wide applications in dimensionality and noise reduction of various types of multidimensional images. The aim of this study is to classify the features derived from hyperspectral data of western Mongolia using machine learning methods. As a data source, a 242 band Hyperion image from August 2015 is used. For the feature extraction from the Hyperion data, the PCA and MNF transformation are applied. Overall, the research indicates that machine learning methods can be effectively used for the separation of the land cover classes along with the selected feature extraction techniques.

1. INTRODUCTION

As it is known, hyperspectral imaging measures the spatial and spectral features of the Earth's objects at different wavelengths ranging from the ultraviolet through visible and near infrared to the middle infrared spectrum (Amarsaikhan et al. 2012). Unlike the traditional multispectral imaging, which uses only several types of sensors sensitive to the optical range of the electro-magnetic spectrum, hyperspectral images usually include hundreds of narrow band channels. Therefore, datasets acquired by hyperspectral sensors can be successfully used for the differentiation of the land surface features that have very similar spectral properties/signatures (Amarsaikhan et al. 2021). However, as it provides so much spectral information and too many spectral bands, many of the existing remotely acquired image analysis approaches may confront a very serious challenge.

Although, different types of feature extraction and hyperspectral image classification techniques have been developed for the differentiation of land cover classes, the PCA still has wide applications in feature extraction and dimensionality reduction of various types of remotely acquired images (Shi et al. 2022). In recent years, MNF transformation has been used for feature extraction along with other advanced methods (Luo et al. 2016, Li et al. 2018). The MNF method computes the normalized linear combinations of the original bands that maximize the ratio of the signal to noise. It was developed for the analysis of multichannel images that can produce orthogonal components ordered by their information content (Berman et al. 2012).

The aim of this study is to evaluate the features derived from 242 band Hyperion data of western Mongolia using two machine learning techniques like support vector machine (SVM), and random forest (RF). For the feature extraction from the hyperspectral image, the PCA and MNF transformation have been applied. Before applying the feature extraction process, the Hyperion bands had been analyzed in terms of radiometric quality, and poor quality bands were excluded. After performing the PCA and MNF transformation, each of the resulting outputs was analyzed, and the components that contained reasonable amount of information were selected to form the final features. Then, the selected features were evaluated using the SVM, and RF techniques. Overall, the research indicated that both machine learning methods could be successfully used for the differentiation of the land cover classes along with the selected feature extraction techniques.

2. TEST SITE AND DATA SOURCES

As a test site, the Sangiin dalai nuur area has been selected. The area is situated in Uvs aimag, western Mongolia in between Khyargas nuur and Khar-Us nuur. In terms of physical geography, the area belongs to a dry-steppe and arid zone, and its land cover mainly includes such types as soil, light soil, sand, lake water and its surrounding vegetation. Although, the entire region is spread over an area of more than 1000sq.km, the study area chosen for the present study

extends from the west to the east about 7.05km and from the north to the south about 16.8km. The test area has low annual precipitation, hot summer, and cold winter.

In the present study, we used Hyperion data of August 04, 2015. Hyperion is a high resolution hyperspectral imaging system that can acquire images of the Earth's features in 242 spectral bands, covering the low optical range from 0.4 µm to 2.5 µm with a pixel resolution of 30 m (EO-1, 2017). For the analysis, the original Hyperion dataset was reduced from 242 bands to 150 after the bands with totally zero values were excluded. In addition, a large scale topographic map and other ground truth data were available. Figure 1 shows a location of test area and its illustration in a Hyperion image frame.

a) b)

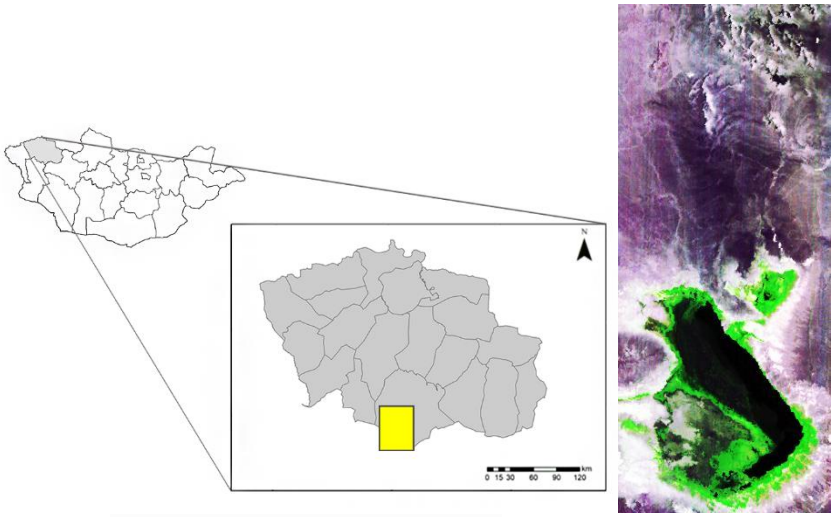

Figure 1. Location of study area (a), Hyperion image (b).

3. MACHINE LEARNING AND FEATURE EXTRACTION METHODS

Machine learning is a data analytics technique which uses different algorithms for building mathematical models and making predictions using historical data or information. It offers a huge potential for effective classification of RS data and land cover discrimination. The main advantages of machine learning in image classification include the capacity to classify data with complex characteristics, modelling of complex class signatures, and capability to handle different types of input predictor data (Maxwell et al. 2018). In the current study, we used the SVM and RF methods for classification of land cover types in a hyperspectral image.

SVM is a set of related supervised learning methods popular for performing classification of different features and regression analysis using data analysis and pattern recognition. Methods vary on the structure and attributes of the classifier, and more accurate definition would state that it builds a set of hyperplanes to classify all inputs in a high-dimensional or even infinite space. The closest values to the classification margin are known as support vectors (Gove and Faytong, 2012). Due to its strong theoretical foundation, good generalization capability, low sensitivity to the curse of dimensionality, and ability to find global classification solutions, SVM has been proved to perform well for classifying hyperspectral data (Fei, 2020).

RF classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation. The decision tree is a hierarchical structure that is built using the independent variables of dataset (Suthaharan, 2016). Each node of the decision tree is split according to a measure associated with a subset of the features. Bootstrapping indicates that several individual decision trees are trained in parallel on various subsets of the training dataset using different subsets of available features (Misra and Li, 2020). It deals with voting that has the effect of correcting for the undesirable property of decision trees to overfit training data. In the training stage, bagging is applied to individual trees in the ensemble and repeatedly selects a random sample with replacement from the training set and fits trees to these samples (Caie et al. 2021).

PCA is a data compression technique used to reduce the dimensionality of the multidimensional datasets. When the PCA is performed, the axes of the spectral space are rotated, changing the coordinates of each pixel in spectral space. The new axes are parallel to the axes of the ellipse. The transect, which corresponds to the major axis of the ellipse, is called the first principal component of the data. The direction of the first principal component is the first

eigenvector, and new axis of the spectral space is defined by this first principal component. In *n* dimensions, there are *n* principal components. Each successive principal component is the widest transect of the ellipse that is orthogonal to the previous components in the *n*-dimensional space (Munkh-Erdene et al. 2021).

MNF is a well-known technique for denoising of hyperspectral images. It transforms a noisy multidimensional data cube into output channel images with steadily increasing noise levels (Luo et al. 2016). Algorithm in the MNF consists of two consecutive data reduction operations. The first is based on an estimation of noise in the data as represented by a correlation matrix. This transformation decorrelates and rescales the noise in the data, by variance. At this stage, the information about between band noise has not been considered. The second operation accounts for the original correlations, and creates a set of components that contain weighted information about the variance across all bands in the raw dataset. The algorithm retains specific channel information because all original bands contribute to each of the components weighting (Boardman, 1993).

4. RESULTS AND DISCUSSION

Initially, the Hyperion bands have been analyzed in terms of the radiometric quality. It was revealed that the water absorption bands and some other bands of the original data had zero values. When these bands have been excluded, the original dataset was reduced from 242 bands to 150 bands. Then, hyperspectral channels were geometrically corrected to a UTM map projection using a topographic map of the study area, scale 1:100,000. The ground control point have been selected on clearly delineated areas around the Sangiin dalai nuur and some other topographic elements. In total 9 points were selected. For the transformation, a second order transformation and nearest neighbor resampling approach have been applied and the related root mean square error was 0.97 pixel.

For the feature selection, the following two approaches have been used:

a) PCA has been performed using all 155 bands and the result showed that the first three principal component (PC)s contained 98,67% of the overall variance (72.08%, 23.26%, 3.53% for the PC1, PC2, and PC3, respectively). The visual inspection of PC4 indicated that it contained noise. However, visual inspection of PC5 showed that though it contained only 0.29% of the overall variance, it still included some useful information. Colour images created by the use of the PC123 and PC135 are shown in Figure 2.

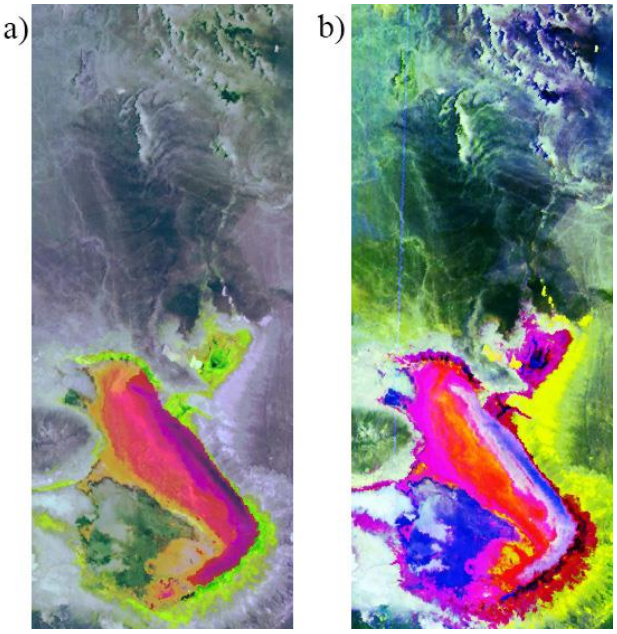

Figure 2. PC123 image (a), PC135 image (b).

b) In the MNF transform, we have initially selected the first 9 bands from the available 150 bands, considering the bands with large eigenvalues (i.e. eigenvalues greater than 1 contain some information, and eigenvalues close to 1 contain noise). Then, visual inspections have been carried out on the selected MNF bands. Although the eigenvalues of all 9 bands greatly exceeded 1, the inspection of each MNF channel indicated that MNF 1,2,3, and 7 contained useful information, clearly delineating the selected land cover classes. Colour images created by the use of the MNF123 and MNF137 are shown in Figure 3.

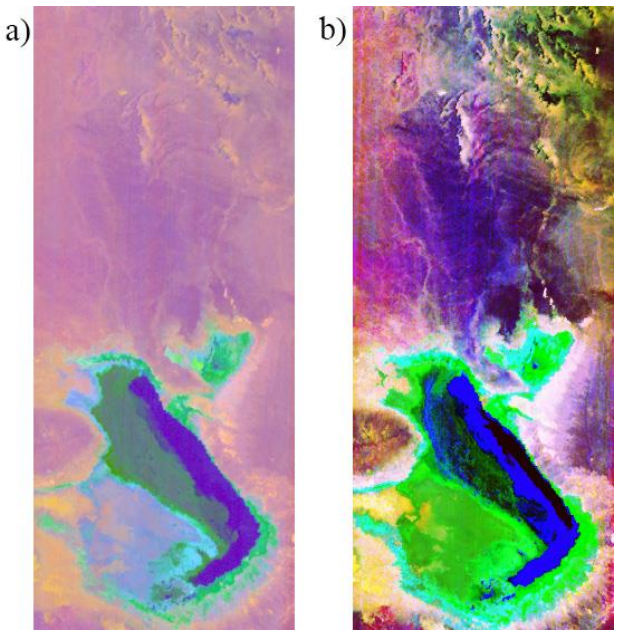

Figure 3. MNF123 image (a), MNF137 image (b).

After selecting the features, the training signatures of the available land cover types such as soil, light soil, sand, water, and vegetation have been selected from the Hyperion image. The separability of the training signatures was firstly checked in a feature space and then evaluated using Jeffries–Matusita distance. The values of Jeffries–Matusita distance range from 0 to 2.0 and indicate how well the selected pairs are statistically separate. The values greater than 1.9 indicate that the pairs have good separability (Amarsaikhan et al. 2012). After the investigation, the samples that demonstrated the greatest separability were chosen for the final signatures. The final signatures included about 157–283 pixels.

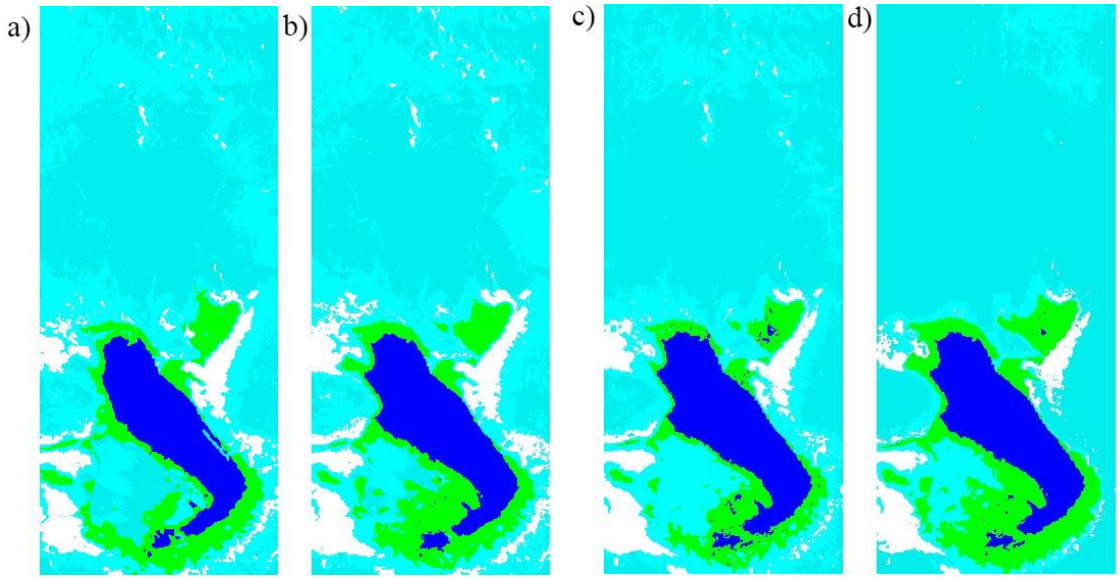

Figure 4. Classification results: (a) RF-PC123, (b) RF-MNF123, (c) SVM-PC123, (d) SVM-MNF123.

For the actual classification, RF and SVM methods have been used. After the classifications, the results have been analyzed in terms of accuracy and separation of the selected land cover types. In the RF classification, both PC123 and MNF123 features illustrated good results. However, the performances of the PC135 and MNF137 features demonstrated poor-quality, because they showed high overlaps among the statistically similar classes. In the SVM classification, PC123 demonstrated better result compared to the result (insufficient quality of output) of MNF123 features. Nonetheless, it was worse than the output of the RF technique, because the result contained a very high overlap between soil and light soil classes. Likewise, the results of the PC135 and MNF137 features were not sufficient to be considered for further analysis. The selected classified images are shown in figure 4a-d.

For the accuracy assessment of the classification results, the overall performance has been used. This approach creates a confusion matrix in which reference pixels are compared with the classified pixels and as a result an accuracy report is generated indicating the percentages of the overall accuracy (Amarsaikhan *et al.* 2012). As ground truth information, different AOIs containing 1893 purest pixels have been selected. AOIs were selected on a principle that more pixels to be selected for the evaluation of the larger classes such as soil and light soil rather than the smaller classes such as sand and vegetation. The overall classification accuracies for the selected classes were 93.12% and 90.05% for the PC123 and MNF123 features using RF classification; and 87,19% and 80.09% for the selected features using SVM technique.

5. CONCLUSIONS

At present, for generating accurate land cover maps using hyperspectral datasets different feature extraction and image classification methods could be applied. The main aim of this study was to evaluate the features derived from Hyperion data of Mongolia using two machine learning techniques. For the feature extraction from the hyperspectral image, the PCA and MNF transformation were selected, while for the classification SVM, and RF techniques were chosen. As could be seen, the PC123 and MNF123 features showed good results, however, the performances of the PC135 and MNF137 features were insufficient. The outputs of these feature combinations contained high overlaps among the statistically similar classes. Overall, the research indicated that both machine learning methods could be successfully used for the differentiation of the land cover classes along with the selected feature extraction techniques, but the RF method could deliver more accurate outputs.

REFERENCES

Amarsaikhan, D., Enkhmanlai, A., Battsengel, V. and Batsaikhan, V., 2012, Feature extraction and classification of hyperspectral images, *Full paper published in CD-ROM Proceedings of the ACRS*, Pattaya, Thailand.

Amarsaikhan, D., Byambadolgor, B., Jargaldalai, E., Enkhjargal, D., Tsogzol, G., 2021, Application of hyperspectral data for land cover classification, Advances in Engineering Research, Vol.206, Proceedings of the ESTIC 2021, Atlantis Press-Springer Nature, pp.86-90.

Berman, M., Phatak, A. and Traylen, A., 2012, Some invariance properties of the minimum noise fraction transform, Chemom. Intell. Lab. Syst. Vol.117, pp.189-199.

Boardman, J.W., 1993. Automating Spectral Unmixing of AVIRIS DATA Using Geometry Concepts, In Summaries of the Fourth Annual JPL Airborne Geoscience Workshop, JPL Publ. 93-26, Vol.1, Jet Propulsion Laboratory, Pasadena, CA, pp.11-14.

EO-1, 2017, EO1 Trajectory Details, National Space Science Data Center *of* NASA, *Available at:* https://nssdc.gsfc.nasa.gov/.

Caie, P., Dimitriou, P. and Arandjelović, O., 2021, Chapter 8 - Precision medicine in digital pathology via image analysis and machine learning, Editor(s): Stanley Cohen, Artificial Intelligence and Deep Learning in Pathology, Elsevier, pp.149-173, ISBN 9780323675383.

Fei, B. 2020, Chapter 3.6 - Hyperspectral Imaging in Medical Applications, J.M. Amigo (Ed.), Data Handling in Science and Technology, Hyperspectral Imaging, Vol. 32, Elsevier (2020), pp. 523-565, 10.1016/B978-0-444-63977-6.00021-3.

Gove R., Faytong J., 2012. Machine learning and event-based software testing: classifiers for identifying infeasible GUI event sequences, Advances in Computers, Vol. 86, Elsevier (2012), pp. 109-135.

Li, Y., Meng, Y., Lu, Y., Wang, K., Xie, B., Cheng, Y. & Zhu, K., 2018, Noise removal for airborne time domain electromagnetic data based on minimum noise fraction, Exploration Geophysics, 49:2, 127-133, DOI: 10.1071/EG15072.

Luo, G., Chen, G., Tian, M., Qin, K. & Qian, S., 2016, Minimum noise fraction versus principal component analysis as a preprocessing step for hyperspectral imagery denoising, Canadian Journal of Remote Sensing, 42:2, 106-116, DOI: 10.1080/07038992.2016.1160772.

Maxwell, A. E., Warner, T. A. & Fang. F., 2018. Implementation of Machine-learning Classification in Remote Sensing: An Applied Review. *International Journal of Remote Sensing* 39 (9): 2784–2817.

Misra, S. and Li, H., 2020. Chapter 9 - Noninvasive fracture characterization based on the classification of sonic wave travel times, Editor(s): Siddharth Misra, Hao Li, Jiabo He, Machine Learning for Subsurface Characterization, Gulf Professional Publishing, pp.243-287, ISBN 9780128177365.

Munkh-Erdene, A., Amarsaikhan, D., Odontuya, G., Enkhjargal, D., E.Jargaldalai, 2021, Feature extraction approach in hyperspectral data, Advances in Engineering Research, Vol.206, Proceedings of the ESTIC 2021, Atlantis Press-Springer Nature, pp.102-108.

Nyamjargal, E., Batbileg, B., Amarsaikhan, D., 2021, Application of random forest approach to biomass estimation using remotely sensed data, Advances in Engineering Research, Vol.206, Proceedings of the ESTIC 2021, Atlantis Press-Springer Nature, pp.109-115.

Shi, H., Yang, Y., Wang, L. and Cao, J., 2022, Two-dimensional functional principal component analysis for image feature extraction, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2035738.

Suthaharan, S. 2016. A cognitive random forest: An intra- and intercognitive computing for big data classification under cune condition. Handb. Stat. 35, pp.207–227