



Original Article

Green vegetation biomass estimation in Mongolia using RS data and machine-learning methods

Damdinsuren Amarsaikhan¹, Dalantai Sainbayar^{1*}, Enkhtuya Jargaldalai¹,
Enkhmanlai Amarsaikhan², Gendaram Odontuya³, Enkhjargal Egshiglen²,
Damdin Enkhjargal¹, Tsedev Bat-Erdene⁴

¹Division of Aerospace Remote Sensing, Institute of Geography and Geocology, Mongolian Academy of Sciences, Ulaanbaatar 14190, Mongolia

²Oyunii Tsomorlig Impex LLC, Ulaanbaatar 14200, Mongolia

³Department of Pharmaceutical Chemistry and Pharmacognosy, Mongolian University of Pharmaceutical Science, Ulaanbaatar 18130, Mongolia

⁴Department of Physical Geography, Mongolian State University of Education, Ulaanbaatar 14191, Mongolia

*Corresponding author: sainbayard@mas.ac.mn, ORCID:0000-0001-8606-6167

ARTICLE INFO

Article history:



Received: 02 April, 2026

Revised: 15 July, 2026

Accepted: 27 July, 2026

ABSTRACT

Green vegetation biomass is a key ecological indicator reflecting ecosystem productivity, rangeland condition, and responses to climate variability, particularly in arid and semi-arid environments. In Mongolia, where vast grassland and desert-steppe ecosystems support both biodiversity and pastoral livelihoods, spatially explicit biomass information is essential for ecological monitoring and conservation-oriented land management. Traditional field-based biomass measurements, however, are labor-intensive and limited in spatial coverage. In this study, we estimated aboveground biomass of green vegetation across Mongolia using moderate-resolution remote sensing data and machine learning methods to support large-scale ecological assessment. Three spectral bands and nine vegetation indices derived from Moderate Resolution Imaging Spectroradiometer MOD13A1 data from August 2021 were combined with field measurements from 275 monitoring sites. Random forest, artificial neural network, and extreme gradient boosting \ models were trained and optimized using grid-search hyperparameter tuning and evaluated through cross-validation. Among the tested models, Random forest achieved the highest predictive accuracy ($R^2 = 0.88$, $MAE = 68.45 \text{ kg ha}^{-1}$, $RMSE = 109.93 \text{ kg ha}^{-1}$), followed by extreme gradient boosting ($R^2 = 0.80$, $MAE = 87.47 \text{ kg ha}^{-1}$, $RMSE = 143.09 \text{ kg ha}^{-1}$), while artificial neural network showed comparatively lower performance. The resulting biomass maps reveal pronounced spatial variability linked to Mongolia's climatic and ecological gradients. Our findings demonstrate that machine learning methods combined with remote sensing provide an efficient framework for estimating vegetation biomass across Mongolia under the environmental conditions represented by the August 2021 dataset. The generated biomass products offer valuable ecological insights and can support rangeland monitoring, conservation planning, and assessments of ecosystem vulnerability under ongoing climate change.

Keywords: AGB, RF, ANN, XGB, pastureland ecosystems

INTRODUCTION

Vegetation plays a crucial role in the health of terrestrial ecosystems, serving as a vital component in the water cycle, energy cycle, and carbon cycle (Nyamsuren et al., 2019). Through processes such as photosynthesis,

respiration, transpiration, and surface albedo, vegetation acts as a key regulator of the water and biogeochemical cycles governing the Earth's surface. Its development is intricately linked to a myriad of natural environmental factors and human activities, reflecting the impacts of climate

change and anthropogenic influences (Meng et al., 2020). Numerous elements shape the growth of vegetation, with climate change exerting a substantial influence. Factors such as precipitation, air temperature, and relative humidity significantly affect the effective accumulation of temperature and available moisture, ultimately regulating plant photosynthesis, respiration, and the decomposition of soil organic carbon - processes that are pivotal in determining vegetation development and distribution (He et al., 2020).

In Mongolia, the progression of vegetation is fundamentally driven by the prevailing temperature and precipitation conditions (Jamiyansharav et al., 2018). The intensity and impact of climate constraints on vegetation vary distinctly, governed by diverse factors including topography, latitude, and air circulation. The climatic gradient transitions from cold semi-humid conditions in the north to a warmer, arid climate in the south, creating a clear latitudinal zonation of vegetation. This is further refined by altitudinal zonation in the mountainous regions. Moving from north to south, vegetation zones transition from the dense taiga to the arid expanses of the Gobi Desert. The taiga, sub-taiga, and fragmented forests found in the forest-steppe region represent the southern boundary of the expansive Euro-Siberian boreal forest. The extensive grasslands dominate the landscape of the Mongolian-Chinese steppe. The distributions of the various vegetation zones, types of boreal forests, and tree-lines are significantly influenced by critical factors such as air temperature, evapotranspiration, and precipitation (Klinge et al., 2018).

Vegetation biomass, defined as the total weight of plant material, is a key ecological indicator frequently measured in environmental studies (Portela et al., 2020). Closely associated with aboveground biomass (AGB), vegetation production plays a vital role in the global carbon cycle and significantly contributes to the regulation of terrestrial carbon sinks (Lin et al., 2023). However, accurately estimating biomass is particularly challenging in heterogeneous landscapes and varied environmental conditions, demanding precise and reliable measurement techniques (Kumar et al., 2015). Biomass

assessment can be carried out using traditional ground-based methods and modern remote sensing (RS) technologies (Chiarito et al., 2021). While ground-based approaches often involve labor-intensive processes and are limited in spatial coverage (Chen et al., 2021), RS-based methods offer scalable, cost-effective, and efficient alternatives for large-area monitoring and change detection (Amarsaikhan et al. 2023). In the past decade, there has been a substantial advancement in the use of RS for the estimation and mapping of AGB. The development of RS techniques has provided powerful tools for rapid assessment and continuous monitoring of biomass dynamics at various scales (Lourenco, 2021). Among the diverse RS data sources available, multiresolution optical imagery stands out due to its broad accessibility and applicability. Commonly used datasets range from low-resolution sensors like the Advanced Very High Resolution Radiometer (AVHRR) and Moderate Resolution Imaging Spectroradiometer (MODIS), to medium-resolution platforms such as Landsat, Satellite for Observation of Earth (SPOT), and Sentinel-2, and high-resolution systems including IKONOS, QuickBird, and WorldView (Amarsaikhan et al. 2023). Numerous techniques have been developed for estimating biomass using optical data, including machine learning algorithms such as Random forest (RF), support vector machines, neural networks, Cubist, K-Nearest Neighbor, and Bayesian network, as well as various statistical models like linear and nonlinear regression. Different forms of vegetation indices are frequently used in combination with these models to enhance predictive performance (Mills, 2011; Matiza et al., 2024).

In Mongolia, RS-based biomass estimation has a long history, beginning in the mid-1980s. One of the earliest studies by Adyasuren (1989) investigated the correlation between field-based biomass measurements and RS data reflectance values. Since then, various studies have adopted both traditional statistical and advanced modeling techniques to assess vegetation biomass across the country (Amarsaikhan et al. 2023). Kogan et al. (2004) utilized vegetation health indices from AVHRR to detect biomass anomalies in pasturelands. Sternberg et al. (2011) analyzed

normalized difference vegetation index (NDVI), alongside field data to examine desertification trends in a dryland region. Lamchin et al. (2016) integrated NDVI, the Topsoil Grain Size Index, and land surface albedo to evaluate land surface conditions and assess desertification in a steppe ecosystem. More recently, Otgonbayar et al. (2018) compared Partial Least Squares Regression (PLSR) and RF for estimating pasture biomass in Mongolia, concluding that PLSR offered more accurate predictions.

Environmental monitoring in Mongolia often requires rapid assessment of vegetation biomass to support timely and informed decision-making. Although high-resolution satellite datasets provide detailed information, their processing is typically time-intensive. As a more efficient alternative, moderate-resolution satellite data can offer quick, albeit less precise, estimations. This study investigates the potential of the RF, artificial neural network (ANN), and extreme gradient boosting (XGB) techniques to estimate green vegetation biomass across Mongolia under small-scale mapping conditions. The models were trained using original spectral bands and vegetation indices derived from MOD13A1 satellite data. Their performance was compared to determine the most effective method for fast and reliable biomass estimation.

Beyond its role as a biophysical variable, green vegetation biomass is a fundamental indicator of

ecosystem functioning, rangeland condition, and ecological resilience. In Mongolia's arid and semi-arid environments, spatial and temporal variability in biomass reflects the combined influences of climate variability, grazing pressure, land degradation, and ecological recovery processes. Spatially explicit biomass estimation using RS is therefore critical for ecological monitoring, conservation planning, and sustainable rangeland management at regional and national scales.

MATERIALS AND METHODS

Study area

The territory of Mongolia constitutes the northern boundary of the Central Asian drylands. Steppe vegetation predominates across much of the region, acting as a transition zone between the Siberian taiga and the Great Gobi Desert. Vegetation zones are distinctly arranged according to latitudinal and altitudinal gradients, reflecting the patterns of climatic zonation (Wesche et al., 2016). In northern Mongolia and the upper mountains of central Mongolia, mountain taiga, consisting of coniferous trees, forms continuous forested areas. Moving southward, the taiga evolves into sub-taiga, which is characterized by a mixture of coniferous and deciduous broad-leaf trees. In central Mongolia, the landscape features steppe in the basins and forest-steppe on the mountains. This region, with its semiarid climate, exemplifies a forest-steppe transition ecotone. In

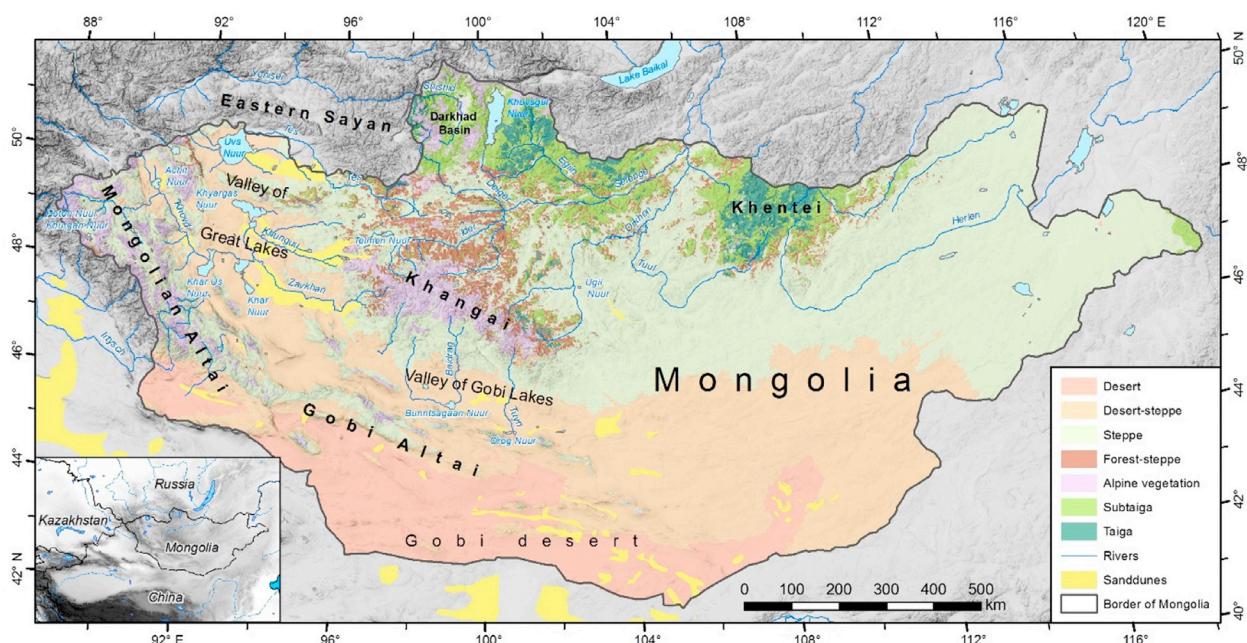


Fig. 1. Vegetation zones of Mongolia (Source: Klinge and Sauer, 2019)

the high mountains, dense alpine shrub vegetation and meadows yield to the periglacial zone of frost debris. Along rivers and areas with groundwater access through capillary rise, lush azonal grasslands and trees such as poplar, willow, and elm are predominant (Klinge and Sauer, 2019). Fig. 1 shows a map that illustrates the various vegetation zones in Mongolia.

The country can be divided into five distinct regions based on variations in climate, geomorphology, vegetation, and hydrology. These regions include: the Eastern Mongolian steppe region, characterized by a semi-humid to semi-arid climate, flat terrains, and extensive basins, featuring steppe and desert-steppe with interior drainage; the Gobi-Altai and southern Mongolian desert, which experiences a semi-arid to arid climate, encompassing wide basins and high mountains, and is home to desert steppe and desert landscapes with interior drainage; the Mongolian Altai in western Mongolia, ranging from semi-humid to arid conditions, marked by high mountains in the west and a broad basin with large interior lakes to the east, supporting steppe, forest steppe, and alpine meadows; the Khangai Mountains in central Mongolia, exhibiting a semi-humid to semi-arid climate, characterized by high mountains and intra-mountain basins, which host meadow steppe, forest-steppe, and

alpine meadows; and finally, the Khuvsgul Lake area and Khentii region in northern Mongolia, noted for its humid to semi-humid conditions, comprising mountainous landscapes with small valleys, along with taiga, forest-steppe, and steppe regions with external drainage (Klinge and Sauer, 2019).

Satellite and field data sets

In the present study, we selected a 16-day average data of Level-3 MOD13A1 of MODIS from August 2021 to estimate the vegetation biomass of Mongolia (Fig. 2). MODIS is a satellite sensor mounted on the National Aeronautics and Space Administration Terra and Aqua satellites. It continually collects data in 36 spectral channels with global coverage every 1 to 2 days. The spatial resolution of the acquired data ranges from 250 m to 1 km. Its exceptionally broad spectral range enables MODIS data to be used in studies across many different disciplines. MOD13A1 contains 12 layers, including NDVI, enhanced vegetation index, vegetation index quality indicator, blue reflectance, red reflectance, near-infrared (NIR) reflectance, short-wave infrared (SWIR) view zenith angle, sun zenith angle, and relative azimuth angle, composite day of the year and pixel reliability.

To estimate vegetation biomass using RS data, we

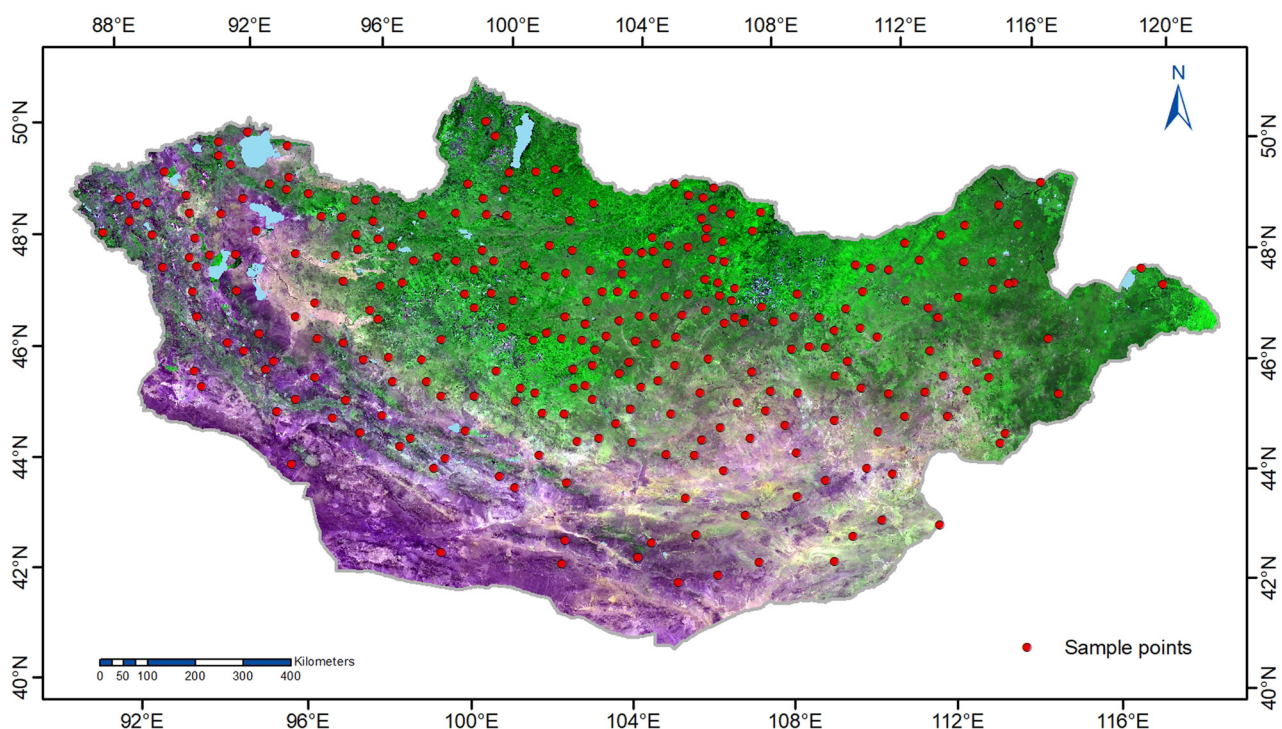


Fig. 2. Locations of the sampling sites positioned on the MODIS image

utilized both field-based biomass measurements and spectral information from satellite imagery. The field data was supplied by the Information and Research Institute of Meteorology, Hydrology, and Environment of Mongolia, which manages nationwide vegetation monitoring. They conduct annual biomass assessments across various monitoring plots using standardized techniques such as line-point intercept, gap intercept, air-dry biomass collection at a 1 cm clipping height, and photo documentation (SDC and GGP, 2015). Trained meteorological staff collect the biomass samples annually. Once collected, the samples are sealed in plastic bags and transported to meteorological stations for analysis. In the laboratory, the field-measured biomass is dried, and the dry weight is determined. This value is then divided by the surface area of the plot, and expressed in kilograms per hectare (kg ha^{-1}). Locations of the selected 275 sampling sites positioned on the MODIS image are shown in Fig. 2.

Determination of indices

Vegetation and soil characteristics are critical components in understanding terrestrial ecosystem functioning and assessing land surface biophysical parameters. Vegetation, in particular, plays a vital role in regulating energy exchanges, nutrient cycling, and hydrological processes.

RS technologies have become indispensable for monitoring these dynamics due to their ability to provide consistent, large-scale, and temporally frequent observations. Among the most effective RS tools are vegetation indices, which are mathematical transformations of spectral reflectance data designed to enhance the signal of vegetation-related variables while minimizing background noise from soil, atmospheric interference, or illumination conditions (Huete et al., 2002; Dorigo et al., 2007).

Vegetation indices are typically computed using combinations of spectral reflectance values in the visible and NIR regions of the electromagnetic spectrum. Recent advancements in sensor technology have allowed researchers to develop a diverse array of vegetation indices tailored to specific objectives, such as assessing chlorophyll content, vegetation cover, biomass, and water stress (Amarsaikhan et al., 2023). These indices serve as valuable input variables in empirical and machine learning models used to estimate vegetation characteristics at multiple spatial and temporal scales (Altanchimeg et al., 2022). In our study, we determined 9 spectral indices commonly used in vegetation analysis (see Table 1). These indices were calculated using spectral bands from the blue, red, and NIR ranges of the electromagnetic spectrum (Enkhtuya et al., 2022).

Table 1. Vegetation indices used for the study

No.	Acronym	Index	Formula
1	NDVI	Normalized Difference Vegetation Index	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$
2	ATSAVI	Adjusted Transformed Soil-Adjusted Vegetation Index	$1.22 \cdot (\rho_{\text{NIR}} - 1.22 \cdot \rho_{\text{red}} - 0.03) / (1.22 \cdot \rho_{\text{NIR}} + \rho_{\text{red}} - 1.22 \cdot 0.03 + 0.08 \cdot (1 + 1.22^2))$
3	BWDRVI	Blue-Wide Dynamic Range Vegetation Index	$(0.1 \cdot \rho_{\text{NIR}} - \rho_{\text{blue}}) / (0.1 \cdot \rho_{\text{NIR}} + \rho_{\text{blue}})$
4	EVI	Enhanced Vegetation Index	$2.5 \cdot \frac{(\text{NIR} - \text{Red})}{(\text{NIR} + 6 \cdot \text{Red} - 7.5 \cdot \text{Blue} + 1)}$
5	DVI	Difference Vegetation Index	$\rho_{\text{NIR}} - \rho_{\text{red}}$
6	kNDVI	Kernel NDVI	$\tanh(\text{NDVI}^2)$
7	TVI	Transformed Vegetation Index	$\sqrt{\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}} + 0.5$
8	OSAVI	Optimized Soil Adjusted Vegetation Index	$(1 + 1.16) \cdot \frac{\text{NIR} - \text{Red}}{\text{R} + \text{NIR} + 0.16}$
9	NNIR	Normalized Near Infrared	$\frac{\text{NIR}}{(\text{NIR} + \text{Red} + \text{Blue})}$

Advanced classification methods

In the current study, we used such advanced classification techniques as a RF, ANN, and XGB for estimation of the vegetation biomass in Mongolia.

The RF classifier is a popular ensemble method that generates multiple decision trees by randomly selecting subsets of both training samples and variables. Each decision tree is structured hierarchically, with nodes divided based on selected features and specific splitting criteria. This method is widely adopted in RS and geospatial fields due to its strong classification performance and reliability. One of its key strengths is its capability to process large and complex datasets effectively, while also being relatively resistant to noise and outliers. In RF, around two-thirds of the data are randomly chosen to train each tree, while the remaining third is used to evaluate the out-of-bag error rate and to determine the importance of different variables (Amarsaikhan et al., 2024).

The ANN is a machine learning model inspired by the structure and function of the human brain. They comprise interconnected nodes known as neurons, which are organized into layers. Each neuron receives input signals, processes them using an activation function, and generates an output signal that can be relayed to other neurons within the network. A multi-layer perceptron is a specific type of feed-forward ANN that includes multiple layers of neurons. In an multi-layer perceptron, neurons typically utilize nonlinear activation functions, enabling the network to learn complex patterns from the given data. This method has been extensively applied in RS digital image processing, and its architectural flexibility, along with its capacity to approximate any function under specific conditions, establishes it as an essential component in the field of neural network research (Chan et al., 2023).

The XGB is a supervised ensemble learning algorithm that supports both classification and regression tasks, and is based on the framework of gradient-boosted decision trees. It offers an effective alternative for modeling a response variable using a set of explanatory variables. As a representative of modern ensemble methods, XGB improves predictive performance while mitigating overfitting during model training

(Ahmetoglu and Das, 2022). The algorithm constructs regression trees sequentially, with each new tree trained to minimize the residual errors from the previous iteration. This iterative process allows the model to incrementally enhance prediction accuracy by integrating outputs from earlier trees. To further prevent overfitting, XGB employs a pruning mechanism that removes low-contributing decision nodes, thereby simplifying the model without significantly compromising its predictive capability (Zhang et al., 2023).

The selected machine learning methods have been used to estimate the vegetation biomass in the entire area of Mongolia. In each model, the vegetation indices along with 3 spectral bands were considered to be independent variables and field measured biomass values were the dependent variable. For implementation of the models, Python algorithm was used.

RESULTS

Mongolia is characterized by a wide range of natural and geographical zones, each supporting unique vegetation communities and associated biomass levels. This natural diversity is reflected in the field-measured AGB data collected across the country. Therefore, the biomass values measured at 275 field measurement points ranged from 10 kg/ha to 1530 kg/ha. The average biomass across all sites was calculated to be 322.95 kg/ha, indicating considerable variation in vegetation density and productivity. The standard deviation of 320 kg/ha further underscores the heterogeneity of biomass distribution across different landscapes. In addition to biomass, the study involved 12 other variables. To ensure comparability and minimize the effects of scale differences among variables, these 12 variables were normalized using feature scaling techniques, resulting in standardized values ranging between -1 and +1. A summary of the descriptive statistics for both the biomass and these normalized variables is presented in Table 2.

As part of the analytical approach, the study aimed to identify which spectral features derived from RS data were most strongly associated with ground-measured biomass. The list of spectral bands and vegetation spectral indices used is provided in Table 1. To quantify the relationships between biomass and these spectral indicators,

Pearson correlation coefficients were computed and summarized in the form of a correlation matrix, shown in Fig. 3. The result revealed that several vegetation indices had relatively strong linear relationships with biomass. Notably, four indices such as the Enhanced Vegetation Index (EVI), Adjusted Transformed Soil-Adjusted Vegetation Index (ATSAVI), Blue-Wide Dynamic Range Vegetation Index (BWDRVI), and Optimized Soil-Adjusted Vegetation Index (OSAVI) demonstrated correlation coefficients (r) equal to or greater than 0.56, suggesting a moderate to strong positive correlation with biomass. In addition, another group of indices, namely, the Normalized Difference Vegetation Index (NDVI), Kernel Normalized Difference Vegetation Index (kNDVI), Triangular Vegetation Index (TVI), and Normalized Near-Infrared Index (NNIR), showed even stronger associations, with correlation values reaching 0.60. These indices appear to be the most effective in capturing variations in AGB among the RS variables evaluated.

On the other hand, some variables exhibited weaker or negative correlations with biomass. The Difference Vegetation Index (DVI), for instance, yielded relatively lower correlation

value compared to other indices, indicating limited usefulness in biomass estimation in this context. Among the individual spectral bands, the NIR band (Band 2) showed the weakest positive relationship with biomass, with a correlation coefficient of only 0.11. Meanwhile, the Red (Band 1) and Blue (Band 3) bands were negatively correlated with AGB, with coefficients of -0.56 and -0.55, respectively. These negative correlations suggest that higher reflectance in the Red and Blue bands is associated with lower biomass, which aligns with known spectral behavior of healthy vegetation absorbing more in these wavelengths.

Following the establishment of correlation patterns between field-measured biomass and a set of spectral indices, machine learning techniques were employed to construct more complex, non-linear models that could effectively predict biomass. These models aimed to capture the intricate relationships between the dependent variable-actual field-measured biomass-and the independent variables. To ensure optimal model performance, a tuning process was conducted for the number of trees used in each ensemble-based algorithm. This was a critical step, as the number of trees in ensemble models such as RF

Table 2. Statistics for biomass measurements and dependent variables

Variables	Mean (kg/ha)	Min (kg/ha)	Max (kg/ha)	Standard Deviation (kg/ha)	Kurtosis	Skewness	Samples
Biomass	322.95	10.00	1530.00	321.00	3.02	1.78	275
Variables	Mean	Minimum	Maximum	Standard Deviation	Kurtosis	Skewness	Samples
Band 1 (B1: Red)	0.13	0.05	0.25	0.04	-0.69	0.41	275
Band 2 (B2: NIR)	0.27	0.16	0.38	0.04	0.44	-0.18	275
Band 3 (B3: Blue)	0.07	0.02	0.13	0.02	-0.43	0.57	275
NDVI	0.37	0.08	0.74	0.16	-0.99	0.09	275
kNDVI	0.40	0.08	0.97	0.19	-0.54	0.38	275
EVI	0.23	0.05	0.49	0.10	-0.75	0.18	275
DVI	0.14	0.03	0.28	0.06	-0.66	0.03	275
TVI	0.93	0.76	1.11	0.09	-1.05	-0.04	275
ATSAVI	0.17	-0.09	0.47	0.13	-0.94	0.02	275
BWDRVI	-0.38	-0.68	0.12	0.16	-0.41	0.34	275
NNIR	0.59	0.42	0.81	0.09	-0.89	0.12	275
OSAVI	0.56	0.12	1.09	0.23	-0.94	0.06	275

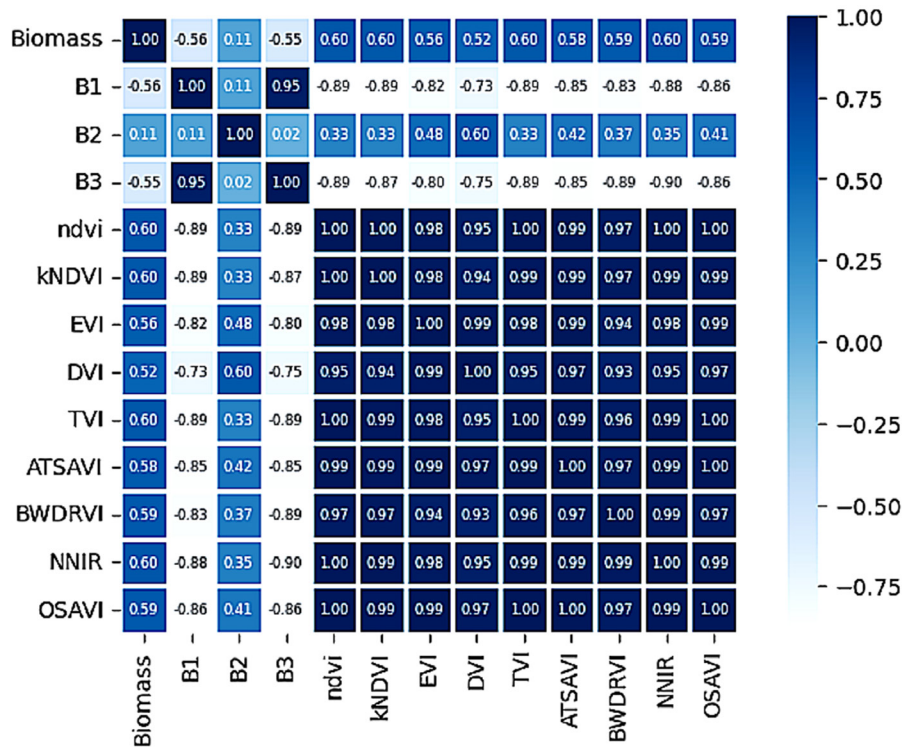


Fig. 3. Correlation coefficients matrix for the selected variables and biomass

or XGB significantly affects model accuracy and computational efficiency.

The biomass estimation was carried out using a 10-fold cross-validation procedure to evaluate model robustness and reduce overfitting. This method involves partitioning the dataset into ten equal subsets. The modeling process is repeated over ten iterations, and in each iteration, nine subsets are used for training the model while the remaining one is used for validation. The process rotates such that each subset is used once as a validation set. Performance metrics are calculated at each iteration and then averaged to provide a more reliable estimate of model performance.

To further enhance the predictive capability of the selected machine learning algorithms, a hyperparameter tuning process was performed. This involved using a grid search strategy, which systematically tests multiple combinations of hyperparameter values to determine the configuration that yields the best performance. The optimized hyperparameters for each model are presented in Table 3. For parameters not explicitly tuned during this process, the default settings provided by the respective algorithms were retained. All modeling and machine learning tasks were executed within a Python environment using Jupyter Notebook, which provided an

interactive and flexible platform for data analysis and visualization. Prior to modeling, data preprocessing and spatial data preparation tasks were conducted using ArcGIS 10.8.

Table 3 outlines the key hyperparameters used in tuning the machine learning models applied in this study. Among them, the `n_estimators` parameter (default = 100) specifies the number of individual trees to be constructed in an ensemble-based learning method. In boosting algorithms, this also corresponds to the number of boosting iterations or rounds. The value assigned to this parameter must be a positive integer (i.e., greater than zero). While the default is 100, adjusting this value can influence both the performance and the generalization ability of the model. The `max_features` hyperparameter defines how many features the model considers when determining the optimal split at each node of a decision tree. This setting affects the randomness and diversity of the trees in the ensemble and can play a significant role in reducing overfitting and improving model accuracy.

Another critical hyperparameter is `eta`, also known as the learning rate (default = 0.3). This parameter controls the magnitude of updates during each boosting round. A smaller `eta` value results in more conservative updates, often requiring more

Table 3. Ranges of tuned hyperparameters

Algorithm	Library	Tuned hyperparameters	Hyperparameter configurations
RF	Scikit-learn	n_estimators max_features	(10, 30, 50, 100, 200) (3, 5, 7, 9, 12)
XGB	xgboost	eta	(0.01, 0.02, 0.5, 0.1)
ANN	Scikit-learn	hidden_layer_sizes	(3), (4), (5), (6), (7), (8)

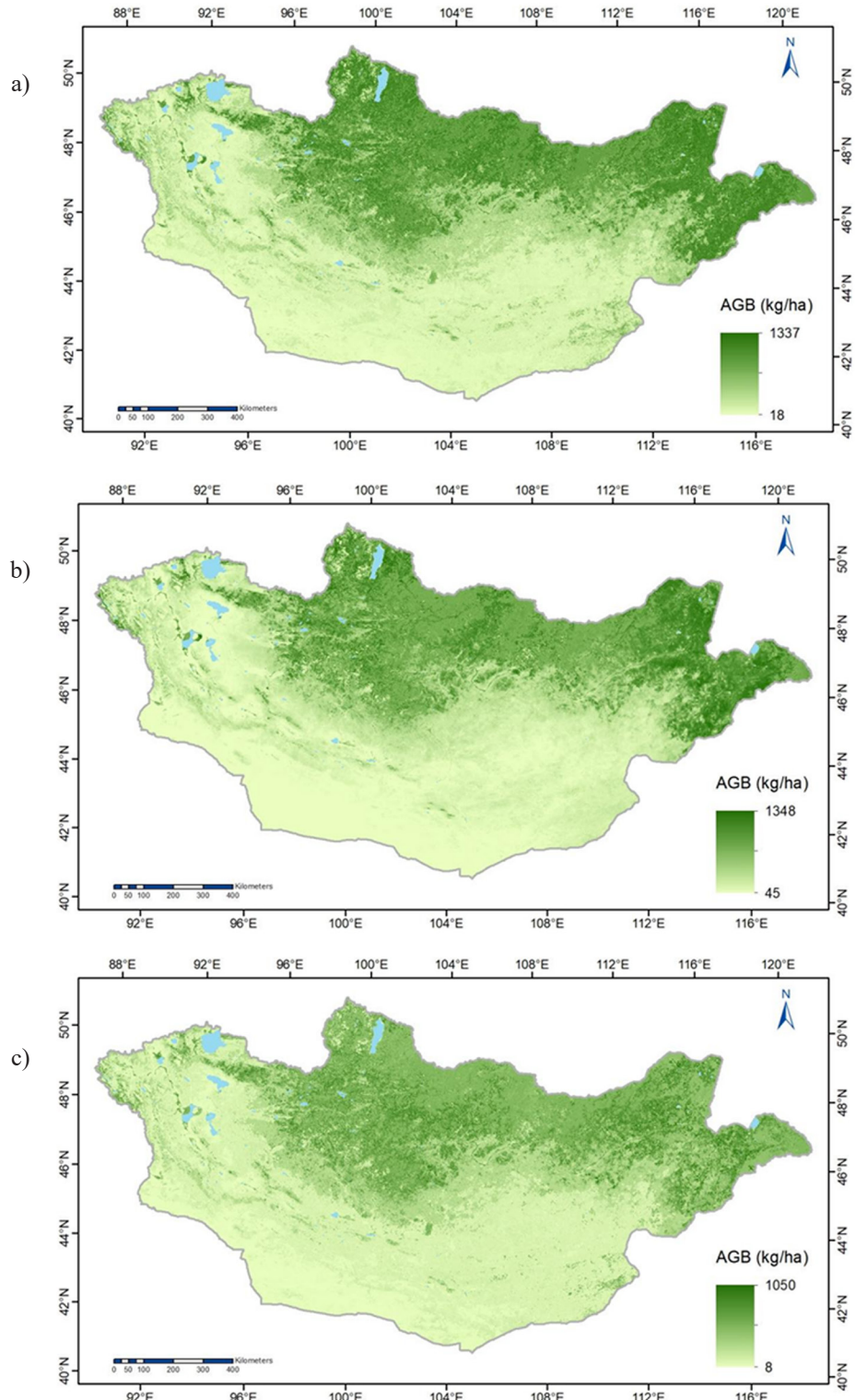


Fig. 4. The results of the classifications (biomass maps) using: (a) RF, (b) ANN, (c) XGB model

boosting rounds to reach a similar level of error reduction as higher learning rates. While a lower learning rate can lead to better generalization, it also increases training time. For neural network-based models, the `hidden_layer_sizes` parameter was used to define the architecture of the network. This parameter indicates the number of neurons in each hidden layer. Tuning this parameter is essential, as it directly impacts the model's capacity to capture complex patterns in the data. In the present study, we used different search values for the hyperparameter configurations.

After conducting a comprehensive statistical analysis of the relationships between the field-observed AGB measurements and RS-based variables across 275 sampling points distributed throughout Mongolia, we proceeded to apply three machine learning algorithms to model and predict biomass values. Each of these models was trained using a 10-fold cross-validation (CV) procedure to ensure reliable model evaluation and generalizability. The input features used for model training included three selected spectral bands and nine spectral indices that had been identified through earlier correlation analysis as being significantly associated with AGB.

After the models were trained and validated, their outputs were used to generate biomass distribution maps across the study region (Fig. 4). The visual representation of these predictions enables more intuitive interpretation of model outputs and supports spatial analysis of ecosystem conditions. To quantitatively evaluate and compare the predictive performance of the RF, ANN, and XGB models, we employed three widely recognized statistical indicators: the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) (Amarsaikhan et al., 2024).

R^2 is commonly interpreted as a measure of how well the model aligns with the observed data, with a higher coefficient indicating a better fit. MAE

represents the average of the absolute differences between the actual and predicted values. On the other hand, RMSE offers insights into the model's performance by comparing the actual differences between the estimated and observed values, with smaller values indicating better model performance (Nyamjargal et al., 2019).

Table 4 summarizes both the training performance and cross-validation performance of the evaluated machine-learning models. Based on the training dataset, the RF model achieved the highest coefficient of determination (Training $R^2=0.88$), followed by XGB (Training $R^2=0.80$) and ANN (Training $R^2=0.61$). However, the 10-fold cross-validation results showed that the XGB model achieved the highest predictive performance (CV $R^2=0.67$), slightly outperforming RF (CV $R^2=0.62$) and ANN (CV $R^2=0.44$). These results indicate that RF provided the best fit to the training data, whereas XGB exhibited superior predictive performance and generalization ability under cross-validation.

Although RF achieved the highest training accuracy metrics, the XGB model obtained the highest cross-validation score, indicating superior predictive robustness and generalization capability. Therefore, considering both model fitting and validation performance, XGB can be regarded as the most reliable model among the three. Furthermore, linear regression analyses were conducted to evaluate the relationship between observed and predicted biomass values obtained from the fitted models (Fig. 5). The resulting regression coefficients describe the goodness-of-fit of the fitted models rather than their cross-validation performance. As shown in Fig. 5, XGB exhibited the strongest regression relationship ($R^2=0.897$), closely followed by RF ($R^2=0.895$), whereas ANN showed the weakest relationship ($R^2=0.431$), indicating greater dispersion of the predicted values.

Table 4. Summary statistics for the selected biomass prediction methods.

	RF	ANN	XGB
Training R^2	0.88	0.61	0.80
Training MAE (kg/ha)	68.45	158.94	87.47
Training RMSE (kg/ha)	109.93	241.66	143.09
10 fold CV R^2	0.62	0.44	0.67

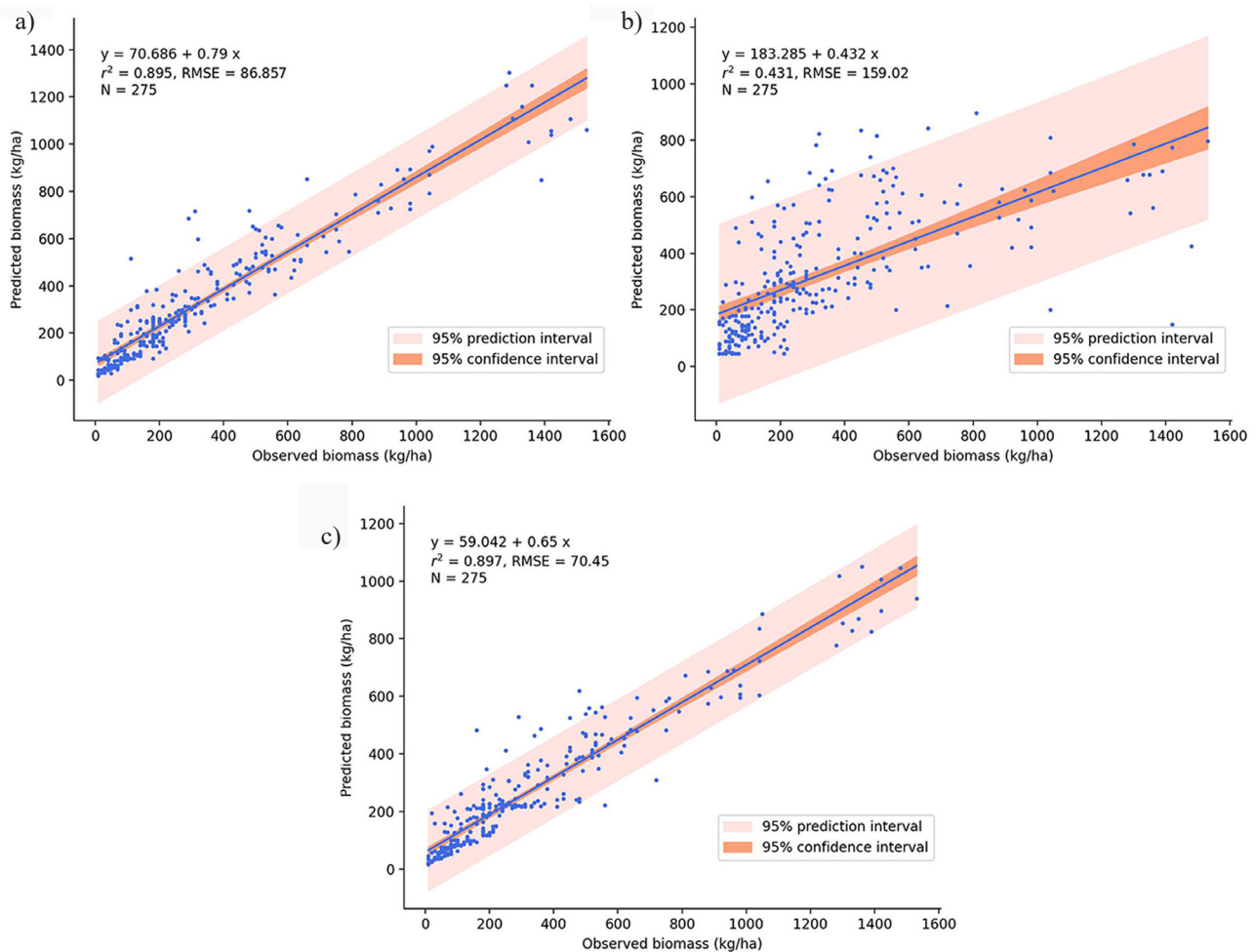


Fig. 5. Regression relationship between observed and predicted biomass values obtained from the fitted models: (a) RF, (b) ANN, and (c) XGB

DISCUSSION

Recent research has delved deeply into various machine learning techniques aimed at accurately estimating pasture biomass utilizing RS data. A study by Amarsaikhan et al. (2023) conducted a comparative analysis of three prominent machine learning algorithms, namely, RF, Support vector machines, and PLSR within the forest-steppe ecosystem of Mongolia. Their results revealed that the PLSR approach consistently outperformed the other models, achieving an impressive predictive accuracy with an R^2 value of 0.899 and a RMSE of 10.560 g/m². In comparison, the Support vector machines model achieved a solid, yet slightly lower performance with an R^2 of 0.837 and an RMSE of 12.881 g/m², while the RF approach yielded the least favorable results, exhibiting a predictive accuracy of $R^2=0.823$ and an RMSE of 13.430 g/m².

In a related investigation, Otgonbayar et al. (2018) explored biomass estimation across Mongolia's vast grassland expanse by applying

both PLSR and RF models and integrating selected spectral indices into their analysis. Their findings indicated that while the PLSR model maintained a satisfactory alignment between predicted and observed biomass values ($R^2=0.750$, RMSE=101.10 kg/ha), the RF model performed marginally better with metrics indicating $R^2 = 0.764$ and an RMSE of 98.00 kg/ha.

However, the variability in the RMSE values reported in this study was significantly greater than those presented by Amarsaikhan et al. (2023), suggesting possible inconsistencies stemming from differing geographic extents, variations in data resolution, or environmental conditions influencing biomass detection. In contrast, our current study observed RF as the most accurate model, yielding an R^2 of 0.88, with an RMSE of 109.93 kg/ha. These discrepancies underscore the sensitivity of model performance to various input variables, including data structure, and radiometric properties of the pixel values.

Vegetation indices have traditionally been

derived from reflectance measurements within the visible and NIR regions of the electromagnetic spectrum, primarily due to the robust correlation between NIR reflectance and vegetation structure or biomass. The growing availability of advanced satellite sensors equipped with SWIR bands has initiated a more comprehensive investigation into the potential of SWIR-based indices.

For instance, Hill et al. (2016) employed linear unmixing of the two-dimensional response envelope of NDVI and ratio (SWIR) 32 indices derived from optical data to retrieve fractional cover of photosynthetic vegetation, non-photosynthetic vegetation, and bare soil for tropical savannah. In a study on accurately modeling the biomass of arid steppe areas in Algeria, Benseghir and Bachari (2021) utilized various vegetation indices that combined visible and NIR OLI bands, visible and SWIR OLI bands, as well as NIR and SWIR OLI bands. Despite these innovative advancements, research efforts such as those by Amarsaikhan et al. (2023) have illustrated that SWIR-based indices do not consistently achieve greater accuracy compared to traditional NIR-based indices. Consequently, our current study emphasizes the importance of utilizing visible and NIR spectral bands to ensure reliable and accurate biomass estimation outcomes.

Ecological and conservation implications of green vegetation biomass mapping

Green vegetation biomass is widely recognized as a fundamental ecological variable that integrates plant productivity, resource availability, and overall ecosystem health. In grassland-dominated systems, such as those covering much of Mongolia, aboveground biomass serves as an effective proxy for net primary productivity and forage availability, while also reflecting key ecological processes including soil moisture dynamics, nutrient cycling, and vegetation structure. Consequently, spatial patterns of biomass derived from remote sensing provide critical insights into ecosystem functioning across climatic gradients, enabling the identification of areas experiencing ecological stress, land degradation, or vegetation recovery (Portela et al., 2020).

Mongolia's rangelands support both pastoral

livelihoods and globally significant steppe ecosystems. However, these landscapes are increasingly threatened by climate variability, overgrazing, and land degradation. Spatially explicit biomass maps generated using machine learning and remote sensing offer a powerful tool for identifying vulnerable grazing areas, monitoring forage availability, and supporting adaptive rangeland management strategies. By enabling rapid assessment of biomass conditions over large areas, the approach presented in this study can contribute to early warning systems for pasture degradation and inform sustainable livestock management practices.

Variability in vegetation biomass is closely linked to climate drivers, particularly precipitation and temperature, which strongly regulate plant growth in Mongolia's dryland ecosystems (Jamiyansharav et al., 2018). Spatial biomass estimation using moderate-resolution satellite data provides an effective means to evaluate ecosystem responses to climate anomalies, such as droughts or extreme temperature events. The modeling framework developed in this study can therefore support long-term ecological monitoring by enabling consistent assessment of biomass dynamics and ecosystem resilience under ongoing climate change (He et al., 2020).

From a conservation perspective, vegetation biomass strongly influences habitat quality and species distribution across Mongolia's diverse ecosystems. Areas characterized by persistently low or declining biomass may indicate habitat degradation, reduced ecosystem services, or increased desertification risk, particularly in environmentally sensitive dryland regions. Conversely, regions exhibiting stable or increasing biomass may represent resilient ecosystems or successful land-management outcomes, and integrating biomass maps with ecological and biodiversity data can therefore enhance conservation planning and support evidence-based decision-making at regional and national scales (Harris et al., 2023).

The strong performance of RF and XGB models observed in this study highlights the value of machine learning approaches for ecological RS applications. Their ability to capture nonlinear relationships between spectral indices and biomass enables more accurate representation

of ecological processes than traditional linear models. Importantly, the demonstrated robustness of these models under small-scale mapping conditions suggests their suitability for operational ecological monitoring, where timely and repeatable biomass estimation is essential for conservation and land-management purposes.

It should be noted that the cross-validation R^2 values (0.62-0.67) were lower than the corresponding training R^2 values (0.80-0.88). This difference suggests a degree of model overfitting and reflects the more conservative nature of cross-validation, which evaluates model performance using independent validation subsets rather than the data used for model fitting. The observed performance gap may also be attributed to the inherent variability of biomass measurements and the complex nonlinear relationships between biomass and spectral predictors. Nevertheless, the 10-fold cross-validation procedure provided a robust assessment of model generalizability, and the XGB model maintained the highest validation performance among the evaluated approaches.

In contrast, the ANN model demonstrated substantially lower predictive performance than RF and XGB. This may be partly attributed to the relatively limited sample size ($n=275$), as ANN generally require larger training datasets to effectively learn complex nonlinear relationships. Additionally, the predictor variables used in this study represent structured environmental and spectral data, for which tree-based algorithms such as RF and XGB often exhibit greater robustness and predictive accuracy. Although ANN hyperparameters were optimized during model development, the selected network architecture may not have fully captured the spatial and ecological variability of biomass conditions across Mongolia, which likely contributed to its lower predictive performance.

Another limitation of this study is its temporal scope. The model development and evaluation were based on field observations and MODIS imagery acquired during August 2021 only. Because vegetation biomass in Mongolia is strongly influenced by interannual climatic variability, the predictive performance reported in this study may not fully represent biomass conditions in other years. Therefore, the proposed framework should be regarded as a

promising approach for biomass estimation under the conditions represented by the study period. Future studies incorporating multi-year observations and a wider range of climatic conditions would be valuable for evaluating the temporal transferability and long-term robustness of the models.

CONCLUSIONS

The study aimed to develop a suitable model for predicting vegetation biomass in Mongolia by comparing the RF, ANN, and XGB techniques. 3 spectral bands and 9 vegetation indices derived from MOD13A1 were used to identify the optimal spectral predictor variables. The indices were calculated from MODIS data of August 2021, utilizing the blue, red, and NIR bands. The training phase employed reference biomass datasets collected from 275 sample plots of a field survey. Hyperparameters for each selected algorithm were tuned to enhance model performance, using a grid search method to identify the optimal hyperparameters. The comparative analysis demonstrated that the RF model achieved the highest predictive accuracy, with an R^2 of 0.88, MAE of 68.45 kg ha⁻¹, and RMSE of 109.93 kg ha⁻¹. In contrast, the ANN model exhibited substantially lower performance ($R^2=0.44$, MAE=158.94 kg ha⁻¹, RMSE=241.66 kg ha⁻¹). The XGB model ranked second, producing reliable estimates with an R^2 of 0.80, MAE of 87.47 kg ha⁻¹, and RMSE of 143.09 kg ha⁻¹. Linear regression analysis comparing observed and predicted biomass values further revealed that XGB yielded the strongest relationship ($R^2=0.897$), closely followed by RF ($R^2=0.895$), while ANN again showed the weakest agreement ($R^2=0.431$). Overall, the results indicate that RF is the most suitable method for vegetation biomass estimation in Mongolia for the analyzed period, while XGB also represents a viable and robust alternative. Given that R^2 values exceeding 0.8 are generally considered satisfactory for ecological modeling, both RF and XGB approaches demonstrate strong potential for operational biomass mapping. However, because the analysis was based on data acquired during August 2021, further validation using multi-year datasets is recommended before extending these findings to long-term national-

scale biomass monitoring applications. Beyond methodological performance, the generated biomass maps provide ecologically meaningful spatial information that can support pastureland management, biodiversity conservation, and climate-change impact assessments across Mongolia's dryland ecosystems.

ACKNOWLEDGEMENTS

This research was conducted within the framework of the project entitled "Assessment of Heatwave Risk in Agricultural Test Fields of Mongolia using Remote Sensing and Machine Learning Methods".

Authorship

Damdinsuren Amarsaikhan formulated the research objectives, designed the research methodology and experimental approach, supervised the project, and wrote the manuscript. Dalantai Sainbayar contributed to the preparation of the original draft of the manuscript and developed the computer programs and algorithms used in the study. E. Jargaldalai contributed to the review and editing of the manuscript and validated the methodology by verifying the accuracy and reliability of the methods and results. Enkhmanlai Amarsaikhan contributed to the design of the software workflow, the development of the research methodology, and the implementation of the experiments. Odontuya Gendaram contributed to data collection and assisted in reviewing and improving the manuscript. Egshiglen Enkhjargal was responsible for data curation, including organizing, cleaning, and maintaining the research data, and assisted in conducting the experiments. Damdin Enkhjargal assisted with the experimental work, field investigations, and survey activities. Tsedev Bat-Erdene assisted with the experimental work and the preparation of figures and illustrations.

Conflict-of-interest

The authors have no potential conflict of interest or no competing interests to declare.

REFERENCES

Adyasuren, Ts. 1989. Remote sensing technology for evaluation of biomass in Mongolia. Doctoral thesis, Ulaanbaatar, Mongolia.

Ahmetoglu, H., Das, R. 2022. A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. *Internet of Things*, vol. 20, 100615. <https://doi.org/10.1016/j.iot.2022.100615>

Altanchimeg, T., Damdinsuren, A., Tseveen, B., Batdorj, B. 2022. Analysis of Relations Between Aboveground Biomass and Vegetation Indices Derived from Sentinel-2 Satellite Data. *Journal of Institute of Mathematics and Digital Technology*, vol. 4(1), p. 94-100 (in Mongolian with English abstract) <https://doi.org/10.5564/jimdt.v4i1.2666>

Amarsaikhan, E., Enkhjargal, D., Jargaldalai, E., Amarsaikhan, D. 2024. Comparison of machine learning and parametric methods for the discrimination of urban land cover types. *Geocarto International*, vol. 39(1), 2380372. <https://doi.org/10.1080/10106049.2024.2380372>

Amarsaikhan, E., Erdenebaatar, N., Amarsaikhan, D., Otgonbayar, M., Bayaraa, B. 2023. Estimation and mapping of pasture biomass in Mongolia using machine learning methods. *Geocarto International*, vol. 38(1), p. 1-17. <https://doi.org/10.1080/10106049.2023.2195824>

Benseghir, L., Bachari, N.E.I. 2021. Shortwave infrared vegetation index-based modelling for aboveground vegetation biomass assessment in the arid steppes of Algeria. *African Journal of Range & Forage Science*, vol. 38(4), p. 281–290. <https://doi.org/10.2989/10220119.2021.1882575>

Chan, K.Y., Salih, B.A., Qaddoura, R., Al-Zoubi, A.M., Palade, V., Pham, D-C., Ser, J.D., Muhammad, K. 2023. Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing*, vol. 545, 126327. <https://doi.org/10.1016/j.neucom.2023.126327>

Chen, Y., Guerschman, J., Shendryk, Y., Henry, D., Harrison, M.T. 2021. Estimating Pasture Biomass Using Sentinel-2 Imagery and Machine Learning. *Remote Sensing*, vol. 13(4), 603. <https://doi.org/10.3390/rs13040603>

Chiarito, E., Cigna, F., Cuzzo, G., Fontanelli, G., Aguilar, A.M., Paloscia, S., Rossi, M., Santi, E., Tapete, D., Notarnicola, G. 2021. Biomass retrieval based on genetic algorithm feature selection and support vector regression in Alpine grassland using ground-based

- hyperspectral and Sentinel-1 SAR data. *European Journal of Remote Sensing*, vol. 54(1), p. 209-225. <https://doi.org/10.1080/22797254.2021.1901063>
- Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E. 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, vol. 9(2), p. 165-193. <https://doi.org/10.1016/j.jag.2006.05.003>
- Enkhtuya J., Altangerel, M., Damdinsuren, A., Oczipka, M., Gurdorj, A., Enkhbayar, B., Tovuusuren B. 2022. Analysis of vegetation indices for pasture biomass evaluation using multi-temporal satellite images and UAV data. 2022. Proceedings of the 43rd Asian Conference on Remote Sensing (ACRS 2022), Ulaanbaatar, Mongolia, ACRS22_38. <https://acrs-aars.org/proceedings2022/>
- Harris, C.D., Lattimore, O., Sanders, A.J.K. 2023. Mongolia. In *Encyclopedia Britannica*. <https://www.britannica.com/place/Mongolia>
- He, D., Huang, X., Tian, Q., Zhang, Z. 2020. Changes in vegetation growth dynamics and relations with climate in Inner Mongolia (2000-2018). *Sustainability*, vol. 12(6), 2534. <https://doi.org/10.3390/su12062534>
- Hill, M.J., Zhou, Q., Sun, Q., Schaaf, C.B., Southworth, J., Mishra, N.B., Gibbes, C., Bunting, E., Christiansen, T.B., Crews, K.A. 2016. Dynamics of the relationship between NDVI and SWIR32 vegetation indices in southern Africa: implications for retrieval of fractional cover from MODIS data. *International Journal of Remote Sensing*, vol. 37(6), p. 1476-1503. <https://doi.org/10.1080/01431161.2016.1154225>
- Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, vol. 83(1-2), p. 195-213. [https://doi.org/10.1016/S0034-4257\(02\)00096-2](https://doi.org/10.1016/S0034-4257(02)00096-2)
- Jamiyansharav, K., Fernández-Giménez, M.E., Angerer, J.P., Yadamsuren, B., Dash, Z. 2018. Plant community change in three Mongolian steppe ecosystems 1994-2013: applications to state-and-transition models. *Ecosphere*, vol. 9(3), e02145. <https://doi.org/10.1002/ecs2.2145>
- Klinge, M., Dulamsuren, C., Erasmí, S., Karger, D.N., Hauck, M. 2018. Climate effects on vegetation vitality at the treeline of boreal forests of Mongolia. *Biogeosciences*, vol. 15(5), p. 1319-1333. <https://doi.org/10.1016/j.quascirev.2019.02.020>
- Klinge, M., Sauer, D. 2019. Spatial pattern of Late Glacial and Holocene climatic and environmental development in Western Mongolia - A critical review and synthesis. *Quaternary Science Reviews*, vol. 210, p. 26-50. <https://doi.org/10.1016/j.quascirev.2019.02.010>
- Kogan, F., Stark, R., Gitelson, A., Jargalsaikhan, L., Dugrajav, C., Tsooj, S. 2004. Derivation of pasture biomass in Mongolia from AVHRR-based vegetation health indices. *International Journal of Remote Sensing*, vol. 25(14), p. 2889-2896. <https://doi.org/10.1080/01431160410001697619>
- Kumar, L., Sinha, P., Taylor, S., Alqurashi, A.F. 2015. Review of the use of remote sensing for biomass estimation to support renewable energy generation. *Journal of Applied Remote Sensing*, vol. 9(1), 097696. <https://doi.org/10.1117/1.JRS.9.097696>
- Lamchin, M., Lee, J-Y., Lee, W-K., Lee, E.J., Kim, M., Lim, C-H., Choi, H-A., Kim, S-R. 2016. Assessment of land cover change and desertification using remote sensing technology in a local region of Mongolia. *Advances in Space Research*, vol. 57(1), p. 64-77. <https://doi.org/10.1016/j.asr.2015.10.006>
- Lin, S., Hu, Z., Wang, Y., Chen, Y., He, B., Song, Z., Sun, S., Wu, C., Zheng, Y., Xia, X., Liu, L., Tang, J., Sun, Q., Joos, F., Yuan, W. 2023. Underestimated Interannual Variability of Terrestrial Vegetation Production by Terrestrial Ecosystem Models. *Global Biogeochemical Cycles*, vol. 37(4), e2023GB007696. <https://doi.org/10.1029/2023GB007696>
- Lourenco, P. 2021. Biomass Estimation Using Satellite-Based Data. In: *Forest Biomass: From Trees to Energy*. IntechOpen. <https://doi.org/10.5772/intechopen.93603>
- Matiza, C., Mutanga, O., Peerbhay, K., Odindi, J., Lottering, R. 2024. Assessing above-ground biomass in reforested urban landscapes using machine learning and remotely sensed data. *Journal of Spatial Science*, vol. 69(3), p. 1047-

1027. <https://doi.org/10.1080/14498596.2024.2343764>
- Meng, X., Gao, X., Li, S., Lei, J. 2020. Spatial and Temporal Characteristics of Vegetation NDVI Changes and the Driving Forces in Mongolia during 1982-2015. *Remote Sensing*, vol. 12(4), 603. <https://doi.org/10.3390/rs12040603>
- Mills, P. 2011. Efficient statistical classification of satellite measurements. *International Journal of Remote Sensing*, vol. 32(21), p. 6109-6132. <https://doi.org/10.1080/01431161.2010.507795>
- Nyamjargal, E., Amarsaikhan, D., Munkh-Erdene, A., Battsengel, V., Bolorchuluun, Ch., 2019. Object-based classification of mixed forest types in Mongolia. *Geocarto International*, vol. 35(14), p. 1615-1626. <https://doi.org/10.1080/10106049.2019.1583775>
- Nyamsuren, B., Nasahara, K.N., Kubota, T., Masaki, T. 2019. Vegetation mapping using GPM/DPR. *Remote Sensing*, vol. 11(20), 2386. <https://doi.org/10.3390/rs11202386>
- Otgonbayar, M., Atzberger, C., Chambers, J., Damdinsuren, A. 2018. Mapping pasture biomass in Mongolia using Partial Least Squares, Random Forest regression and Landsat 8 imagery. *International Journal of Remote Sensing*, vol. 40(8), p. 1-23. <https://doi.org/10.1080/01431161.2018.1541110>
- Portela, M.G.T., de Espíndola, G.M., Valladares, G.S., Amorim, J.V.A., Frota, J.C.O. 2020. Vegetation biomass and carbon stocks in the Parnaíba River Delta, NE Brazil. *Wetlands Ecology and Management*, vol. 28, p. 607-622. <https://doi.org/10.1007/s11273-020-09735-y>
- SDC (Swiss Agency for Development and Cooperation), GGP (Green Gold Program). 2015. National Report on the Rangeland Health of Mongolia. Ulaanbaatar, Mongolia, 66 p.
- Sternberg, T., Tsolmon, R., Middleton, N., Thomas, D. 2011. Tracking desertification on the Mongolian steppe through NDVI and field-survey data. *International Journal of Digital Earth*, vol. 4(1), p. 50-64. <https://doi.org/10.1080/17538940903506006>
- Wesche, K., Ambarlı, D., Kamp, J., Torok, P., Treiber, J., Dengler, J. 2016. The Palaearctic steppe biome: a new synthesis. *Biodiversity and Conservation*, vol. 25(12), p. 2197-2231. <https://doi.org/10.1007/s10531-016-1214-7>
- Zhang, W., Gu, X., Hong, L., Han, L., Wang, L. 2023. Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges. *Applied Soft Computing*, vol. 136, 110066. <https://doi.org/10.1016/j.asoc.2023.110066>