Davaasuren Bayarmagnai, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Land Cover Classification Using a Random Forest Model

Davaasuren Bayarmagnai, Professor Bayanjargal Darkhijav, Professor Tsolmon Renchin

Department of Applied Mathematics[1,2]
Department of Physics[3]

Abstract- Earth's surface forms the outermost layer of our planet, and it changes due to both natural processes and human activities. Therefore, the classification of land surface features is essential for many environmental applications and serves as the basis for survey studies. Various methods are used for land cover classification using satellite data, such as random forest classification and the decision tree method of machine learning. In this study, we used the Random forest(RF) classification, which shows robustness and provides high accuracy compared to other image classification methods in remote sensing. The study area is in Khangal soum of Bulgan province of Mongolia, a forest-steppe zone with mountains and hills. Land cover types of the study area include bare land, forest, and grass. Spectral bands of Blue, Green, Red, and Near Infrared(NIR) of Landsat 8 data and ground observation data were used in the research. A confusion matrix was obtained by comparing the results obtained by the random forest method with the ground observation values, and the result was 86.4 percent. Using the results, we applied Random forest results to create a land cover map for 2017-2021. However, larch forests are estimated to be the most significant percentage in the study area. RF can be applied to different forest classifications in any forested region of Mongolia to save time and money.

Keywords- Land cover classification, Random Forest (RF)

I. INTRODUCTION

In recent years, rapid and uncontrolled population growth has multiplied the rate of land use/land cover change(LULCC) due to economic and industrial development, especially in developing countries (Swapan Talukdar et al. 2020). In this regard, user-friendly programming tools, high-level user computing power, and free satellite data are available to represent land cover changes (Abdulhakim Mohamed Abdi 2019). Remote sensing(RS) is widely used to accurately update and improve space data using powerful Geographic Information System(GIS) applications. G. Byambakhu, a Professor at the National University of Mongolia, defined GIS in one of his presentations as "GIS is a complex computer system that collects, analyzes, models, and visualizes data on the human social environment on Earth and space" (G. Byambakhuu 2013). Likewise, in the book "Introduction to Remote Sensing", Remote Sensing is defined as a science that allows collecting information about the Earth's surface without contacting the Earth's surface (S. Jenicka 2021 p 1-16). It is important that we process remote sensing data using the latest advanced methods of GIS and further apply that research. On the other hand, with the rapid increase in the amount of satellite data and the variety of applications, there is a need to constantly update and improve them, speed up processing, and automate them. Therefore, researchers are still looking for new methods in addition to the current methods of satellite data

processing. Then, artificial intelligence(AI) and machine learning(ML) methods and techniques are trying to solve these problems with the help of computers in a human-like way (Sisodiya et al. 2020). Also, ML methods are effective methods for solving RS and geoscience problems, and are widely used for classification on satellite data. ML methods are divided into supervised and unsupervised, and unsupervised method are similarly divided into clusters. However, the supervised method is divided into classification and regression (David et al. 2016; Sarker 2021). For instance, Chuny Zhong and other researchers used supervised ML methods to map the phytoplankton layer of marine ecosystems on large amounts of LiDAR-generated data. Furthermore, ML algorithms with automatic recognition of deep phytoplankton layers were studied to reduce the burden of manual tasks. These methods used Support Vector Machine(SVM), Llinear Discriminant Analysis(LDA), neural network, decision tree and RUSBoost. The results of the study showed that the RUSBoost method showed more accurate results and the neural network method processed information faster (Chunyi Zhong et al. 2024). Also, ML methods have been applied to soil quality research in the agricultural sector (Freddy et al. 2022). And as the mining industry grows, machine learning is being used in the industry to make geologists' explorations more efficient. In doing so, developmental methods are used to map the ground cover of mineral prospects. (Hojat Shirmard et al. 2022). For the air temperature change study, a method was developed to calculate the regularity of monthly average surface temperature changes in Mongolia using MODIS satellite data and land surface temperature (LST) time series data. (Otgonbayar et al. 2019). Researchers are increasingly applying machine learning (ML) methods to land cover classification. In 2020, several researchers jointly produced a map of the Bogd Khan Mountain forest area in Mongolia using Sentinel-2A satellite data (Nyamjargal et al. 2020). In 2021, Batnyam conducted land cover classification in Uvs Province, Mongolia, using Landsat 8 satellite data and the Support Vector Machine (SVM) classification method, demonstrating that the SVM method performs better in arid and semi-arid mountainous regions (Batnyam 2021). Classifying land surface features is crucial for various environmental applications and forms the basis for survey studies. For example, in their 2020 research article, A. Kalaivani and Rashmita Khilar classified agricultural land cover in India using satellite data. They highlighted the significance of accurately mapping agricultural regions and crop types, given the importance of agriculture in India. The researchers employed several supervised ML classifiers, including J48, B-Tree, J48 graft, NB tree, Random Forest (RF), and LAD tree, with the RF classifier yielding the best results (Kalaivani & Khilar 2020).Numerous studies have emphasized the effectiveness of RF for land cover classification and change detection. The RF method, known for its high accuracy and suitability for large datasets, constructs a hierarchical structure based on primary data attributes or similarities (J. Enkhtuya et al. 2022). In 2020, Thanh Noi Phan used Google Earth Engine (GEE) software and Landsat 8 satellite data for land cover classification in Mongolia, achieving promising results with the RF method (Thanh Noi Phan 2020). Additionally, some researchers have applied deep learning (DL) methods to land use/land cover change classification. For instance, Ce Zhang et al. (2019) used a joint deep learning approach. Others, like Subhra Swetanisha et al. (2022), have employed SVM and Extreme Gradient Boosting (XGBoost) for classification in land use/land cover change studies. Despite the variety of ML methods applied to land cover classification, studies specifically using the RF classification method are relatively sparse. Therefore, this research aims to classify land cover in the Khangal Sum area of Bulgan Province, Mongolia, using the RF classification method on Google Earth Engine cloud technology.

II. MATERIALS AND METHODS

1. Study Area

The study area of the research is Khangal Soum, shown in Figure 1, one of the main sounds of the Bulgan Province, located in the northern part of Mongolia. It was established in 1924. The territory of Soum is mainly a steppe region with mountains and hills, suitable for livestock, agriculture, and

tourism. Khangal sum of Bulgan province has an area of 165.6 thousand hectares, belonging to the Khangai mountainous region in the central part of Mongolia. The territory belongs to the eastern part of Bulgan province and is bordered by Orkhon province to the west, Bugat soum to the north, Selenge soum, and Baruunburen soum in Selenge province. The study area mainly consists of the four land classes: Larch, Birch, Vegetation, and Bare. Larch and Birch have different spectral values suitable as inputs for the spectral mixture model. The model allows different spectral values for vegetation, larch, and birch.

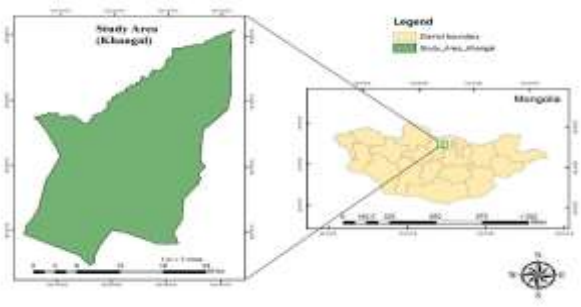

Figure 1. Study area: Khangal Soum in Bulgan province

2. Satellite Data Used

In this survey, we used the Satellite 8 in the series of Landsat satellite data 2018, downloaded from the US Geological Survey (USGS) home page. In Remote sensing data, Landsat 8 image data files consist of 11 spectral bands with a spatial resolution of 30 meters for Coastal aerosol, Blue, Green, Red, Near Infrared(NIR), Shortwave Infrared (SWIR) 1, Shortwave Infrared (SWIR) 2 and Cirrus bands. The cirrus band is beneficial for cirrus cloud detection. The resolution for the Panchromatic band is 15 meters. Thermal Infrared (TIRS) 1 band and Thermal Infrared (TIRS) 2 band help provide more accurate surface temperatures and are collected at 100 meters. The approximate scene size is 170 km north-south by 183 km east-west. The instruments on Landsat 9 are improved copies of those on Landsat 8 (https://www.usgs.gov/). Our study utilized Blue, Green, Red, and Near Infrared(NIR) (Table 1).

Table 1. Spectral bands of the Landsat 8.

Spectral bands		Wavelength (micrometer)	Resolution (meters)
SR_B2	Band 2 (blue) surface reflectance	0.452-0.512 µm	30
SR_B3	Band 3 (green) surface reflectance	0.533-0.590 µm	30
SR_B4	Band 4 (red) surface reflectance	0.636-0.673 µm	30
SR_B5	Band 5 (near infrared) surface reflectance	0.851-0.879 µm	30

3. Training and Validation Data Set

As previously mentioned, we categorized the study area's ground cover into four dominant types: Larch, Birch, Bare Ground, and Vegetation. All training and validation data were collected from field measurements taken between 2016 and 2018, and supplemented with high-resolution images from Google Earth Engine. A total of 427 points were selected for the training data, comprising 118 points in vegetation areas, 132 points in bare ground areas, 101 points in larch areas, and 76 points in birch areas. Of these, 70% were used for training data and 30% for validation data.

4. Methodology

Our research was conducted using data from Landsat 8 satellites, covering the period from June to September 2018. This timeframe corresponds to the growing season, making it ideal for our study.

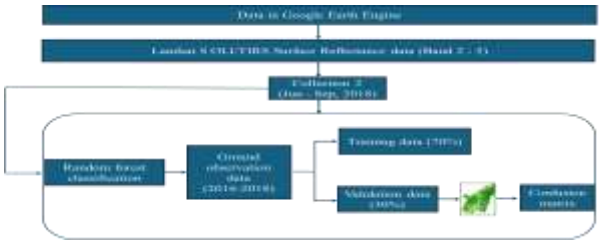

Figure 2. Schema of methodology

5. Random Forest Classification

A random decision tree method consists of a large number of decision trees, each with its own characteristics, and is shaped like a tree to show them more clearly. Moreover, this method is a machine-learning algorithm that combines several tree classifiers. Each tree votes for the most popular class, and the final result is determined by combining these votes (Liu et al. 2012 pp 246–252) (Figure 3). However the decision tree consists of root notes, decision notes, and leaf notes. So which will be the primary node? The question arises, and this problem is answered by the Gini index (1). If all ⟦pˆ¬⟧_mkvalues are zero or very close to one, the Gini index is very low. For this reason, the Gini index is used to determine the cleanliness of a node. A low value of the Gini index indicates that the node is dominated by observation values of the same class(1). Another index similar to the Gini index is entropy. Entropy and the Gini index are almost identical in numerical terms.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right) \qquad (1)$$

Then Bagging is a vital part of RF. This creates new subsets of the original data set from the original data, which is called Bootstrapping. Clustering reduces variation and improves accuracy by allowing hundreds or thousands of decision trees to be combined together (2) (Anshul, 2022), (James et al., 2013).

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \qquad (2)$$

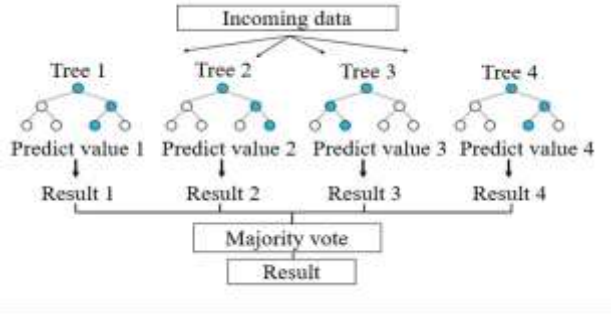

Figure 3. Random forest classification

The advantages of the random forest method are that it is easy to interpret, closer to decision making, and can easily predict quality variables without using dummy variables. However, the disadvantage is that this method does not provide sufficient accuracy compared to other classification methods, and the final result may change even if the data is changed slightly.

III. RESULTS

In our study, we conducted a Random forest classification method using Google Earth Engine (GEE). We selected four land cover classes based on the characteristics of the study area: larch, birch, vegetation, and bare land. After classification, we obtained results that were further evaluated using a confusion matrix, which compares the field measurement pixel values with the values of the classified pixels (Figure 4).

Figure 4. Result of random forest in the GEE

In our study, we employed a Random Forest classification method using Google Earth Engine (GEE). We specifically identified four land cover classes based on the characteristics of the study area: larch, birch, vegetation, and bare land. Following the classification process, we evaluated the results using a confusion matrix. This matrix compares the pixel values obtained from field measurements with those classified by our model (Figure 4).

Table 2. Confusion matrix for the RF in GEE

		Ground observation data					
		Larch	Birch	Vegetation	Bare land	Total	
d value	Larch	31	3	3	1	38	
	Birch	4	12	1	0	17	
	Vegetation	3	0	23	0	26	
	Bare land	0	2	0	42	44	
	Total	38	17	27	43		
	Overall accuracy				86.4%		

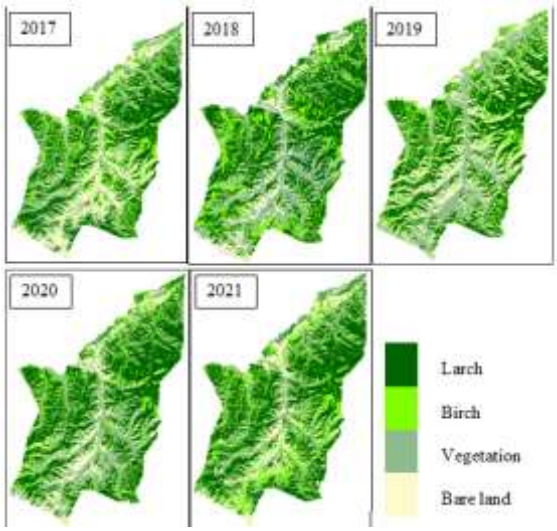

Figure 5. Land cover change map over the years 2017-2021 in the Khangal soum in Bulgan province

The figure below illustrates that larch occupies a significant area consistently across all years. Conversely, birch covers the smallest percentage of the area. Analysis of fitted values reveals that bare land constitutes an average of 27% of the total land cover. Vegetation, on the other hand, occupies approximately 25% of the entire area of Khangal Soum (Figure 6).

Figure 6. Land cover change for the study area during 2017-2021

IV. DISCUSSION

In this study, we used the Random forest method on Google Earth Engine to classify the land cover, and the overall accuracy was 86.4%. As mentioned earlier, the study of land cover classification forms the basis of most research. For example, it is important that the choice of crop area in agriculture in the future depends on LCLU, and it is used to predict how the choice of crop area in agriculture will change in the future due to LCLU (Eduardo et al, 2019).

In 2022, Researcher Amarsaikhan and other researchers worked together to classify land cover using machine learning classification methods such as Random forest and Support vector in multi-zonation data from Sentinel-2 satellite in Dzunburen Sum area of Selenge province. As a result of the research, it was concluded that the Random forest method and Support vector method, which are machine learning methods for processing modern remotely sensed image information, are effective in obtaining accurate and thematic information on land cover and land use (Jargaldalai, 2022). Compared to this research, our research has the advantage of using Google earth engine software for classification, although only Random forest method is used.

However, in future research, it is concluded that classification using more accurate satellite data and other machine learning methods will give better results.

V. CONCLUSION

Land use and land cover change (LUCC) significantly impacts climate. For instance, urban expansion leads to increased urban heat emissions, while deforestation disrupts the carbon cycle and raises global temperatures (Bununu et al., 2023). Therefore, optimizing land cover classification is crucial. Applying satellite-based classification in remote and inaccessible areas offers economic benefits compared to field visits. As highlighted in the Introduction, satellite data enables efficient mapping of mineral prospects (Hojat Shirmard et

al., 2022). In this study, we utilized Landsat 8 satellite data to classify land cover in the Khangal Soum area of Bulgan Province. Using the Random Forest method, we compared the classification results with field measurements using a confusion matrix. The land cover was categorized into four primary classes: larch, birch, vegetation, and bare land. The confusion matrix yielded an overall accuracy of 86.4%. Our findings demonstrate that the Random Forest method implemented in Google Earth Engine (GEE) is effective for land cover classification. Future research in this field should consider integrating high-resolution satellite data with other methods to enhance research outcomes.

Acknowledgment

This work was conducted under project P2019-3753, supported by the National University of Mongolia. We gratefully acknowledge our colleagues at the NUMITC-UNESCO Laboratory for Remote Sensing and Space Science for providing field measurement data from Khangal Soum, Bulgan province.

REFERENCES

1. A.Kalaivani, Rashmita Khilar., 2020. Crop Classification and Mapping for Agricultural Land from Satellite Images DOI:10.1007/978-3-030-24178-0_10
2. Abdulhakim Mohamed Abdi., 2019. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data https://doi.org/10.1080/15481603.2019.1650447
3. Anshul Saini., 2022. An Introduction to Random Forest Algorithm for beginners. Data Science Blogathon https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/
4. Batnyam, Ts, 2020. Using the help vector method for land cover classification (on the example of Uvs province) http://portal.igg.ac.mn/dataset/ra3pbih-bypxebhnnh-ahrnjiajid-tycjiax-bektopbih-aprbir-awnrjiax-hb/resource/7b450265-1bf8-4b6a-b1b9-38e36a4c017a
5. Bununu, Y.A., Bello, A., Ahmed, A., 2023. Land cover, land use, climate change and food security. Sustain Earth Reviews 6, 16 (2023). https://doi.org/10.1186/s42055-023-00065-4
6. Ce Zhang., Isabel Sargent., Xin Pan., Huapeng Li., Andy Gardiner., Jonathon Hare., Peter M. Atkinson., 2019 Joint Deep Learning for land cover and land use classification https://doi.org/10.1016/j.rse.2018.11.014
7. Chunyi Zhong., Peng Chen., Siqi Zhang., 2024. Enhancing Subsurface Phytoplankton Layer Detection in LiDAR Data through Supervised Machine Learning Techniques .https://doi.org/10.3390/rs16111953
8. David J. Lary., Amir H. Alavi., Amir H. Gandomi., Annette L.Walker., 2016. Machine learning in geosciences and remote sensing https://doi.org/10.1016/j.gsf.2015.07.003
9. Freddy A. Diaz-Gonzalez., Jose Vuelvas., Carlos A. Correa., Victoria E. Vallejo., D. Patino., 2022. Machine learning and remote sensing techniques applied to estimate soil indicators – Review https://doi.org/10.1016/j.ecolind.2021.108517
10. G. Byambakhuu., 2021. Urban intelligence in development policy the problem of reflecting research concepts. Department of Geography, National University of Mongolia http://nda.gov.mn/backend/files/Q0yWsVJ1Q6w0oAP.pdf
11. GIS definition, 2021. The problem of incorporating the concept of urban studies into development policy http://nda.gov.mn/backend/files/Q0yWsVJ1Q6w0oAP.pdf
12. Hojat Shirmard., Ehsan Farahbakhsh., R. Dietmar Müller., Rohitash Chandra., 2022. A review of machine learning in processing remote sensing data for mineral exploration https://doi.org/10.1016/j.rse.2021.112750
13. James G, Witten D, Hastie T, & Tibshirani R., 2013 "Introduction to Statistical Learning with applications in R", Springer.
14. Jargaldalai Enkhtuya., Amarsaikhan Damdinsuren., Bilgun Ulziibat., Munkh-Erdene Altangerel., 2022. Land cover classification

using machine-learning method and vegetation indices. Mongolian Geography-Geoecology Magazine DOI:10.5564/mjgg.v59i43.2532
15. Jenicka, S., 2021. Introduction to Remote Sensing. In: Land Cover Classification of Remotely Sensed Images. Springer, Cham. https://doi.org/10.1007/978-3-030-66595-1_1
16. Liu, Y., Wang, Y., Zhang, J., 2012. New Machine Learning Algorithm: Random Forest. In: Liu, B., Ma, M., Chang, J. (eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science, vol 7473. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32
17. Nyamjargal, E., Amarsaikhan, D., Munkh-Erdene, A., Battsengel, V., Bolorchuluun, C., 2020. Object-based classification of mixed forest types in Mongolia DOI:10.1080/10106049.2019.1583775
18. Otgonbayar, M., Atzberger, C., Mattiuzzi, M., Erdenedalai, A., 2019. Estimation of climatologies of average monthly air temperature over Mongolia using MODIS land surface temperature (LST) time series and machine learning techniques. https://doi.org/10.3390/rs11212588
19. Sarker, I.H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160. https://doi.org/10.1007/s42979-021-00592-x .
20. Sisodiya, N., Dube, N., & Thakkar, P., 2020. Next-generation artificial intelligence techniques for satellite data processing. DOI:10.1007/978-3-030-24178-0_11
21. Subhra Swetanisha., Amiya Ranjan Panda., Dayal Kumar Behera., 2022. Land use/land cover classification using machine learning models. DOI:10.11591/ijece.v12i2.pp2040-2046
22. Swapan Talukdar., Pankaj Singha., Susanta Mahato., Shahfahad., Swades Pal., Yuei-An Liou., Atiqur Rahman., 2020. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. https://doi.org/10.3390/rs12071135
23. Thanh Noi Phan., Verena Kuch., Lukas W Lehnert., 2020. Land cover classification using Google Earth Engine and Random forest classifier- The role of image composition. Remote Sensing https://doi.org/10.3390/rs12152411
24. Eduardo Gomes , Patrícia Abrantes , Arnaud Banos , Jorge Rocha , Michael Buxton., 2019. Farming under urban pressure: Farmers' land use and land cover change intentions. https://doi.org/10.1016/j.apgeog.2018.12.009