

Integrating in Situ Measurements with Multi-Sensor SAR–Optical Data for Above-Ground Biomass Estimation in Azerbaijan’s Mountain Pastures

Karim Mirzayev^{1*}, Bulent Bayram², Tolga Bakirman³

¹ Ph.D Researcher, Yildiz Technical University (YTU), Istanbul, Turkey

^{2,3} Faculty of Civil Engineering, Geomatic Engineering, Photogrammetry, YTU, Istanbul, Turkey

*Corresponding Author Email: kmirzeyev30@gmail.com

Abstract

Accurate estimation of above-ground biomass (AGB) in mountainous pastures is important for monitoring of grazing livestock management. From previous studies it can be observed that integration of field data with the satellite images (sentinel-1 and/or sentinel-2) is limited. Hence, in this study our goal is to fuse both optic and SAR satellite images at the feature level as inputs in order to evaluate machine learning models to predict biomass in the mountainous of Azerbaijan. Results will be compared with the field-destructive biomass measurements collected in September 2020. Remotely sensed features were extracted from temporally aligned Sentinel-1 SAR ($\gamma^0_{VV_dB}$, $\gamma^0_{VH_dB}$, $VH-VV_dB$, RVI) and Sentinel-2 optical imagery (NDVI, NDMI, MSAVI), alongside topographic covariates (altitude, slope). Considering number of features used as an input for the ordinary least squares (OLS) regression and Random Forest (RF) models, we run the models with original number of features as well as dimensionality reduction analysis. After the initial analysis, we observed that, vegetation indices, SAR-derived features and grass, legume, forb as separate features are highly correlated with each other, therefore applying principal component analysis (PCA) became imminent. OLS and RF models achieved R^2 values up to 0.63 and 0.30 (cross-validated), respectively, with RMSEs corresponding to approximately 271 - 400 kg/ha. Based on feature importance analysis optical vegetation indices and altitude turns out to be most important predictors, while SAR-derived features add stability to the model rather than being significant predictor. Our results demonstrated the integrating multi sourced satellite data with on field dataset for stable summer pasture biomass estimation.

Keywords

Biomass estimation, Data Fusion, Machine Learning, NDVI, Sentinel-2, Sentinel-1, PCA.

INTRODUCTION

Changes in climate are especially important for the agriculture considering it directly impacts the land which resulting decrease in agriculture production [1]. Summer pastures are crucial part of the agriculture production. Considering location of the summer pastures it is hard to monitor them thoroughly. Having information only part of the summer pastures (where roads are available) makes hard to assess situation of the given pastures (excessive animal grazing and lack of sufficient water supply, etc.). In response to mounting needs for repeatable, objective, and spatially comprehensive monitoring to deal with issues with sustaining management, degradation, and effects from climate change, remote sensing has offered a transformative capability to obtain synoptic and temporally repetitive data coverage over large and remote rangelands [2].

The acquisition date of satellite imagery is particularly important over semi-arid and mountainous regions, where phenological variations as well as land management measures

can lead to quite rapid vegetation cover modification [3]. Imagery was captured during late September, coinciding with just after the termination of the major grazing season and destructive biomass sampling throughout the region. The date is optimum since senescing vegetation and onset of dormancy increase live biomass and bare ground spectral separability, especially across chlorophyll- and moisture-sensitive bands [4]. Greater discrimination across visible and shortwave infrared wavelengths during the season allows better residual pasture productivity and land cover fractions to be delineated with increased precision.

Accurate Aboveground Biomass (AGB) quantification is vital to understand how much the summer pastures changed over time. Estimation of accurate AGB is challenging mostly due to tough terrain and traversing through it for collection of data is time-consuming and energy-intensive [5]. Hence, applying remote sensing tools such as satellite images, airborne images (drones and planes) to estimate the biomass information from every part of the pastures makes monitoring easier and more efficient [6].

Instrumental properties of some satellite sensors are also important to improve the analysis. The Sentinel-2 constellation, for instance, provides imagery with up to 10-meter spatial resolution and captures data across twelve spectral bands, from visible (blue, green, red) and near infrared (NIR) to shortwave infrared (SWIR) bands. In these, however, red, NIR, and particularly, SWIR bands are particularly valuable to vegetation observation, facilitating computation of indices such as NDVI (Normalized Difference Vegetation Index), NDMI (Normalized Difference Moisture Index), and MSAVI (Modified Soil Adjusted Vegetation Index) [7]. The indices gain from disparities in reflectance by virtue of plant pigments, plant moisture, and canopy structure, and permits precise measurements to be made concerning vegetation health, phenology, and cover [8]. According to Ismayilov et al. [8] there are over 160 existing vegetation indexes which are used to monitor and assess the situation for the environmental problems.

While landscape-cover scales of pasture monitoring are optimally matched by higher-resolution sensors like Sentinel-2, other platforms, including MODIS (Moderate Resolution Imaging Spectroradiometer), have been extremely popular within rangeland research. MODIS provides daily global coverage but with a coarse-to-moderate spatial resolution (250–500 m for vegetation products in general) and so is ideally matched to broad regional or continental analysis but less useful to resolving heterogeneity within pastures or to monitoring small experimental areas. The high revisit rate of MODIS provides very important strengths to time-series analysis and to monitoring drought, but a large pixel size serves to obscure patterns occurring at fine scales and can lead to contamination by mixed pixels within fragmented landscapes [9]. In contrast, Sentinel-2's higher spatial granularity makes it easier to integrate with field observation, plot experiments, and management units but more susceptible to cloud contamination and having a lower revisit rate.

Even though there are a lot of sources for accessing optical data, trying to attain cloud free version from said sources are not always available. Furthermore, it is even harder to get them, if the area of interests are mountainous regions. Passive optical sensors are only sensitive to the uppermost canopy layer [10]. As a result, information provided by passive sensors are limited to vegetation's vertical structure, which may affect the biomass estimation [11].

The limitations of optical satellite sensors can be overcome with the emergence of synthetic aperture radar (SAR) for the estimation of biomass [11, 12]. Recent interest research of the application of different SAR sensors in agriculture sector and amount of data accumulated made it possible to expand its potential. From previous studies, it can be observed that, SAR sensors are active sensors, meaning

they can monitor the summer pastures day and night to capture data [11]. These sensors emit pulses of microwave energy [13]. Furthermore, atmospheric conditions have no effect on them and they also have cloud penetration characteristics [14].

From the literature we can see that, there are some cases where authors only used different or multiple SAR sensors at the same time for the estimation of biomass in summer pastures [15, 16, 17, 18].

Data integration across several sources, or so-called “data fusion,” has been becoming more well-known among newly emergent studies of pasture biomass. Besides feature-level fusion, used here by stacking, as predictors, derived indices with values of SAR backscatter, several other approaches to data fusion are feasible [19]. In image-level fusion, optical and radar data are fused, pixel by pixel, to produce synthetic images [20]. In decision-level fusion, products from individual models or classifiers are combined. Each process has certain advantages and compromises: image-level fusion can incorporate better spatial information but may incorporate artifacts or heavy radiometric correction; decision-level data fusion offers flexibility but runs a risk of overlooking input feature-ground truth relationships.

Feature-level fusion is a very compelling compromise within this scenario. It involves assembling a large, multi-dimensional predictive feature set with a mixture of both Sentinel-2 spectral indices and Sentinel-1 SAR backscatter, as well as with topographical covariates, to benefit to the fullest extent from each type of sensor [21]. Optical indices are sensitive to green biomass and canopy structure but are susceptible to saturation with high leaf area and are impacted by clouds. SAR backscatter, however, cuts through clouds and responds to canopy moisture, structure, and surface roughness, providing additional information during seasons or locations where optical data are limited.

Feature-level fusion sidesteps pre-processing challenges at the image level and loss of physically meaningful relationships during abstraction at the decision-level. It, rather, optimizes each input's capability to predict by taking them together within the modeling framework [22]. This approach comes naturally to machine learning applications with tasks to capture as much variance as possible within a target variable, our target being destructively sampled dry biomass. Feature-level fusion accepts input from different sources and try to see pattern through combination of given features, therefore increasing model ability to be able to get better results.

After thoroughly researching the available data sources and understanding their strengths and limitations, in this study our goal is to fuse those data at the feature level by integrating SAR and optical datasets as well as the elevation information of the field data.

METHODOLOGY

Study Areas

The study areas are the summer pastures of Gakh and Yardimli in Azerbaijan (Figure 1).

Gakh district is located in the north side of the country. The passage of the Greater Caucasus Mountain range through the district results in a mildly warm climate in the region itself, and cold in the summer pastures (in the mountainous regions). In Gakh pastures, dominant perennial pasture grass species are *Trifolietia-Alchemiletum-Poaosum*, *Stachysetum-Amoriosum* and *Heracleum trachyloma* which make up a relatively closed and productive sward. According to Yashar, [7] the vegetation of the *Trifolietia-Alchemiletum-Poaosum* formation is widespread in the high-mountainous pastures of the area, accounting for 4.5% of the total area of summer pastures in the district.

Yardimli district is located south-east side of the Azerbaijan. Mountainous parts of this region are belong to Talish highlands. According to Gurbanov, [23] Yardimli highlands have *Festuceta-Poaetum-Thymosum* formation group and two *Festuceta ovina* – *Poaetum meyeri* – *Thymusosum trautvetteru* and *Poaetum meyeri* – *Thymus osum trautvetteri* associations which are belong to wheat-grassy different-grassy stepped subalpine meadow formation class.



Figure 1: Study Areas

Azerbaijan's high-mountain summer pastures, play significant roles regarding forage supply to livestock. In order to improve the biomass and analyze the current pasture conditions across contrasting ecological zones were conducted in September 2020 when 60 destructive biomass samples with 0.25 m x 0.25 m quadrats (cages) were collected. 30 of them were created at Gakh and the others at Yardimli. On-ground estimates of pasture biomass can be acquired by destructive and non-destructive (e.g., visual) means. Destructive methods involve laborious cutting and drying of a large number of samples of a known metric such

as a quadrat, where the dried biomass values are scaled up to estimate a larger area [24]. Each sampling site was precisely localized by GPS to allow spatial agreement that can later be used for remote sensing analysis. Even though Gakh and Yardimli's district altitude is different, selected places for the cages were the same in general, hovering between around 1600 – 2400 m elevation (Figure 2 and Figure 3). Each of these biomass harvesting samples were dried using an oven to approximate dry weight and were then used to authenticate satellite-derived biomass measurements.

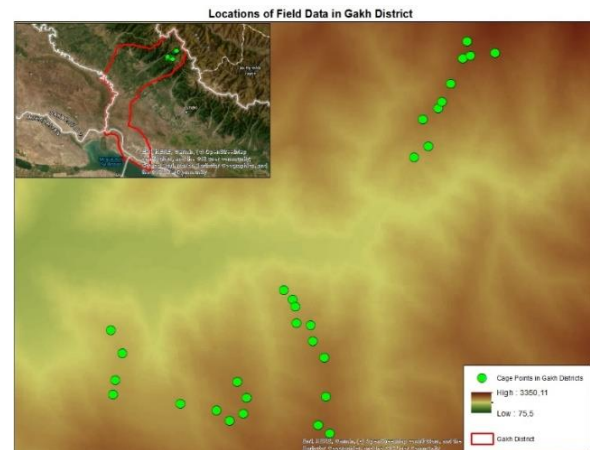


Figure 2: Location of samples in Gakh District

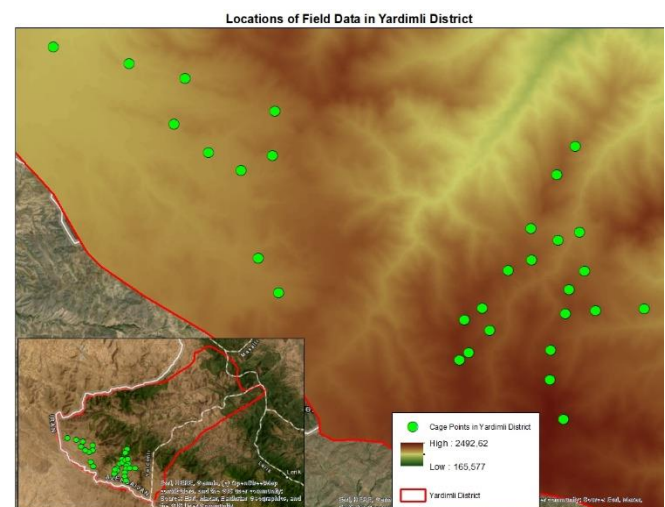


Figure 3: Location of samples in Yardimli District

Preprocessing of the Satellite Data

Due to the extensive preprocessing requirements for Synthetic Aperture Radar (SAR) imagery, Sentinel-1 scenes were manually downloaded from the Copernicus Open Access Hub (<https://scihub.copernicus.eu/>) as Level-1 Ground Range Detected (GRD) products (Figure 4, 5). Manual SAR preprocessing was necessary because Google Earth Engine (GEE) currently does not support advanced Sentinel-1 operations such as radiometric terrain correction and refined speckle filtering, which are essential for quantitative analysis in heterogeneous mountainous terrain.

More information regarding downloaded SAR scenes can be seen from Table 1.

Table 1: Metadata of SAR scenes

SAR SCENE	GAKH	YARDIMLI
File Name	S1A_IW_GRDH_1SDV_20200921T030028_20200921T030053_034451_04020F_D535.SAFE	S1A_IW_GRDH_1SDV_20200928T025253_20200928T025318_034553_04059B_4B34.SAFE
File Size	1634 MB (compressed)	1640 MB (compressed)
Sensing Time	2020-09-21T03:00:28.729000Z	2020-09-28T02:52:53.434000Z
Satellite Platform	Sentinel - 1A	Sentinel - 1A
Sensor	C-band Synthetic Aperture Radar (SAR)	C-band Synthetic Aperture Radar (SAR)
Operational Mode	IW (Interferometric Wide Swath)	IW (Interferometric Wide Swath)
Product Type	IW_GRDH_1S	IW_GRDH_1S
Processing Level	Level - 1	Level - 1
Polarisation Channels	VV & VH	VV & VH
Orbit Direction	Descending	Descending
Relative Orbit Number	79	6
Absolute Orbit Number	34451	34553
Slice Number	4	6
Swath Identifier	IW	IW
Spatial Coverage	40.417912,47.690102; 40.818931,44.716545; 42.317741,45.003090; 41.923538,48.045845	38.121429,49.165157; 38.520275,46.279713; 40.019180,46.609959; 39.621170,49.557648
Acquisition Start	2020-09-21T03:00:28.729000Z	2020-09-28T02:52:53.434000Z
Acquisition End	2020-09-21T03:30:53.727000Z	2020-09-28T03:18:53.434000Z
Publication Date	2020-09-21T06:56:44.220132Z	2020-09-28T07:08:15.149407Z
Origin Date	2020-09-21T06:45:30.239000Z	2020-09-28T07:05:03.295000Z



Figure 4: Footprint of the SAR Scene over Gakh District

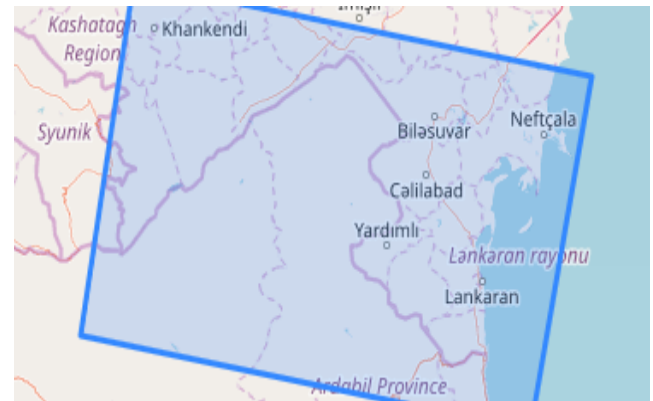


Figure 5: Footprint of the SAR Scene over Yardimli District

Data Conversion and Feature Engineering

Synthetic Aperture Radar (SAR) data from Sentinel-1 were preprocessed using the ESA SNAP toolbox to ensure data quality and comparability. Workflow of the preprocessing of the SAR scenes can be seen from Figure 6. The processing steps includes:

- Precise Orbit Correction
- Thermal and Border Noise Removal
- Radiometric Calibration (β^0)
- Radiometric Terrain Flattening (γ^0)
- Refined Lee Speckle Filtering
- Range-Doppler Terrain Correction

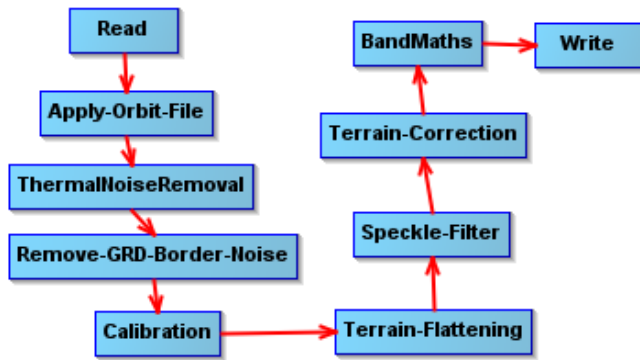


Figure 6: Workflow of the preprocessing of the SAR scenes

After radiometric correction and terrain correction, all the SAR backscattering values were then converted to the decibel (dB) unit in order to obtain a standardized scale for later analysis. It was done through the standard logarithmic transformation,

$$10\log_{10}(\text{linear value}) \quad (1)$$

since the SAR backscattering is given in linear power value to derive vegetation structure information, two of the most important characteristics were computed through the SNAP Band Math tool.

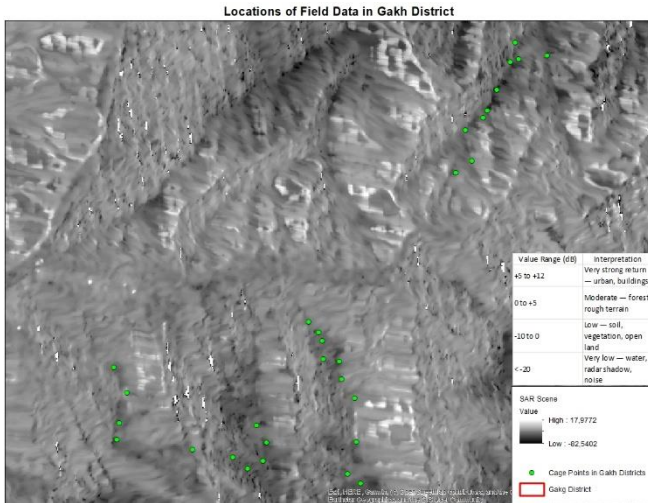


Figure 7: Field data locations in Gakh District overlaid on Sentinel-1 VV backscatter image. Bright areas indicate strong returns (urban/rough terrain), dark areas show low returns (vegetation, water, or shadow).

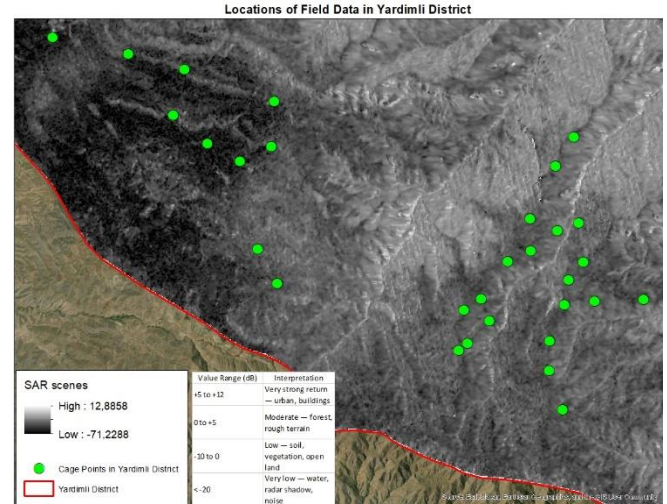


Figure 8: Field data locations in Yardimli District overlaid on Sentinel-1 VV backscatter image. Bright areas indicate strong returns (urban/rough terrain), dark areas show low returns (vegetation, water, or shadow).

The Vertical Horizontal (VH) and Vertical Vertical (VV) polarization difference was also calculated for every pixel. Process, reflects differences in the scattering mechanisms related to vegetation cover and vegetation structure, since VH backscatter is generally most responsive to volume scattering within plant canopies, whereas VV is most responsive to surface scattering (Figure 7,8). Furthermore, Radar Vegetation Index (RVI) was derived as an additional indicator of vegetation conditions, following conversion of dB bands back to linear scale [25]. The RVI is defined as:

$$RVI = \frac{4 \times VH}{VH + VV} \quad (2)$$

Where both VH and VV represent the backscatter intensities in linear units. RVI values help quantify the relative contribution of volume to surface scattering and have demonstrated sensitivity to both biomass and canopy structure [26].

These SAR-derived features were exported as GeoTIFF rasters and then sampled in the locations of field plots, allowing for integration with in situ biomass observations for the purpose of training and validating models.

In contrast, Sentinel-2 optical data and associated vegetation indices such as NDVI, NDMI, MSAVI were accessed and processed directly via Google Earth Engine (GEE) for both of the study areas. Using GEE via JavaScript code made easier to access and download necessary indices automatically rather than trying to download them one by one through Copernicus official site. Automation process made easier for us to filter and attain the optical data similar to SAR scenes and the field data dates.

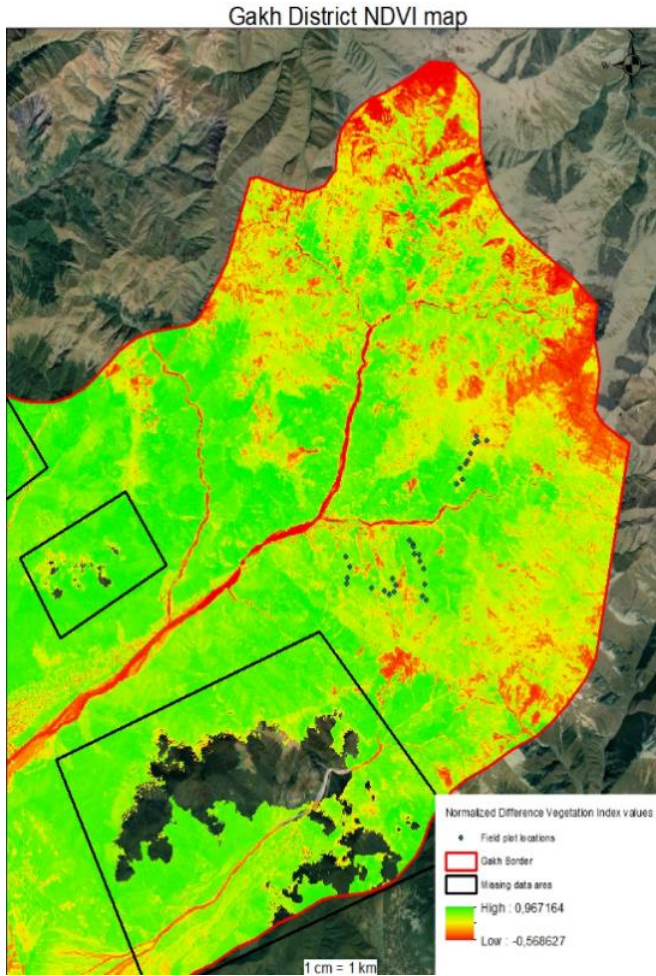


Figure 9: NDVI values in Gakh district

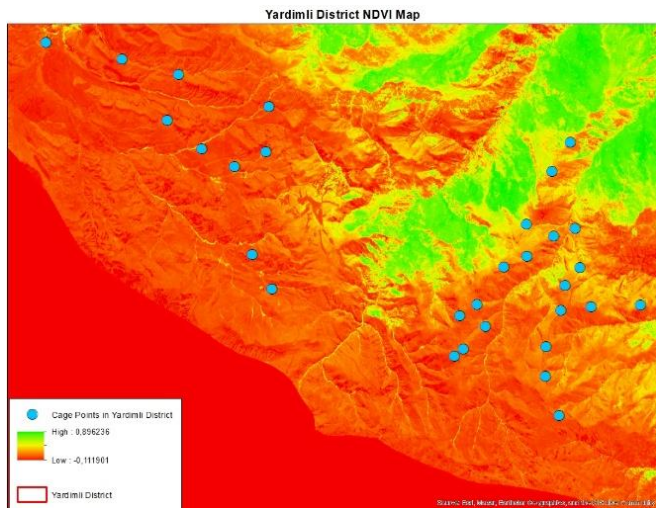


Figure 10: NDVI values in Yardimli district

As mentioned above we downloaded cloud-free Sentinel-2 Level-2A satellite images for each study area using Google Earth Engine. All images were within ± 5 days of the 2020 field sampling. For the filtering processes we used s2cloudless algorithm and mosaicked via median composition.

The following spectral indices were calculated from Sentinel-2 Level-2A surface reflectance values to calculate canopy condition and vegetation moisture status in the study area.

NDVI computed as:

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (3)$$

where NIR and RED represent surface reflectance in the near-infrared (Band 8) and red (Band 4) wavelengths, respectively. NDVI values are always in -1 and 1 interval. If NDVI is in between -1 and 0 it means that, those regions are covered by water, snow or ice covers and if it comes closer to 1 it would mean canopy condition and vegetation moisture have good status.

NDMI calculated as:

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR} \quad (4)$$

with SWIR corresponding to the shortwave-infrared band (Band 11). NDMI is used to determine vegetation water content. It is calculated as a ratio between the NIR and SWIR values in traditional fashion. [27].

Soil Adjusted Vegetation Index (SAVI) is created to minimize soil background influence, specifically areas with weak vegetation cover.

$$SAVI = \frac{(NIR - SWIR)}{(NIR + RED + L)} \times (1 + L) \quad (5)$$

where the soil adjustment factor L was set to 0.5 as recommended for intermediate vegetation densities [8].

The Modified SAVI (MSAVI) was also calculated using the following formula:

$$MSAVI = \frac{2 NIR + 1 - \sqrt{(2 NIR + 1)^2 - 8 \times (NIR - RED)}}{2} \quad (6)$$

Modified Soil Adjusted Vegetation Index (MSAVI) increases the dynamic range of the vegetation signal while minimizing the soil background influences further than the Soil Adjusted Vegetation Index (SAVI). [8].

The Sentinel-2 spectral indices were derived using Google Earth Engine (GEE). While the NDVI and NDMI values appeared as expected, the MSAVI dataset exhibited an unusually low minimum value of -2.29.

After visual inspection of the obtained data, we did not come across any problems regarding the spatial correlation between dataset and the field plots, hence we continued with the analysis. As it can be seen from Fig 9, all three indices (NDVI, NDMI, and MSAVI) had areas where data was not available, however these gaps was not near to any of the field data. Therefore, we determined that, they would not affect the outcome and proceed with the analysis. Also, it should be mention that, gap was only in Gakh district, Yardimli didn't have any problem with obtained data.

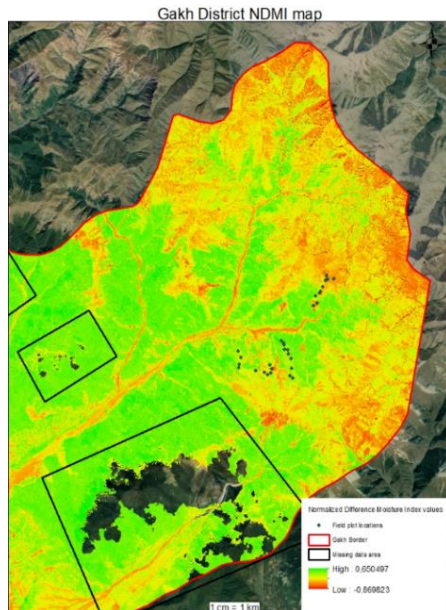


Figure 11: NDMI values in Gakh district

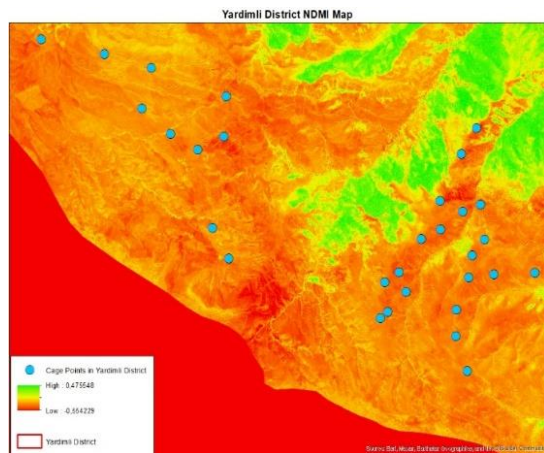


Figure 12: NDMI values in Yardimli district

Optical vegetation index (VI) maps for the study areas reveal obvious differences between Gakh and Yardimli districts. In Gakh, VI maps have more greenness meaning NDVI (Figure 9), NDMI (Figure 11) and MSAVI (Figure 14) values are higher, particularly across on field data plots (quadrat cages). From summer pastures perspective, this results for the Gakh's highlands, indicating healthier and denser vegetation cover. Considering all optical VI maps are identical to each other, for the next steps in our analysis we decided to apply dimensionality reduction to optical VI features and use them as 1 feature instead of treating them 3 different features. Reason for this step is to avoid multicollinearity problem. When we inspect the Yardimli district optical VI maps (Figure 10, 12, 13) we can see different results from the Gakh district. Yardimli exhibits lower and more dispersed optical VI values, particularly across the field plots and from the summer pasture perspective we can conclude that highland is overgrazed. Suggesting tenuous or stressed vegetation.

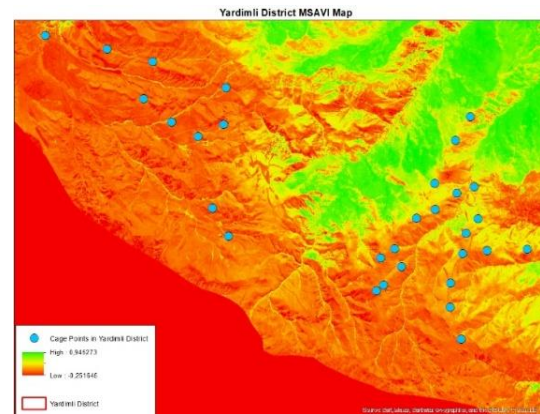


Figure 13: MSAVI values in Yardimli district

For the last features other than optical and SAR images we added altitude and slope information as a feature to the model as well. These features extracted from Digital Elevation Model (DEM) with 30 m resolution. Copernicus official site used to download the raster data. We used ArcGIS Pro to process the DEM and extract both slope and elevation information for each field plot.

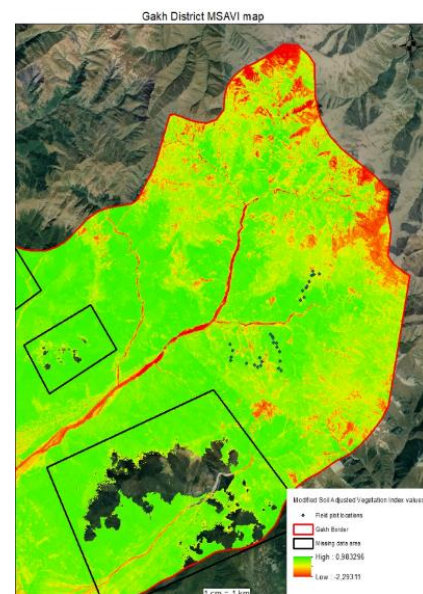


Figure 14: MSAVI values in Gakh district

Model Selection and Evaluation

For each georeferenced cage plot, we created a 7 m buffer to the cage plots using ArcGIS PRO software in order to match field plot resolution with the spatial resolution of the satellite data. In the next step, we extracted median from SAR, Sentinel-2 and DEM values per each buffer. Considering median compositing is not affected by the outliers, we decided to use median for more robust model results.

As mentioned previously in order to calculate the relationship between satellite features and field-measured biomass, we used several statistical and machine learning models.

Ordinary Least Squares (OLS) is simple and widely used regression method [28]. OLS provides a global model for the variable or process that users want to understand or predict, in our case would be the AGB values. By fitting the regression equations to each feature in the dataset model can estimate the direct relationships between all candidate predictor variables and AGB. Even though as previously mentioned, OLS models are widely used, one of the limitations of this method is being sensitive to multicollinearity. Considering extracted features from raster data for our study this limitation creates big obstacle for us to overcome. Hence, we decided to use another machine learning method, namely Random Forest (RF).

Random Forest is a machine learning algorithm that uses many decision trees to make better predictions. Each tree looks at different random parts of the data and their results are put together by averaging the regression. [29]. Key features of RF method is it can be used for different tasks such as classification and regression, for understanding the data better, it tells us which features are most useful and it can handle complex datasets.

After choosing the regression models, next step in our study was to modify parameters of each model. Namely, we used 5-fold cross validation to make sure model does not overfit. In order to keep tabs on the process as a whole we calculated the coefficient of determination (R^2) and root mean square error (RMSE) for each fold.

For this study, we decided to apply Principal Component Analysis (PCA) [30] and Variance Inflation Factor (VIF) analysis to our features before we run the model. By applying PCA and VIF our goal to address the multicollinearity problem considering how close were both our VI features and features that extracted from SAR scenes.

“PCA is a dimensionality reduction method that transforms the data in the input bands from the input multivariate attribute space to a new multivariate attribute space whose axes are rotated with respect to the original space” [31]. Main reason for applying PCA is to compress data by removing redundancy. First principal component will have the largest variance while second principal component will have second most variance which are not described by first. For our case, we applied PCA to different groups of features which first one being optical VI (all optical VI had similar values, hence we applied PCA to remove redundancy while preserve the data itself) and second group was features extracted from SAR scenes namely; 1) γ^0 _VV_dB, 2) γ^0 _VH_dB, 3) VH-VV_dB, 4) RVI.

VIF analysis, on the other hand, is a diagnostic tool for detecting multicollinearity and giving information on which predictors are causing the issue [32]. For our study case, as soon as we look at the VI maps and SAR-derived features we concluded that, these features will have strong correlation

among themselves therefore, dimensionality reduction was necessary. But we decided to apply VIF to see how much are forementioned features are correlated with each other (Table 2).

Table 2: Quantification of inflated features

Features	VIF
Grass	∞ (infinite)
Legume	∞ (infinite)
Forb	∞ (infinite)
Slope	1.619
Altitude	1.796
NDVI	247.3697
NDMI	21.32927
MSAVI	179.5835
gamma_vv_median_20	52,367,440,000,000
gamma_vh_median_20	35,885,260,000,000
vh_vv_db_20	14,251,900,000,000
rvi_20	15

Grass, Legume and Forb are the places where field plots located, we first decided to separate them, however seeing how those values have infinite correlation (Table 1) we applied PCA and used only compressed version of it, namely: productive. Same processing steps goes for both VI and SAR derived features as well. Therefore, we ended up using productive, slope, altitude, optical and radar features as input for our models.

Uncertainty and error in field-derived biomass is mainly related to potential edge effects from pixel values of raster images and sampling error peculiar in destructive harvests. Even though we created 7 m buffer area for each field plots to not to have spatial mismatches, due to terrain (e.g., steep) or from residual preprocessing artifacts it is still possible for issue to happen. Model-related uncertainties/error originates from multicollinearity that can be solved by PCA and feature selection. Lastly like every research field amount of data used for the model also have effect on the how well model can predict.

RESULTS

Integrating both VIF diagnostic tool and dimensionality reduction analysis into the workflow of our study, overall performance of our models namely; linear (OLS) and tree-based (RF) increased. This approach highlights the importance of feature engineering and diagnostic checking in remote sensing-based biomass estimation, especially when working with multi sourced datasets.

Ordinary Least Squares (OLS) regression provided the best apparent fit, with an R^2 of approximately 0.63. However, it should be noted that this estimate is without cross-validation and may therefore be a little optimistic. Therefore, we applied VIF to understand what's going on with the data and turned out just as predicted, multicollinearity was actually off the charts. Hence why, we tried again but this time R^2 dropped to 0.57, but multicollinearity was also dropped significantly (Table 3).

Random Forest (RF) model with all raw features had a moderate fit to the training data with an Out-of-Bag (OOB) R^2 of 0.45. Even so, its cross-validated R^2 fell significantly lower (0.21 ± 0.70), with indications of overfitting and instability. The implication is that, within our dataset, RF model did not generalize as well as did OLS.

In contrast, application of the Random Forest (RF) model using PCA attributes yielded improved OOB and cross-validated performance (OOB $R^2 = 0.52$, CV $R^2 = 0.30 \pm 0.54$). We also achieved a reduction of the Root Mean Squared Error (RMSE). The above results suggest that PCA-facilitated dimensionality reduction benefited RF to generalize by allowing it to focus on principal sources of variation as well as by mitigating redundancy and noise (Table 4).

Table 3: Results of the OLS model

Model	R^2	Adj. R^2
OLS	0.63	0.52
OLS (After applying PCA)	0.57	0.52

Table 4: Results of the RF model

Model	OOB R^2	CV R^2	CV RMSE
RF	0.45	0.21	281 kg/ha
RF (after applying PCA)	0.52	0.30	274 kg/ha

The comparison of variable importance between the two models revealed a few differences. OLS, the strongest association by far belonged to Altitude, with MSAVI and NDVI being the second strongest associations. The rest of Synthetic Aperture Radar (SAR) variables and Radar Vegetation Index (RVI) were significantly less important. For RF with raw data attributes, NDMI (Normalized Difference Moisture Index) performed as the strongest predictor, followed by altitude and NDVI. The SAR attributes contributed little new information to this model.

When applying PCA features to RF model, Spectral_PC1 (the first principal component between NDVI, NDMI, and MSAVI) was most important, then came altitude. The lack of importance by SAR_PC1 means that most variability that can be linked to biomass within our study region had been taken up by optical indices.

Briefly, irrespective of which particular model had been employed, optical indices (notably NDMI, NDVI, and MSAVI) and altitude invariably were by far the best predictors of biomass (Figure 15). Although SAR attributes,

even as compressed PC form, provided minuscule individual predictive capacity, incorporation of all of them via PCA still improved overall model robustness and performance. The latter advantage probably stemmed from them serving to reduce noise and multicollinearity within the dataset.

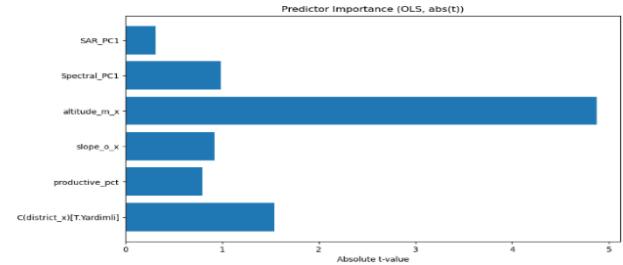


Figure 15: Importance of the Predictors for the OLS Model

Much of this enhanced performance of RF model with PCA features can be attributed to elimination of redundancy and collinearity achieved by compressing optical and SAR features into principal components. This thus enabled RF model to focus on prominent gradients within data. Thus, even though SAR_PC1 wasn't a very significant predictor by individual consideration, a clear improvement was achieved for both OOB and cross-validated statistics. The consequence is that PCA helped to "clean up" this feature set such that the non-linear model fared better even when individual SAR variables weren't very predictive (Figure 16).

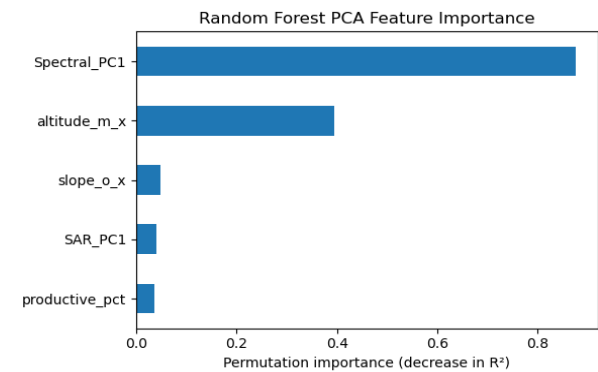


Figure 16: Importance of the Predictors for the RF Model

Small sample size ($n = 59$) very likely contributed to making the OLS model outperform the raw RF model. In limited data and predominantly linear relation scenarios, OLS generally will outperform RF. Random Forest is generally more liable to fit noise within smaller datasets, especially whenever there are substantially collinear predictors.

DISCUSSIONS

Although there are recognized uncertainties, our models yield screening-quality biomass maps appropriate for management prioritization. Following the adjustment for multicollinearity, OLS accounted for a significant portion of plot-level variance ($R^2 \approx 0.57$), whereas all predicted values fall below 1.5 t ha^{-1} , flagging the season/sites as generally

low-biomass and/or the model as conservative at the high end. That makes the map suitable for monitoring potential overgrazing zones but not for separating moderate vs. very high AGB. This pattern also mirrors the scatter's high-end under-prediction; thresholds should be interpreted with field knowledge. If management decisions hinge on the 5-t ha^{-1} line, additional calibration data at high AGB would be valuable (Figure 17).

The cloud lies near the 1:1 line, resulting in a moderate fit. Points diverge as biomass gets larger, and there are larger uncertainties higher in AGB (non-homoscedasticity). The model over-predicts the highest observations and under-predicts the lowest, a typical shrinkage behavior for small samples. The dashed ± 300 kg lines indicate that most plots are contained in a expected error band (Figure 18).

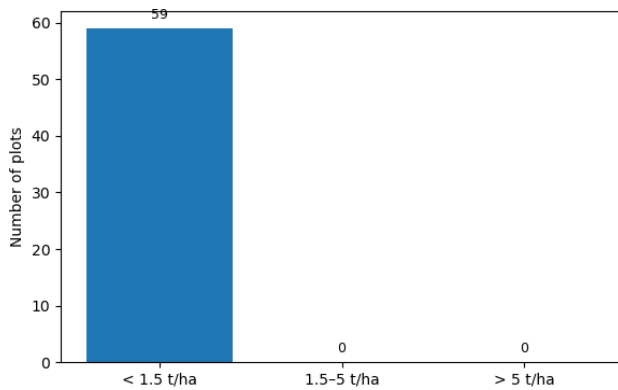


Figure 17: Predicted management classes

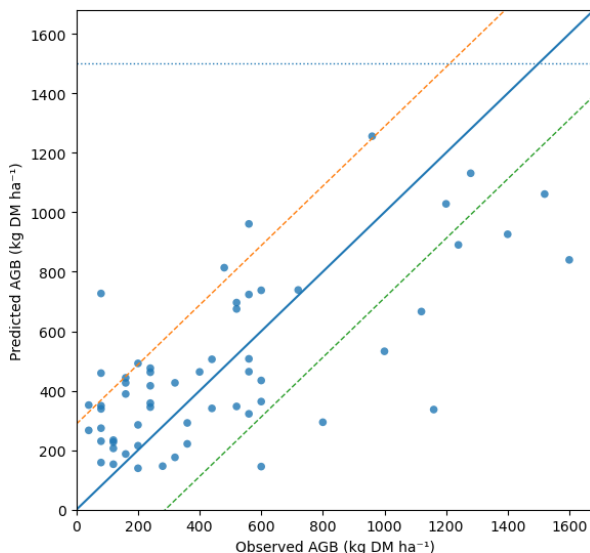


Figure 18: Forecasted versus observed AGB (scatter)

As it can be seen from Figure 16, out of every feature which we used as a predictor for our study case, optical VI exhibited strong predictive capability. Furthermore, altitude also had impact on the predictive capability of the model, while SAR-derived features provided stability rather than being strong predictor in a highland relief and possibility of cloud cover variable.

Figure 18 exhibit that Cross-validated RMSE for biomass prediction ranged from 274 to 400 kg DM ha^{-1} which considering reality of generally accepted threshold for summer pasture monitoring via remote sensing tool.

This curve (Figure 19) gives the proportion of plots whose error is absolute less than a desired tolerance on the x-axis. Roughly 70% are in ± 300 kg ha^{-1} , and $\sim 50\%$ are already in $\sim \pm 220$ kg ha^{-1} . The vertical dashed line corresponds to the model RMSE, so the y-value there corresponds to proportion of plots that do “as well as RMSE.” The constant increase shows errors are not being pulled by a minority of outliers (Figure 19).

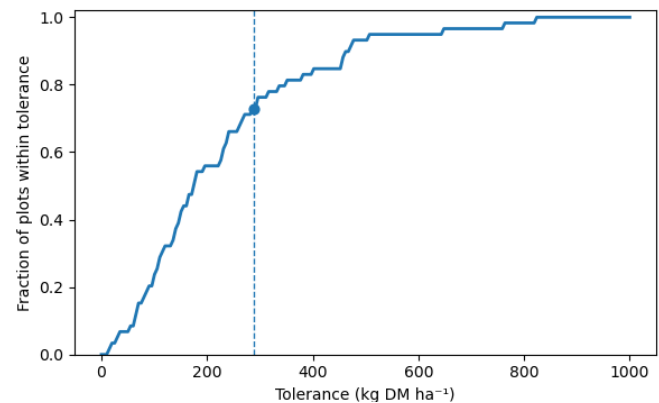


Figure 19: Error-tolerance curve

To mitigate current limitations and boost accuracy, we propose: (i) extending in situ sampling over a range of elevation/aspect layers and across several years to distinguish between site and year effects; (ii) combining phenological trajectories and object-oriented grouping to reduce noise; and (iii) considering higher order SAR descriptors (e.g., multi-date coherence, LIA-informed textures) and regularized non-linear models. In general, a simple OLS remains satisfactory under small-sample, late-season situations, while RF uses PCA to remove redundancy. The workflow offers an efficient approach to conventional 10 m biomass mapping combining field plots, phenology responsive optical indices, and carefully pre-processed SAR in topographically heterogeneous rangelands.

REFERENCES

- [1]. Dong, P., Jing, C., Wang, G., Shao, Y., & Gao, Y. (2024). The Estimation of Grassland Aboveground Biomass and Analysis of Its Response to Climatic Factors Using a Random Forest Algorithm in Xinjiang, China. *Plants*, 13(4), 548. <https://doi.org/10.3390/plants13040548>
- [2]. Guan, K.; Wu, J.; Kimball, J.S.; Anderson, M.C.; Frolking, S.; Li, B.; Hain, C.R.; Lobell, D.B. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sens. Environ.* 2017, 199, 333–349. <https://doi.org/10.1016/j.rse.2017.06.043>
- [3]. Guo, R., Gao, J., Fu, S., Xiu, Y., Zhang, S., Huang, X., ... & Liang, T. (2023). Estimating aboveground biomass of alpine

- grassland during the wilting period using in situ hyperspectral, Sentinel-2, and Sentinel-1 data. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-16. 10.1109/TGRS.2023.3341956.
- [4]. Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International journal of remote sensing*, 25(19), 3999-4014 DOI:10.1080/01431160310001654923.
 - [5]. Allbed, A., & Kumar, L. (2013). Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: a review. *Advances in remote sensing*, 2(4), 373-385. <http://dx.doi.org/10.4236/ars.2013.24040>. <https://doi.org/10.1016/j.ecoinf.2024.102732>
 - [6]. Rex, F. E., Silva, C. A., Dalla Corte, A. P., Klauberg, C., Mohan, M., Cardil, A., ... & Hudak, A. T. (2020). Comparison of statistical modelling approaches for estimating tropical forest aboveground biomass stock and reporting their changes in low-intensity logging areas using multi-temporal LiDAR data. *Remote Sensing*, 12(9), 1498. <https://doi.org/10.3390/rs12091498>.
 - [7]. Yashar A. (2024). Multispektral peyk təsvirləri əsasında otlaq sahələrində biokütlənin (otun) miqdarının hesablanması. Calculation of biomass (grass) quantity in pasture areas based on multispectral satellite images. *Agricultural Economics*, 45(1). <https://agroecconomics.az/az/article/2079/multispektral-peyk-tesvirleri-esasinda-otlaq-sahel/>.
 - [8]. İsmayilov, A. İ., İsmətova, X. R., Abdullayev, X. İ., & Yashar, A. (2017). Multispektral kosmik təsvirlərin emalı əsasında şorlaşmış torpaqların müəyyən edilməsi üçün müxtəlif indekslərin hesablanması [Calculation of various indices for salinized soil identification based on multispectral satellite imagery processing]. *Cöğrafiya və təbii resurslar*, (5). https://gsaz.az/ci/Kitabxana/Jurnallar/Cografiya_ve_tebii_resurslar_2017_N_1.pdf, 118.
 - [9]. 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*, 83(1-2), 195-213. [https://doi.org/10.1016/S0034-4257\(02\)00096-2](https://doi.org/10.1016/S0034-4257(02)00096-2).
 - [10]. F. Zhang and N. Cao, "Application and Research Progress of Geographic Information System (GIS) in Agriculture," 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, 2019, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2019.8820476.
 - [11]. Breunig, F. M., Dalagnol, R., Galvão, L. S., Bispo, P. d. C., Liu, Q., Berra, E. F., Gaida, W., Liesenberg, V., & Sampaio, T. V. M. (2024). Monitoring Cover Crop Biomass in Southern Brazil Using Combined PlanetScope and Sentinel-1 SAR Data. *Remote Sensing*, 16(15), 2686. <https://doi.org/10.3390/rs16152686>
 - [12]. Parag, M., Lottering, R., Peerbhay, K., Agjee, N., & Poona, N. (2024). The use of synthetic aperture radar technology for crop biomass monitoring: A systematic review. *Remote Sensing Applications: Society and Environment*, 33, 101107. <https://doi.org/10.1016/j.rsase.2023.101107>.
 - [13]. Yun, S. H. A. O., & Hasi, T. (2019). Research advances of SAR remote sensing for agriculture applications: A review. *Journal of integrative agriculture*, 18(3), 506-525. [https://doi.org/10.1016/S2095-3119\(18\)62016-7](https://doi.org/10.1016/S2095-3119(18)62016-7).
 - [14]. Baghdadi, N., Boyer, N., Todoroff, P., El Hajj, M., & Bégué, A. (2009). Potential of SAR sensors TerraSAR-X, ASAR/ENVISAT and PALSAR/ALOS for monitoring sugarcane crops on Reunion Island. *Remote sensing of environment*, 113(8), 1724-1738. <https://doi.org/10.1016/j.rse.2009.04.005>.
 - [15]. Moreau, S., & Le Toan, T. (2003). Biomass quantification of Andean wetland forages using ERS satellite SAR data for optimizing livestock management. *Remote Sensing of Environment*, 84(4), 477-492. [https://doi.org/10.1016/S0034-4257\(02\)00111-6](https://doi.org/10.1016/S0034-4257(02)00111-6).
 - [16]. Dhar, T., Menges, C., Douglas, J., Schmidt, M., & Armston, J. (2010, July). Estimation of pasture biomass and soil-moisture using dual-polarimetric X and L band SAR-accuracy assessment with field data. In 2010 IEEE International Geoscience and Remote Sensing Symposium (pp. 1450-1453). IEEE. <https://doi.org/10.1109/IGARSS.2010.5650570>.
 - [17]. McNeill, S. J., Pairman, D., Belliss, S. E., Dalley, D., & Dynes, R. (2010, July). Robust estimation of pasture biomass using dual-polarisation TerrASAR-X imagery. In 2010 IEEE International Geoscience and Remote Sensing Symposium (pp. 3094-3097). IEEE. <https://doi.org/10.1109/IGARSS.2010.5649266>.
 - [18]. Li ZhaoQin, L. Z., & Guo XuLin, G. X. (2017). Can polarimetric Radarsat-2 images provide a solution to quantify non-photosynthetic vegetation biomass in semiarid mixed grassland? <https://doi.org/10.1080/07038992.2017.1394181>
 - [19]. Singh, S., Tiwari, R. K., Sood, V., Gusain, H. S., & Prashar, S. (2021). Image fusion of Ku-band-based SCATSAT-1 and MODIS data for cloud-free change detection over Western Himalayas. *IEEE transactions on geoscience and remote sensing*, 60, 1-14. <https://doi.org/10.1109/TGRS.2021.3123392>.
 - [20]. Shah, E., Jayaprasad, P., & James, M. E. (2019). Image fusion of SAR and optical images for identifying Antarctic ice features. *Journal of the Indian Society of Remote Sensing*, 47(12), 2113-2127. <https://doi.org/10.1007/s12524-019-01040-3>
 - [21]. Cao, X., Chen, J., Matsushita, B., & Imura, H. (2010). Developing a MODIS-based index to discriminate dead fuel from photosynthetic vegetation and soil background in the Asian steppe area. *International Journal of Remote Sensing*, 31(6), 1589-1604. <https://doi.org/10.1080/01431160903475274>
 - [22]. Schmitt, Michael & Zhu, Xiao. (2016). Data Fusion and Remote Sensing – An Ever-Growing Relationship. *IEEE Geoscience and Remote Sensing Magazine*. 4. 6-23. <https://doi.org/10.1109/MGRS.2016.2561021>
 - [23]. Gurbanov E. (2023) YARDIMLI, O. SUBALPINE GRASSLAND VEGETATION. *Bulletin of Science and Practice*. <https://doi.org/10.33619/2414-2948/92/08>
 - [24]. Woodland, O. S. (2019). Integration of Optical and X-Band Radar Data for. *Remote Sensing of Above Ground Biomass*, 41. <https://doi.org/10.3390/rs8120989>
 - [25]. Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J. M., ... & Rao, Y. S. (2020). Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data. *Remote Sensing of Environment*, 247, 111954. <https://doi.org/10.1016/j.rse.2020.111954>
 - [26]. Nasirzadehdizaji, R., Balik Sanli, F., Abdikan, S., Cakir, Z., Sekertekin, A., & Ustuner, M. (2019). Sensitivity analysis of multi-temporal Sentinel-1 SAR parameters to crop height and canopy coverage. *Applied Sciences*, 9(4), 655. <https://doi.org/10.3390/app9040655>.
 - [27]. Lykhovyyd, P. V., & Sharii, V. O. (2024). Normalised difference moisture index in water stress assessment of maize crops. *Agrology*, 7(1), 21-26. <https://doi.org/10.32819/202403>
 - [28]. SuperMap. (2025). OLS In iDesktopX Help. <https://help.supermap.com/iDesktopX/en/SpecialFeatures/Analyst/SpatialStatisticalAnalysis/OLS.html>.

-
- [29]. GeeksforGeeks. (n.d.). *Random forest algorithm in machine learning*. <https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning>.
- [30]. Roshmita Dey, (2023). Understanding Principal Component Analysis (PCA) <https://medium.com/@roshmitadey/understanding-principal-component-analysis-pca-d4bb40e12d33>.
- [31]. Esri, “How Principal Components Works,” ArcGIS Pro Documentation, 2024. <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-principal-components-works.htm>.
- [32]. DataCamp. (n.d.). *Tutorial: Variance Inflation Factor (VIF)*. <https://www.datacamp.com/tutorial/variance-inflation-factor>.