DEVELOPMENT LAND COVER CLASSIFICATION APPROACH USING LINEAR MIXING MODEL AND RANDOM FOREST IN GOOGLE EARTH ENGINE

Davaasuren Bayarmagnai[1], Tsolmon Rentsen[1], Bayanjargal Darkhijav[2]
[1] Department of Physics, National University of Mongolia
[2] Department of Applied Mathematics, National University of Mongolia
Email: dawka99.magnai@gmail.com

KEY WORDS: Land cover class, Linear mixing model (LMM), random forest, classification accuracy, Google Earth Engine (GEE)

ABSTRACT: The purpose of the research is to develop and compare land cover classification methods using linear mixing model (LMM) and random forest in Google Earth Engine (GEE). The study area is Khangal soum of Bulgan province and the site is situated in a forest-steppe zone with mountains and hills. The area such land cover types as bare land, forest, and grass. Spectral bands 2, 3, 4, and 5 of Landsat 8 data of 2018 and ground truth data have been used in the research. Result of the LMM was compared with the result of the random forest methodology along with ground truth measurements. The overall accuracy of the LMM using ground truth data was 75.2%. Unlike the model, the result of the random forest technique and ground observation data had a good agreement, resulting in overall accuracy of 94.4%.

1. INTRODUCTION

Generally, multispectral satellite remote sensing (RS) data sets with different spectral, spatial and temporal resolutions have been efficiently used for land cover classification and analysis since the operation of the first Landsat satellite launched in 1972. Image classification is one of the extensively used RS data processing methods, and the traditional methods mainly involved supervised and unsupervised methods and hence, a great number of techniques have been developed (Nyamjargal et al. 2020). However, most traditional classification algorithms have not sufficiently managed the difficulties represented in the RS images, if the given areas are complex and diverse in nature, and the illustrated Earth's and artificial features or classes of objects have similar spectral characteristics (Amarsaikhan, 2020). In other words, it is was not easy to separate these spectrally similar classes by the use of common feature combinations or by applying ordinary techniques.

Over the past years, different advanced methods and machine learning techniques such as such as random forest, support vector machine, artificial neural network, cubist, K-nearest neighbour, Bayesian network, LMM, linear or nonlinear regression models have been developed for an efficient land cover classification and other land-related analysis to overcome the traditional difficulties, and solve many of the existing problems. Different authors have used either one or combination of these methods along with some other suitable features (Otgonbayar et al. 2019 Enkhmanlai, 2022). Usually, in order to successfully classify any RS data into a number of class labels it is very important to determine reliable features combinations and design a suitable image processing procedure. The effective use of appropriate features and the selection of a reliable classification technique can be a key significance for the improvement of classification accuracy (Amarsaikhan et al. 2012).

In recent years, LMM and random forest methods along with various feature combinations have been efficiently used for land cover discrimination. The LMM offers an effective framework to analyse mixed pixels, and its underlying assumption is that knowledge of end members with known spectral profiles yields the fractions of the end members within a pixel. The success of the model result relies on the quality of the priori knowledge, and depends on a proper identification of the main components present in the scene. This identification might be difficult if the image has been acquired over very heterogeneous landscapes or when we work with coarse resolution data because in these cases most of the pixels are mixed (Clevers and Zurita-Milla, 2008). Random forest is a supervised machine learning algorithm that builds decision trees on different samples and takes their majority vote for classification. It can be used for regression or classification depending on the type of variable to be estimated. Compared with linear regression techniques, the method has lower bias and avoids overfitting (Tian et al. 2017). The advantage of the method is that it can run effectively on large data sets and it is relatively robust to outliers, a reduction of training data and noise (Hastie, 2009).

The aim of this research is use the LMM and random forest algorithm in the GEE for the classification of land cover types and compare the performances of the selected methods. For this purpose, four multispectral bands of the 2018 Landsat data acquired over the Khangal sum area of Bulgan province, northern part of Mongolia, and ground truth information have been selected. Overall, the study indicated that although both methods can be applied for the classification, the random forest has a superior performance compared to the LMM.

2. STUDY AREA AND USED DATA

The study area is Khangal soum, Bulgan province of Mongolia. Bulgan is one of the northern provinces of Mongolia, located between the latitudes 47°14’ - 50°23’ N and longitudes 101°37’ - 104°45’ E, geographically its territory belongs to the Khangai mountain forest-steppe zone. The soil type is sandy with semi-desert features in the southern part. In the current study, we have selected four different classes of land cover types at the study site such as larch, birch, vegetation and bare land. Figure 1 shows the slected study area illustrated in a satellite imagery (Bayanmunkh et al 2019).

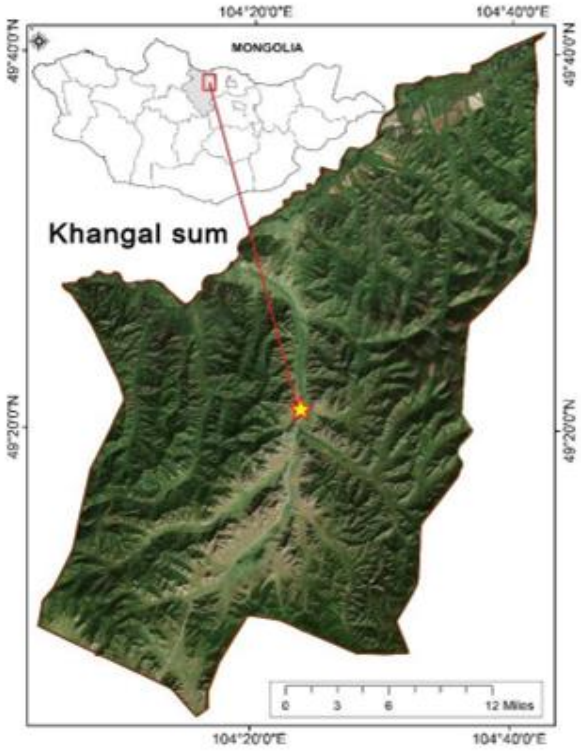

Figure 1. Study area: Khangal soum in Bulgan province.

We have used the following 3 types of data:

- Landsat 8 satellite data of 2018
- Google Earth Engine platform
- Ground observation data of 2018

A field survey for ground observation data collection was carried out in 2018. Landsat 8 satellite data were downloaded from the homepage of the United States Geological Survey (USGS). For the study, four multispectral bands with 30m resolution have been selected.

3. METHODOLOGY

3.1 Satellite data preparation

In the current study, we used the Landsat 8 data. The instruments of Landsat represent an evolutionary advance in technology and has two science instruments—the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). OLI improves on past Landsat sensors using a technical approach demonstrated by a sensor flown on NASA’s experimental EO-1 satellite. It is a push-broom sensor with a four-mirror telescope and 12-bit quantization, and collects data in visible, near infrared, and short wave infrared spectral bands as well as a panchromatic band. TIRS provides coverage of the Earth’s surface features at a spatial resolution of 100 m (Landsat 8, 2013). We have selected 4 bands of OLI with a pixel resolution of 30 m. As one pixel of satellite data is 30 × 30 m by 900 sq.m. Therefore, the selected models can find the percentage of each land cover type in an area of . For example, in a one-pixel image: red is a larch, green is a birch, blue is vegetation, and yellow is bare land, and our model can find the percentage of larch, birch, vegetation, and bare land on one pixel (Figure 2).

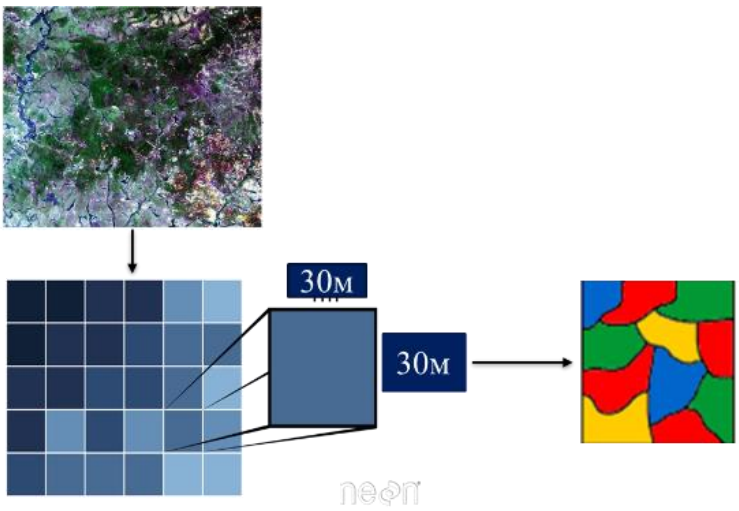

Figure 2. Land cover classes per pixel.

3.2 LMM for land cover classes

We used the LMM to classify the land cover classes. The model is an extension of simple linear models and considers two concepts: fixed effect and random effect. A fixed effect refers to the covariance of dependent variables among the data as a whole, while a random effect refers to clusters within the data. This method is also used to separate one item from a whole or complex item. In this way, the percentage of land cover type per pixel can be calculated using the model.

The linear system of equations of the LMM for the land cover type can be described as follows:

$$\mathrm{R} = A * X + E \tag{1}$$

Where R is a $m \times 1$ matrix whose elements are the reflectance values of spectral bands, A is a $m \times n$ matrix where the elements are ground observation data, X is a $n \times 1$ variable matrix, E is a $m \times 1$ matrix which has error values, n denotes the land cover types (classes) and m is the number of spectral bands of Landsat 8 data.

In our case, the number of spectral bands is 4 and land cover types are birch, larch, vegetation and bare land. Therefore, the detailed form of the linear system can be written as below:

$$\begin{cases} r_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + e_1 \\ r_3 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 + e_2 \\ r_3 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 + e_3 \\ r_4 = a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 + e_4 \end{cases} \tag{2}$$

Where r_i is a reflectance value on Band i ($i = 1,2,3,4$), a_{ij} is a ground observation data (j = Birch, Larch, Vegetation and Bare land) on Band i, x_j is variable for the land cover types, and e_i is error value of Band i.

In order to find the percentage of land cover classes, we consider the following optimization problem with additional constraints where the sum of the squares of the errors should be minimum as:

$$f(x) = \sum_{i=1}^{4} e_i^2 = \sum_{i=1}^{4} (r_i - \sum_{j=1}^{4} a_{ij}x_j)^2 \rightarrow \min \tag{3}$$

$$x_1 + x_2 + x_3 + x_4 = 1$$

$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, \tag{4}$$

The matrix form of the non-linear optimization problem is:

$$\begin{cases} f(x) = E^2 \rightarrow min \\ \sum X = 1 \\ X \geq 0 \end{cases} \tag{5}$$

Where $E = R - AX$.

3.3 Random forest algorithm

The random forest is a family of tree-based models; in the first one, data are stratified into homogeneous subsets by decreasing the within-class entropy, whereas in the second one, a large number of regression trees are constructed by selecting random bootstrap samples from the discrete or continuous datasets (Pflugmacher et al. 2014). The advantage of technique is that it can run effectively on large data sets, relatively robust to outliers, and efficient to complex training data (Hastie et al. 2009; Rodriguez-Galiano et al. 2012). For each tree, approximately two-thirds of the original data was randomly chosen to build the tree, and the remaining data was used for estimating out-of-bag error and calculating variable importance.

4. ANALYSIS AND RESULTS

The more accurate the ground observation data for each land cover class are, the better the results in the model. MATLAB program was used to solve the non-linear optimization problem (5) for total of 27030 pixels. The results are expressed as a percentage between 0 and100 and for the mapping the ArcGIS program was applied.

The results of the nonlinear optimization problem (5) using Matlab software are shown in Table 1. As from the table 1, birch has 30%, while larch has 8%. Moreover, it is seen that vegetation occupies 49.6%, whereas bare land has 11.4% in the first row of the table.

1	Birch	Larch	Vegetation	Bare land
2	0.300435548	0.089431016	0.495932022	0.114201414
3	0.245158989	0.084344822	0.515126579	0.15536961
4	0.415467932	0.010339484	0.377439425	0.196753159
5	0.214874779	0.113064716	0.50486717	0.167193335
6	0.245177812	0.222314316	0.219980345	0.312527528
7	0.238705202	0.212702664	0.277013624	0.27157851
8	0.245444519	0.203689476	0.308338209	0.242527796
27023	7.33878E-05	0.999812045	2.78208E-05	8.67468E-05
27024	9.43215E-05	0.999747998	3.58443E-05	0.000121836
27025	0.000236747	0.970891143	8.64381E-05	0.028785672
27026	0.000163815	0.947323594	5.98217E-05	0.05245277
27027	0.000182613	0.976570057	6.66997E-05	0.02318063
27028	0.000269099	0.974486851	9.82399E-05	0.02514581
27029	0.000877083	0.891097363	0.000319569	0.107705985
27030	0.013833275	0.742376813	0.004964754	0.238825157

Table 1. Percentage of each land cover class.

After defining the percentages of the land cover classes, mapping was carried out by the use of ArcGIS system (Figure 3). As could be seen from the Figure 3, larch is the dominant class in Khangal sum of Bulgan province. The birch class is sparsely distributed, and vegetation is spread along ravines. Meanwhile, bare ground can be seen on the mountain slopes. The low proportion of some land cover classes can be hardly seen because of the small number of pixels.

The result of the random forest classification using the GEE are shown in Figure 4. The GEE is a cloud computing platform designed to store and process huge data sets (at petabyte-scale) for analysis and ultimate decision-making (Raich and Schlesinger, 1992). The Engine also has a Code Editor which can be applied to run a machine learning algorithm such as the random forest algorithm. The current archive of data includes those from other satellites, as well as geographic information systems (GIS) based vector data sets, social, demographic, weather, digital elevation models, and climate data layers. In order to use GEE, ground observation data of the study area was entered manually and calculated for each pixel using the standard functions of the machine learning method.

In the study, to evaluate the performances of the classifications a confusion matrix was used. We used the ground observation data selecting 40 points for the larch area, 10 points for the birch area, 30 points for the vegetation cover,

and 25 points for the bare land, respectively. After that we created a confusion matrix for the LMM. In the random forest algorithm, 118 points for the vegetation, 132 points for the bare land, 101 points for the larch cover, and 76 points for the birch cover have been selected as ground information, and a confusion matrix was constructed to test the model. The overall accuracy of the LMM was 75.2%, while for the random forest algorithm it was 94.4%. (Table 2).

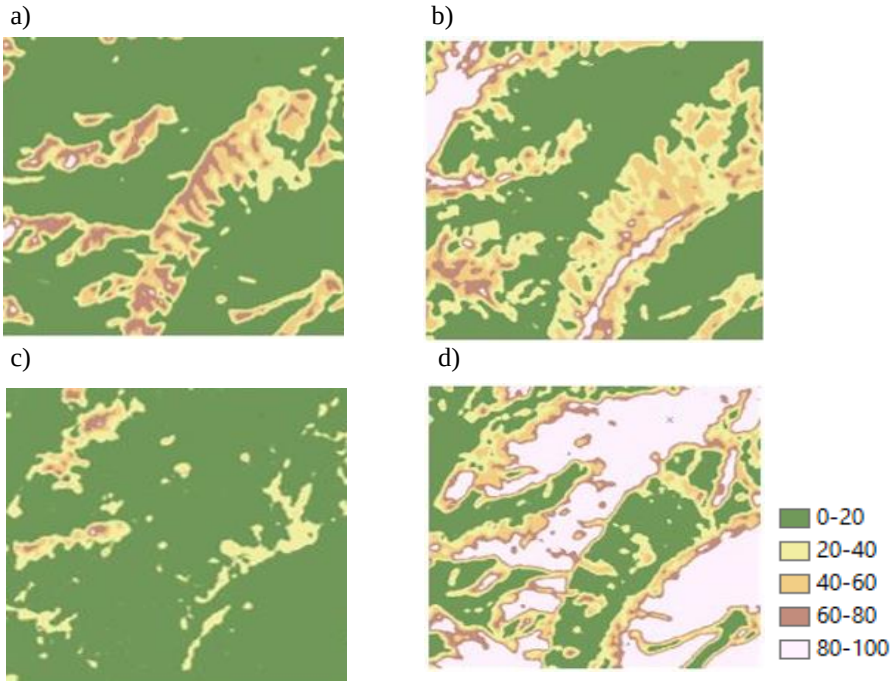

Figure 3. Land cover map: (a) Bare land, (b) Vegetation, (c) Birch, (d) Larch

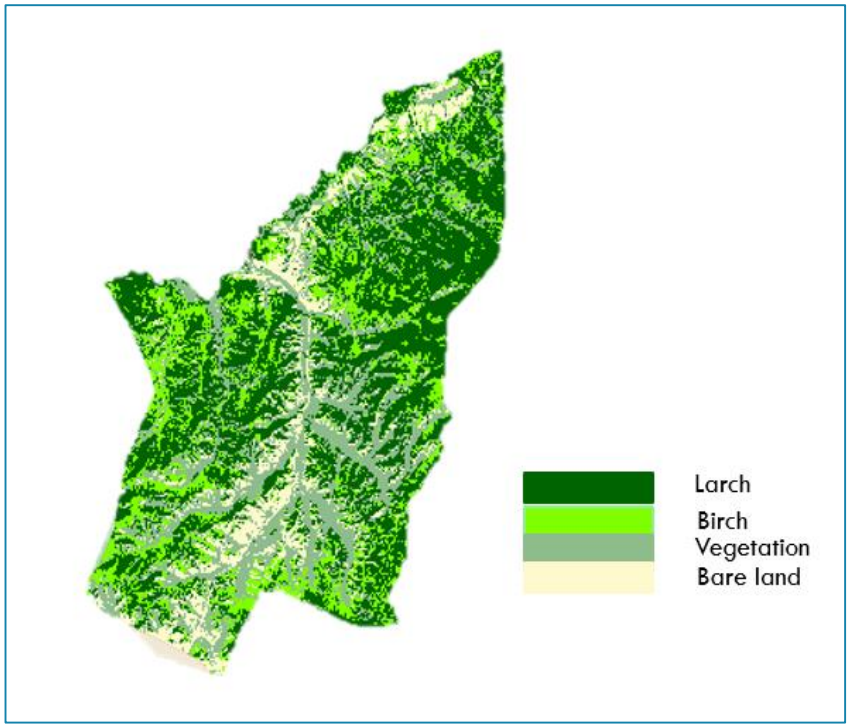

Figure 4. Result of random forest in the GEE.

		Ground observation data			
		Larch	Birch	Vegetation	Bare land
Predicted value	Larch	38	1	0	0
	Birch	1	6	3	0
	Vegetation	1	7	14	0
	Bare land	0	4	9	21

b)

		Ground observation data			
		Larch	Birch	Vegetation	Bare land
Predicted value	Larch	54	5	4	1
	Birch	3	44	1	0
	Vegetation	0	0	85	0
	Bare land	0	1	1	87

a)

Table 2. Confusion matrix for the LMM (a), Confusion matrix for the random forest in GEE (b).

Moreover, in order to check the result of the LMM, it was also compared with natural color image created by the visible bands of Landsat 8 data.

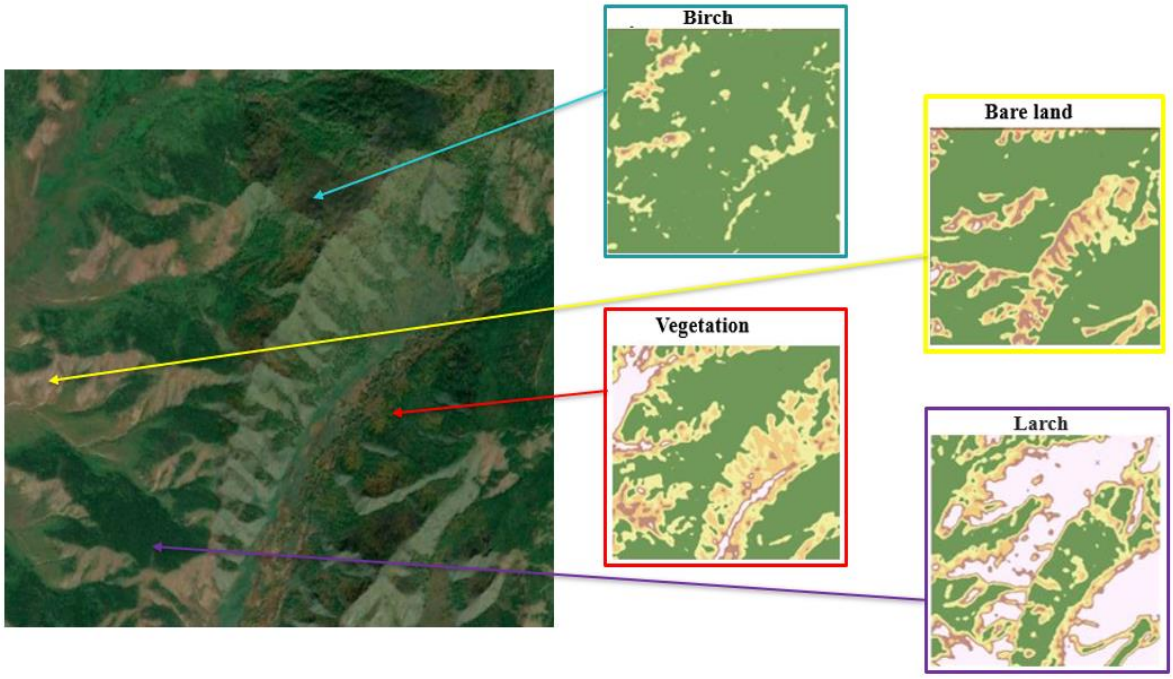

Figure 5. Comparison between the Landsat 8 image and LMM results.

5. CONCLUSION

Over the last few years, different advanced techniques and machine learning algorithms have been developed for the efficient land cover classification. The aim of the research was to develop the LMM and random forest algorithm for classifying land cover classes such as larch, birch, vegetation, and bare land in Khangal soum of Bulgan province, northern Mongolia. As could be seen from the analysis, both techniques could be used for land cover discrimination and other land-related analysis, complementing each other. The overall accuracies of the land cover classifications ranged between 75.2% and 94.4%, however, the random forest application showed better result compared to the LMM. Therefore, we could conclude that the random forest method might be more suitable to further analysis, specially if we consider large areas within the territory of Mongolia.

REFERENCES

Amarsaikhan, D., 2020. Advanced classification of optical and SAR images for urban land cover mapping, International Archives of the Photogrammetry, RS and Spatial Information Sciences, XXIV ISPRS Congress, Nice, France, pp.1417-1421.

Amarsaikhan, D., Ganzorig, M., Saandar, M., Blotevogel, H.H., Egshiglen, E., Gantuya, R., Nergui, B. and Enkhjargal, D., 2012. Comparison of multisource image fusion methods and land cover classification, International Journal of Remote Sensing, Vol.33(8), pp.2532-2550.

Bayanmunkh N., Tsolmon R., Philippe De Maeyer., Tsolmon A.,2019. Estimation methodology for forest biomass in mongolia using remote sensing.

Clevers., JG. P. W. and Zurita-Milla, R., 2008. Multisensor and multiresolution image fusion using the linear mixing model, Academic Press, pp. 67-84,

Enkhmanlai, A., 2022. Application of machine learning approaches for land cover classification using remotely sensed data, Ulaanbaatar, Mongolia, pp.26.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction; Springer Science + Business Media: New York, NY, USA.

Landsat 8, 2013. Landsat science, available at: https://landsat.gsfc.nasa.gov/satellites/landsat-8/

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, No.14, pp.1615-1626.

Otgonbayar, M., Clement, A., Jonathan, C., & Damdinsuren, A., 2018. Mapping pasture biomass in Mongolia using Partial Least Squares, Random Forest regression and Landsat 8 imagery. International Journal of Remote Sensing. 40. 1-23. 10.1080/01431161.2018.1541110.

Pflugmacher, D., Cohen, W.B., Kennedy, R.E. & Yang, Z., 2014. Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. Remote Sensing of Environment, 151:124-137.

Raich, J.W. and Schlesinger, W.H. The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate, Chemical and Physical Meteorology, 4, pp.81-99.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67:93–104.

Tian, X., Yan, M., van der Tol, C., Li, Z., Su, Z., Chen, E., Li, X., Li, L., Wang, X., Pan, X. & Gao, L. 2017. Modeling Forest above-ground biomass dynamics using multi-source data and incorporated models: A case study over the qilian mountains. Agricultural and Forest Meteorology, 246:1–14