EVALUATION OF ADVANCED DATA MINING ALGORITHMS IN LAND USE/LAND COVER MAPPING

Ali Jamali[1]* and Alias Abdul Rahman[2]

[1] Faculty of Surveying Engineering, Apadana Institute of Higher Education, Shiraz, Iran- ali.jamali.65@gmail.com
[2] Universiti Teknologi Malaysia (UTM), Faculty of Geoinformation and Real Estate- alias@utm.my

KEYWORDS: Land use Land cover; LULC; machine learning; data mining; image classification

ABSTRACT:

For environmental monitoring, land-cover mapping, and urban planning, remote sensing is an effective method. In this paper, firstly, for land use land cover mapping, Landsat 8 OLI image classification based on six advanced mathematical algorithms of machine learning including Random Forest, Decision Table, DTNB, Multilayer Perceptron, Non-Nested Generalized Exemplars (NN ge) and Simple Logistic is used. Then, results are compared in the terms of Overall Accuracy (OA), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for land use land cover (LULC) mapping. Based on the training and test datasets, Simple Logistic had the best performance in terms of OA, MAE and RMSE values of 99.9293, 0.0006 and 0.016 for training dataset and values of 99.9467, 0.0005 and 0.0153 for the test dataset.

1. INTRODUCTION

Land cover is one of the main effecting variables of earth climate (Bojinski et al., 2014; Shih, Stow, Weeks, & Coulter, 2016). In the land cover mapping, detailed land cover maps are an essential input for various research groups working on climate change, sustainable development as well as monitoring the agricultural lands. The capability of high-resolution land cover monitoring is improved by using efficient, time series, and cost-effective classification approaches due to the availability of free optical and radar images including Landsat 8 images.

To monitor and analyze human and physical environment, there are accurate and up-to-date Land Use/Landcover (LULC) information. In various fields including health, ecology, agriculture, risk analyzing and management policy, LULC information plays a significant role (Bégué et al., 2018).

Land cover change is a significant factor that connects to climate change. It can affect ecological methods (Vitousek, 1994) as well as the earth conditions which both are related to climatic change (Skole, 1994). Earth observation satellites sensor data is known as an effective factor to research results of climate change. Land cover mapping (Grippa et al., 2018) and analyzing are one of the important use of earth observation satellites sensor data and changing the land cover may affect the climate based on the changing the composition of pollutant emissions like carbon dioxide ((Betts, Falloon, Goldewijk, & Ramankutty, 2007); (Bonan, 2008); (Bala et al., 2007)). For policy and decision making, up-to-date, land use land cover (LULC) statistics are a need which has an effect on economy and society (Costa, Almeida, Vala, Marcelino, & Caetano, 2018).

For image classification methods, there are different methods from unsupervised algorithms including algorithms of K-means clustering in the case of parametric supervised like maximum likelihood (Otukei & Blaschke, 2010); for algorithms of machine learning like artificial neural networks (ANN) and Support Vector Machines (SVMs) ((Duro, Franklin, & Dubé, 2012); (Mountrakis, Im, & Ogole, 2011)), decision trees ((Breiman, 1984); (Hua, Zhang, Chen, Yin, & Tang, 2017)), and classifiers (Breiman, 1996). In comparison to common parametric algorithms for dealing with large and assembled databases, machine learning algorithms are more efficient (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). To generate a thematic map of the land cover, land cover classification is used. It consists of the material at the ground such as water, soil, vegetation, and man-made structures (Fisher & Unwin, 2005).

For land cover mapping , supervised classification methods have better performance compared to unsupervised ((Hansen & Loveland, 2012; Inglada et al., 2017; Khatami, Mountrakis, & Stehman, 2016)), however, they require accurate and sufficient training information. In various researches, machine learning algorithms (e. g., SVMs (V. Vapnik, 1998; V. N. Vapnik, 1995), Random Forests (Breiman, 2001) and ANNs) have been used for classification tasks. Image classification is presented as an image processing method which defines features in each considered image based on their spectral signatures. Each feature has a specific signature that can be called as feature classification (National Aeronautics and Space Administration of USA) (NASA, 2013). In the case of environmental studies, images of Landsat have been widely utilized. Landsat consists of a collection of multispectral satellites previously expanded by the NASA organization.

In this paper, for image classification, six advanced mathematical and machine learning algorithms including Random Forest, Decision Table, DTNB, Multilayer Perceptron, NN ge and Simple Logistic to propose a fit-for-purpose algorithm have been used which are evaluated in terms of Overall Accuracy (OA), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Following this section, study area is introduced in Section 2. Methodology including pre-processing and classification methods are discussed in Section 3. In Section 4, accuracy assessment and validation methods are discussed. Image classification algorithms are ranked in Section 5. In Section 6, conclusion is presented

2. STUDY AREA AND DATA COLLECTION

Information collected by the Landsat 8 OLI satellite (for 22nd August 2018 data collection of Shiraz city case in WGS 84 / UTM area 39N) is presented in Figure 1. Shiraz city is located in the south of Iran and is constructed at the foot of Zagros Mountains (on a green plain), 4,900 feet above sea level. This city consists of a significant number of gardens which is because of clime change, droughts, as well as population growth in the city, many of these gardens are destroyed. Land cover monitoring is commonly considered as a key factor to evaluate protection of garden zones over time. Additionally, in this paper, surface reflectance characteristics of bands were utilized as follows: Blue (0.452–0.512 µm), Green (0.533–0.590 µm), Red (0.636–0.673 µm), Near Infrared—NIR (0.851–0.879 µm), Shortwave Infrared—SWIR 1 (1.566–1.651 µm) and Shortwave Infrared—SWIR 2 (2.107–2.294 µm) as well as the Normalized Difference Vegetation Index (NDVI).

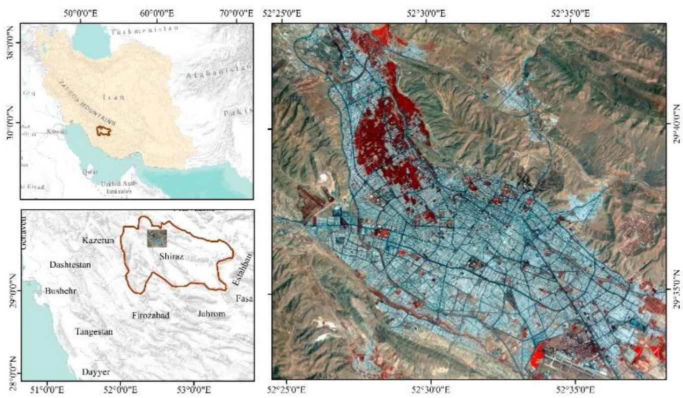

Figure 1. The study area: (a) location map; (b) Shiraz city; (c) Image of the study area in a false-color combination.

3. METHODOLOGY

Methodology of this research is presented in Figure 2. The radiometric and atmospheric effects are corrected in the first step. Second, to represent four LULC classes including built-up areas, bare soil, vegetation, and roads, the reference data which include the training and testing samples were created. Several machine learning algorithms were used to classify the image of the study area in the third step. Fourth, the outputs of best predictive models were statistically assessed. Finally, the results of different image classification algorithms were discussed.

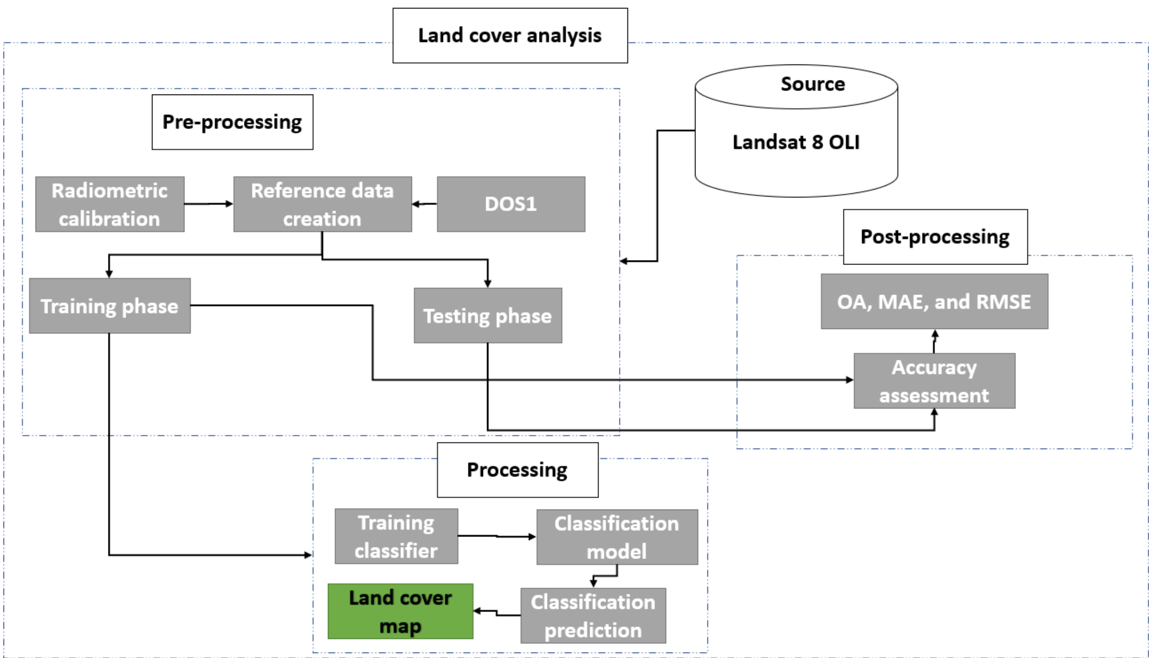

Figure 2. Workflow of this research.

3.1 Preprocessing

In this paper, to predict the reflectance to the ground (ρ) within the pre-processing phase of the pictures, an atmospheric correction (i.e. Dark Object Subtraction) as an image-based atmospheric correction is used.

3.2 Classification

Six advanced machine learning algorithms including Random Forest, Decision Table, DTNB, Multilayer Perceptron, NN ge and Simple Logistic to propose a fit-for-purpose image classification algorithm are used.

3.2.1 Random Forest

Random Forest is an ensemble learning method that is used in the case of land-cover classification of multispectral and hyperspectral satellite sensor imagery. Random Forest generates several trees corresponding to random bootstrapped of the training database patterns. This method performs random binary trees which creates a training subset above bootstrapping method. Additionally, a random choice of the training information is used to generate the model from the initial database, however, out of bag (OOB) is known as the data that is not involved (Catani F., Lagomarsino D., Segoni S., & V., 2013). The tree numbers (n tree) and the variable numbers (m try) are two factors that are required to be adapted in a Random Forest algorithm.

3.2.2 Decision Table

Decision Table (DT) is a classifier that uses a simple DT majority classifier. DTs are one of the most easy to understand hypothesis spaces possible (Kohavi, 1995). It has two parts including a set of characteristics which are involved in the table along with a body including labeled samples of the space specified using the samples. A DT classifier finds exact features in the DT by utilizing only the properties in the schema with taking into consideration of an unlabeled samples. However, it should be noticed that there can be other matching samples in the table.

3.2.3 DTNB

To create and employ a decision table or naive bayes hybrid arranger, DTNB as an appropriate classifier can be used (Hall & Frank, 2008). In this paper, the employed algorithm analyses the merit of separating the properties into two disintegrate parts of the decision table and also the naive bayes. For each phase, a forward option search was used, selected characteristics were modeled using naive Bayes and other by the decision table. Initially, whole properties were modeled using the decision table.

3.2.4 Multilayer Perceptron

Commonly, using several interconnected nodes (i.e. neurons), artificial neural networks are designed. There can be three layers including input layer as well as hidden and output layer in artificial neural network (ANN). Information is required to be classified into three databases as training, validation also test data in order to train the ANN. Training several networks is suggested as the most appropriate method to specify the most proper number of hidden neurons in an ANN algorithm. Neural network of Multi-layer perceptron (MLP) is widely utilized sort of ANNs that identifies itself by using three layers (Hayati, 2018); (Govindaraju & Rao, 2013)).

3.2.5 NN ge

Brent, 1995 introduced the algorithm of Non-Nested Generalized Exemplars (NN ge). NN ge generates generalization task based on the combining samples. However, it creates hyper-rectangles in property space which presents conjunctive rules together with internal disjunction. By connecting this algorithm to its nearest neighbor of the similar class, the algorithm generates an extension, each time a novel sample is added to the dataset, by joining it to its nearest neighbor of the same class.

3.2.6 Simple Logistic

To fit the logistic models, in the simple Logistic algorithm, LogitBoost together with simple regression functions as basis learners were used (Landwehr, Hall, & Frank, 2005). This algorithm was classified in the class learning methods that used additive logistic regression by using instance regression functions as basis learners. This algorithm finds a function which can fit the training information as well, appropriately, using measuring the weights which amplifies the log likelihood function of the logistic regression.

4. ACCURACY ASSESSMENT AND VALIDATION

OA, MAE, and RMSE are used to evaluate the proposed algorithms including Random Forest, Decision Table, DTNB, Multilayer Perceptron, NN ge, and Simple Logistic (see Equations 1 to 3). The numbers of training and testing objects including build-up, soil, roads, and vegetation regions are presented in Table 1.

$$OA = \frac{x}{y} \quad (1)$$

$$MAE = \frac{\sum_{i=1}^{n} |ov_i - pv_i|}{n} \quad (2)$$

$$RMSE = \frac{\sum_{i=1}^{n} (ov - pv_i)^2}{n} \quad (3)$$

Where x is number of correctly classified values;
y is total number of reference values;
ov is observed values and;
pv is predicted values.

Classes	Training set size	Validation set size
Build-up	3528	2606
Soil	633	369
Roads	28138	19502
Vegetation	3057	1895

Table 1: The number of training and testing data for image classification.

Using data mining algorithms, to prevent overfitting issue, which is an undesirable event in utilizing the soft computing approaches, a 10-fold cross-validation technique is used.

5 RESULTS AND DISCUSSION

In this Section, to rank the best classification algorithm, six different advanced mathematical and machine learning techniques namely, Random Forest, Decision Table, DTNB, Multilayer Perceptron, NN ge, Simple Logistic are used. These methods are employed for image classification. Finally, the outputs of image classifications are compared in term of OA, MAE, and RMSE.

5.1 Image classification ranking

The result of proposed methods (e.g., Multilayer Perceptron, Simple Logistic, J48, Lazy IBK, Random Forest, Decision Table, DTNB, NN ge) for both of the training and testing datasets are assessed based on their predictive network results. The result of R^2, MAE, and RMSE for training datasets of proposed methods such as Multilayer Perceptron, Simple Logistic, Random Forest, Decision Table, DTNB, NNge were (99.9321, 99.9293, 99.8529, 98.9536, 99.2534, and 99.703), (0.0007, 0.0006, 0.0015, 0.0171, 0.0051, and 0.0015), and (0.0161, 0.016, 0.0232, 0.0734, 0.0552, and 0.0385), respectively. Similarly, the result of R^2, MAE, and RMSE for testing datasets of proposed methods such as Multilayer Perceptron, Simple Logistic, Random Forest, Decision Table, DTNB, NN ge were (99.9302, 99.9467, 99.8646, 99.1835, 99.5405, and 99.085), (0.0007, 0.0005, 0.0018, 0.0161, 0.0037, and 0.0046), and (0.0183, 0.0153, 0.0244, 0.0683, 0.0454, and 0.0676), respectively. Table 5 presents confusion matrix of the proposed data mining algorithm based on the test dataset where misclassification values are seen. As seen in Table 4, based on the training and test dataset, Simple Logistic classifier has the best performance in terms of OA, MAE, and RMSE with a total value of 35. Multilayer Perceptron classifier is ranked second with a total value of 31. The third-ranked classifier is Random Forest classifier with a total value of 24. DTNB classifier with a total value of 16 is ranked fourth. With a total value of 15, NN ge classifier is ranked fifth. Decision Table has the worst performance with a total value of 8 (see Tables 2 to 5).

	Proposed models	Network results			Ranking the predicted models			Total ranking score	Rank
		R^2	MAE	RMSE	R^2	MAE	RMSE		
1	Multilayer Perceptron	99.9321	0.0007	0.0161	6	5	5	16	2
2	Simple Logistic	99.9293	0.0006	0.016	5	6	6	17	1
3	Random Forest	99.8529	0.0015	0.0232	4	4	4	12	3
4	Decision Table	98.9536	0.0171	0.0734	1	2	1	4	6
5	DTNB	99.2534	0.0051	0.0552	2	3	2	7	5
6	NN ge	99.703	0.0015	0.0385	3	4	3	10	4

Table 2. Image classification algorithms evaluation based on the training dataset.

	Proposed models	Network results			Ranking the predicted models			Total ranking score	Rank
		R^2	MAE	RMSE	R2	MAE	RMSE		
1	Multilayer Perceptron	99.9302	0.0007	0.0183	5	5	5	15	2
2	Simple Logistic	99.9467	0.0005	0.0153	6	6	6	18	1
3	Random Forest	99.8646	0.0018	0.0244	4	4	4	12	3
4	Decision Table	99.1835	0.0161	0.0683	2	1	1	4	6
5	DTNB	99.5405	0.0037	0.0454	3	3	3	9	4
6	NN ge	99.085	0.0046	0.0676	1	2	2	5	5

Table 3. Rank of Image classification algorithms evaluation based on the test dataset.

	Proposed models	Network result						Total score	Rank
		Training dataset			Testing dataset				
		R2	MAE	RMSE	R2	MAE	RMSE		
1	Multilayer Perceptron	6	5	5	5	5	5	31	2
2	Simple Logistic	5	6	6	6	6	6	35	1
3	Random Forest	4	4	4	4	4	4	24	3
4	Decision Table	1	2	1	2	1	1	8	6
5	DTNB	2	3	2	3	3	3	16	4
6	NN ge	3	4	3	1	2	2	15	5

Table 4: Total ranking score and ranking of the proposed classification models based on the both training and testing datasets.

https://doi.org/10.5194/isprs-archives-XLII-4-W16-283-2019 | © Authors 2019. CC BY 4.0 License.

	Model	Build-up	Roads	Soil	Vegetation
Decision Table	Build-up	2265	1	40	0
	Roads	26	299	42	2
	Soil	84	0	19418	0
	Vegetation	3	0	1	1891
Multilayer Perceptron	Build-up	2603	3	0	0
	Roads	2	357	0	10
	Soil	2	0	19500	0
	Vegetation	0	0	0	1895
Simple Logistic	Build-up	2	0	0	2
	Roads	361	0	6	361
	Soil	0	19499	0	0
	Vegetation	0	0	1895	0
Random Forest	Build-up	2596	2	8	0
	Roads	2	360	0	7
	Soil	14	0	19488	0
	Vegetation	0	0	0	1895
DTNB	Build-up	2574	8	24	0
	Roads	1	354	9	5
	Soil	59	5	19438	0
	Vegetation	0	1	0	1894
NNge	Build-up	2542	1	162	1
	Roads	1	322	38	8
	Soil	10	0	19490	2
	Vegetation	0	1	0	1894

Table 5: Confusion matrix of advanced data mining algorithms.

5.2 FIT FOR PURPOSE ALGORITHM

Images classification based on the six advanced mathematical and machine learning algorithms including Random Forest, Decision Table, DTNB, J48, Lazy IBK, Multilayer Perceptron, NN ge, Simple Logistic is presented in Figure 4.

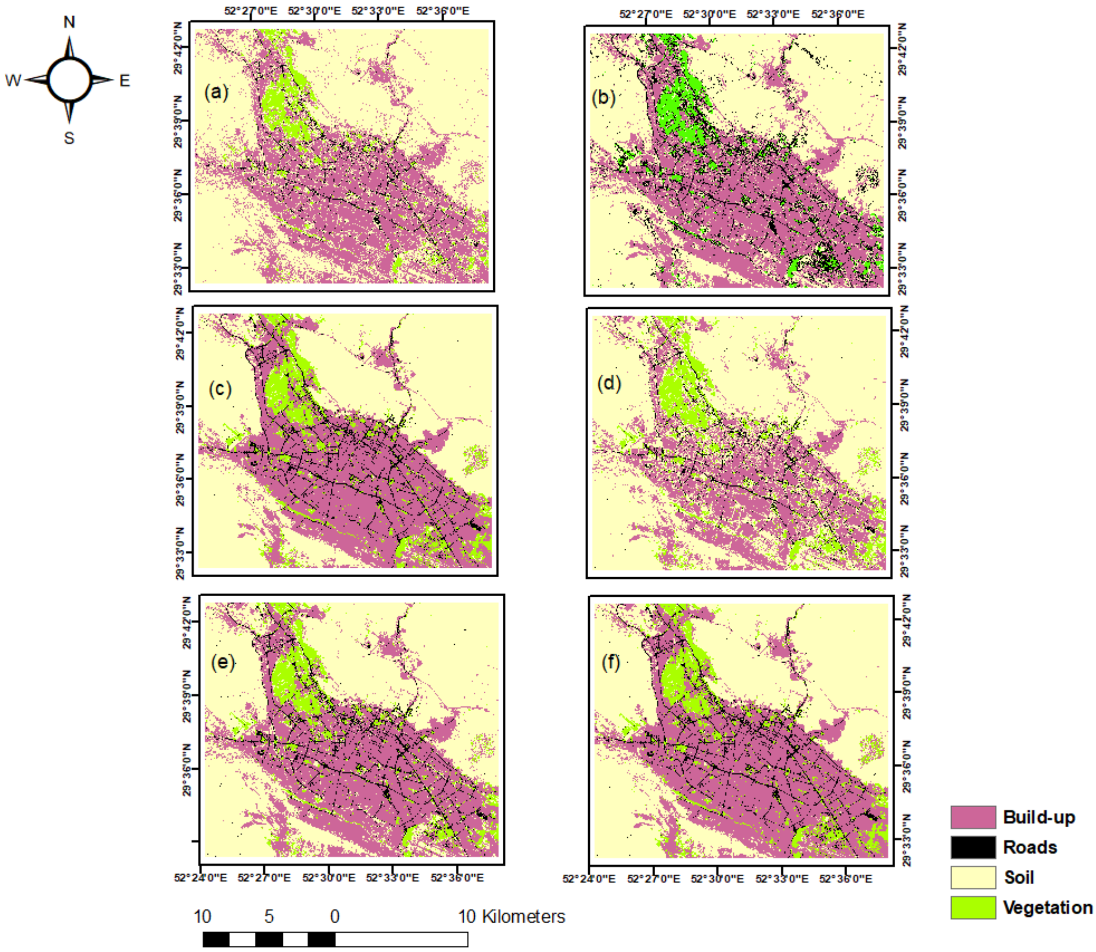

Figure 4. Results of classification algorithms including a) Decision Table b) DTNB c) Multilayer Perceptron d) NN ge e) Random Forest f) Simple Logistic.

6 CONCLUSIONS

Image classification methods in large area environments are highly recommended by researchers as a necessity, due to various climate change phenomena including increase of the temperature of the earth due to pollutant emissions like carbon dioxide (and also their influence on the land cover change and against the influence of land cover changes on earth climate can be analyzed by the method of image classification).

From the above discussion, it is determined that a fit-for-purpose algorithm is required to be suggested for a specific task such as vegetation extraction and flood modeling also man-made zone prediction. In order to identify trees over other materials (for example water and man-made zone), for the Shiraz city case, there is a need to propose an algorithm with high precision for monitoring protection of garden zones.

REFERENCES

Bala, G., Caldeira, K., Wickett, M., Phillips, T., Lobell, D., Delire, C., & Mirin, A. (2007). *Combined climate and carbon-cycle effects of large-scale deforestation.* Paper presented at the Proceedings of the National Academy of Sciences.

Bégué, A., Arvor, D., Bellón, B., Betbeder, J., de Abelleyra, D., Ferraz, R. P. D., . . . Verón, S. R. (2018). Remote Sensing and Cropping Practices: A Review. *Remote Sens, 10*, 99.

Betts, R., Falloon, P., Goldewijk, K., & Ramankutty, N. (2007). Biogeophysical effects of land use on climate: model simulations of radiative forcing and large-scale temperature change. *Agricultural and Forest Meteorology, 142 (2-4)*, 216–233.

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bull. Am. Meteorol. Soc.*, *95*, 1431–1443.

Bonan, G. (2008). Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science, 320 (5882)*, 1444–1449.

Breiman, L. (1984). *Classification and Regression Trees*: Chapman & Hall/CRC.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24 (2)*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning, 45(1)*, 5-32.

Catani F., Lagomarsino D., Segoni S., & V., T. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat Hazards Earth Syst Sci, 13*, 2815–2831.

Costa, H., Almeida, D., Vala, F., Marcelino, F., & Caetano, M. (2018). Land cover mapping from remotely sensed and auxiliary data for harmonized official statistics. *ISPRS International Journal of Geo-Information, 7(4)*, 157.

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment, 118*, 259-272.

Fisher, P. F., & Unwin, D. J., eds. (2005). *Representing GIS*. Chichester, England: John Wiley & Sons.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of the Environment, 80*, 185–201.

Foody, G. M., & Mathur, A. (2004). Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification. *Remote Sens. Environ, 93*, 107–117.

Foody, G. M., & Mathur, A. (2006). The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sens. Environ, 103*, 179–189.

Govindaraju, R. S., & Rao, A. R. (2013). *Artificial neural networks in hydrology* (Vol. 36): Springer Science & Business Media.

Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., . . . Wolff, E. (2018). Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS International Journal of Geo-Information, 7(7)*, 246.

Hall, M. A., & Frank, E. (2008). *Combining naive bayes and decision tables.* Paper presented at the In FLAIRS conference
Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sens. Environ, 122*, 66–74.

Hua, L., Zhang, X., Chen, X., Yin, K., & Tang, L. (2017). A Feature-Based Approach of Decision Tree Classification to Map Time Series Urban Land Use and Land Cover with Landsat 5 TM and Landsat 8 OLI in a Coastal City, China. *ISPRS International Journal of Geo-Information, 6(11)*, 331.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., & Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sens, 9*, 95.

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sens. Environ, 177*, 89–100.

Kohavi, R. (1995). *The power of decision tables.* Paper presented at the In European conference on machine learning
Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59(1-2)*, 161-205.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing, 66 (3)*, 247–259.

NASA. (2013). *Landsat 7 Science Data User's Handbook*. Retrieved from http://landsathandbook.gsfc.nasa.gov

Otukei, J. R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation, 12*, S27-S31.

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.

Shih, H., Stow, D. A., Weeks, J. R., & Coulter, L. L. (2016). Determining the Type and Starting Time of Land Cover and Land Use Change in Southern Ghana Based on Discrete Analysis of Dense Landsat Image Time Series. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 9*, 2064–2073.

Skole, D. L. (1994). Data on global land cover change: acquisition assessment and analysis. In I. Turner, W.B. (Ed.), *Changes in Land Use and Land Cover: A Global Perspective* (pp. 437–471): Cambridge University Press, Cambridge.
Sylvain, R. (2002). *Nearest neighbor with generalization.* University of Canterbury, Christchurch, New Zealand.

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY, USA: Wiley.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory* NewYork, NY, USA: Springer.

Vitousek, P. M. (1994). Beyond global warming: ecology and global change. *Ecology, 75*, 1861–1876.

Revised August 2019