A Fit-for-Purpose algorithm for Environmental Monitoring based on Maximum likelihood, Support Vector Machine and Random Forest

Ali Jamali[1]*

[1] Faculty of Surveying Engineering, Apadana Institute of Higher Education, Shiraz, Iran- ali.jamali.65@gmail.com

Commission III, WG III/10

KEYWORDS: Image classification, Earth Observation, Support Vector Machine, Random Forest, R

ABSTRACT:

Due to concerns of recent earth climate changes such as an increase of earth surface temperature and monitoring its effect on earth surface, environmental monitoring is a necessity. Environmental change monitoring in earth sciences needs land use land cover change (LULCC) modelling as a key factor to investigate impact of climate change phenomena such as droughts and floods on earth surface land cover. There are several free and commercial multi/hyper spectral data sources of Earth Observation (EO) satellites including Landsat, Sentinel and Spot. In this paper, for land use land cover modelling (LULCM), image classification of Landsat 8 using several mathematical and machine learning algorithms including Support Vector Machine (SVM), Random Forest (RF), Maximum Likelihood (ML) and a combination of SVM, ML and RF as a fit-for-purpose algorithm are implemented in R programming language and compared in terms of overall accuracy for image classification.

1. INTRODUCTION

Land cover is a fundamental factor that links and affect with many parts of the human and physical environment (Foody, 2002). The change in land cover is considered as an important factor of global change affecting ecological systems (Vitousek, 1994) with an impact on the earth that is linked with climatic change (Skole, 1994).

Land cover mapping (Grippa et al., 2018) and monitoring is one of key applications of earth observation satellites sensor data which is an important factor to asses results of climate change in the recent years. On the other hand, changes in land cover affect the climate through changes in the composition of greenhouse gasses such as carbon dioxide (Betts et al., 2007; Bonan, 2008; Bala et al., 2007).

According to Rodriguez-Galiano et al. (2012), there are several issues for large area land cover monitoring including:

1. First, complex landscapes are difficult to monitor due to sudden changes in environmental gradients (e.g. moisture, elevation and temperature) and a legacy of past interference (Rogan and Miller, 2006). Such heterogeneous landscapes are defined by land-cover categories that are complicated to be defined spectrally due to low inter-class separability and high intra-class variability.

2. Second, there is a need for algorithms that can be interpreted readily and automated as well as to be easily run with user defined parameters that are simple to adjust.

3. Third, a promising land-cover classification algorithm for large area mapping rely on the capability of the algorithm to work with noisy observations, a complex measurement space, and a few numbers of training data compared to the size of the study area (DeFries and Chan, 2000; Rogan et al., 2008).

A wide range of classification methods have been used to map land cover using remotely sensed data. Classification methods vary from unsupervised algorithms such as K-means klustring to parametric supervised algorithms such as maximum likelihood (Otukei and Blaschke, 2010); to machine learning algorithms such as artificial neural networks (Duro et al, 2012), SVMs (Mountrakis et al., 2011), decision trees (Breiman, 1984; Hua et al., 2017), and ensembles of classifiers (Breiman, 1996).

Dealing with large and complex datasets, machine learning algorithms are more accurate and efficient compared to conventional parametric algorithms, (Rodriguez-Galiano et al., 2012). The usual purpose of land cover classification is to produce a thematic map of the land cover.

Land cover is the material at the ground, such as vegetation, water, soil and man-made structures. (Fisher and Unwin, 2005). The number and kind of land cover classes in the image

* Corresponding author: ali.jamali.65@gmail.com

that can defined vary significantly depending on the sensor resolutions.

An image classification is an image processing technique that according to their spectral signatures, identifies materials in an image. The spectral signature is the reflectance as a function of wavelength where each material has a unique signature, therefore it is called as material classification (NASA, 2013). For environmental research, Landsat images are widely used. Landsat is a set of multispectral satellites developed by the NASA (National Aeronautics and Space Administration of USA), since the early 1970's.

Several pixels and/or regions of pixels of satellite images with known classes as training data are required to predict classes of other regions using the supervised/unsupervised classification model (e.g. Random Forest). Normally, supervised classifications need the user to identify one or more Regions of Interest (ROIs, also Training Areas) for each land cover class identified in the image. ROIs are polygons drawn over homogeneous areas of the image that overlay pixels belonging to the same land cover class.

The spectral signatures (spectral characteristics) of reference land cover classes are defined considering the values of pixels under each ROI having the same Class ID. As a result, the classification algorithm classifies the whole image by comparing the spectral characteristics of each pixel to the spectral characteristics of reference land cover classes.

Assessing the accuracy of land cover classification, in order to identify and measure map errors after the classification process is a usual step. Usually, accuracy assessment is done with the calculation of an error matrix, which is a table that compares map information with reference data for several sample areas (Congalton and Green, 2009).

In this research, for image classification, ML, RF, SVM and a combination of SVM, ML and RF are implemented in R programming language and compared in term of overall accuracy of image classification. Following this introduction, in Section 2, classification methods are discussed. Results of image classification are presented in Section 3. In Section 4, conclusion and future research are discussed.

2. METHODS

Three classification methods including ML, RF and SVM are researched. SVM and RF methods are implemented within R programming language where ML is implemented in QGIS using semi-automatic classification plug-in developed by Luca Congedo.

RF (Breiman, 2001) method is an extension of classification and regression trees (CART; Breiman *et al.*, 1984). RF method is an ensemble learning technique which is increasingly used in land-cover classification using multispectral and hyperspectral satellite sensor imagery. RF creates several trees based on random bootstrapped of the training dataset samples. RF runs random binary trees that creates a subset of the trainings over bootstrapping method, from the initial dataset, a random selection of the training data is selected and implemented to construct the model, out of bag (OOB) is the data which is not included (Catani et al. 2013). The number of trees (ntree), and the number of variables (mtry) are two parameters which are needed to be tuned in a RF method.

SVMs (Vapnik, 1998) uses a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space, but in practice, it does not use any computations in that high-dimensional space. The combined the state-of-the-art performance and simplicity on many learning problems (regression and classification) has increased the popularity of the SVMs (Leo *et al.*, 2006). SVM is a supervised machine learning technique that is implemented based on the Structural Risk Minimization (SRM) principle and statistical learning theory (Tehrany et al. 2015). SVMs have higher accuracies compared with the traditional approaches but the results rely on the kernel used, choice of parameters for the chosen kernel and the method used to generated SVM (Huang et al., 2002).

A maximum likelihood classification algorithm is a parametric supervised classifier. The algorithm for computing the weighted distance or likelihood D of unknown measurement vector X belong to one of the known classes Mc is based on the Bayesian equation (see equation 1).

$$D = \ln(a_c) - [0.5\ln(cov_c) - [0.5X - M_c]T(cov_c - 1)(X - M_c)] \quad (1)$$

The probability distributions for the classes are assumed as a form of multivariate normal models (Richards & Jia, 2006). A sufficient number of pixels are needed for each training area for the calculation of the covariance matrix. The discriminant function, described by Richards and Jia (2006), is calculated for every pixel as (see Equation 2):

$$g_k(x) = \ln p(c_k) - \frac{1}{2} ln |\sum k| - \frac{1}{2} (x - y_k)^t \sum_k^{-1} (x - y_k) \quad (2)$$

where: c_k = land cover class k;
x = spectral signature vector of an image pixel;
$p(c_k)$ = probability that the correct class is c_k;
$|\sum k|$= determinant of the covariance matrix of the data in class Ck;
$\sum_k^{-1}$ = inverse of the covariance matrix;
y_k = spectral signature vector of class k.

The advantage of the ML is that it considers the variance–covariance within the class distributions and for normally distributed data, the ML performs better than the other known parametric classifies (Erdas, 1999). For data with a non-normal distribution, ML may have unsatisfactory results.

3. RESULTS

Shiraz is located in the south of Iran which is built in a green plain at the foot of the Zagros Mountains, 1,500 meters (4,900 feet) above sea level. Shiraz has a considerable number of gardens which due to clime change (i.e. droughts) and population growth in the city, many of these gardens may be lost for new urban city developments. Although the Municipality has taken some measures to preserve these gardens, land cover monitoring is a key factor to monitor preservation of garden regions over time.

As seen in Figure 1, to visualize the Shiraz urban (build-up) growth, Normalized Difference Build-up Index (NDBI) is calculated (see Equation 3).

$$NDBI = (B - SWIR)/(B + SWIR) \quad (3)$$

Where SWIR is Short Wave Infra-Red band and;
B is Blue band.

Most of classification methods take a formula such as $Y = X_1 + X_2$, to find dependent and independent variables where Y is a function of X_1 and X_2. In case of satellites images classification, several levels of classes (i.e. build-in, wet lands, crop) are estimated by a combination of several spectral signature of different bands with/without Normalized Vegetation Difference Index (NDVI) (see Equation 4).

$$NDVI = (NIR - R)/(NIR + R) \quad (4)$$

Where NIR is Near Infra-Red band and;
R is Red band.

Considering six bands plus NDVI for Landsat 8, in R programming language, the classification formula is written by (see Equation 5):

classes ~ band2 + band3 + band4 + band5 + band6 + band7 + NDVI (5)

Landsat images are required to be pre-processed for sensor, solar, atmospheric and topographic effects (Yang et al., 2017). In this research, semi-automatic classification plug-in is used for Dark Object Subtraction (DOS) which is an image-based atmospheric correction. The data for image classification is from the Landsat 8 OLI satellite belonging to 22nd August 2018 data set of Shiraz city in WGS 84 / UTM zone 39N (Figure 2). Images classification using RF (see Figure 3), SVM (see Figure 4), ML (see Figure 5) and a combination of three methods (see Figure 6) are presented.

Figure 1. Normalized Difference Build-up Index (NDBI) for Shiraz city in 1990 (left picture, Landsat 5 image), 2000 (middle picture, Landsat 7 image) and 2018 (right picture, Landsat 8 OLI image).

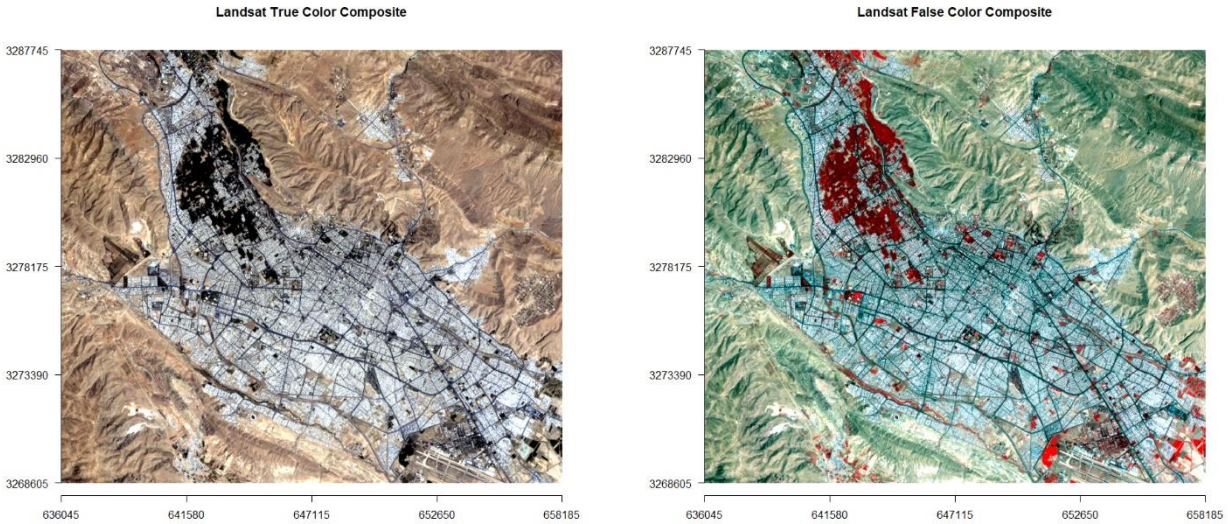

Figure 2. Shiraz city.

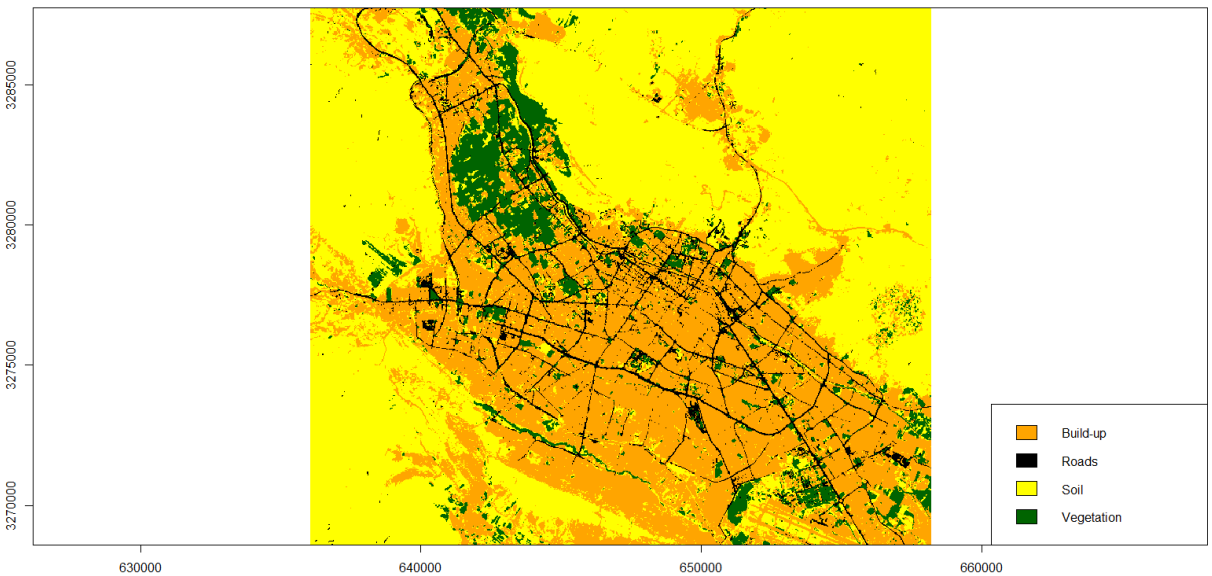

Figure 3. Image classification using Random Forest classification technique within R programming language.

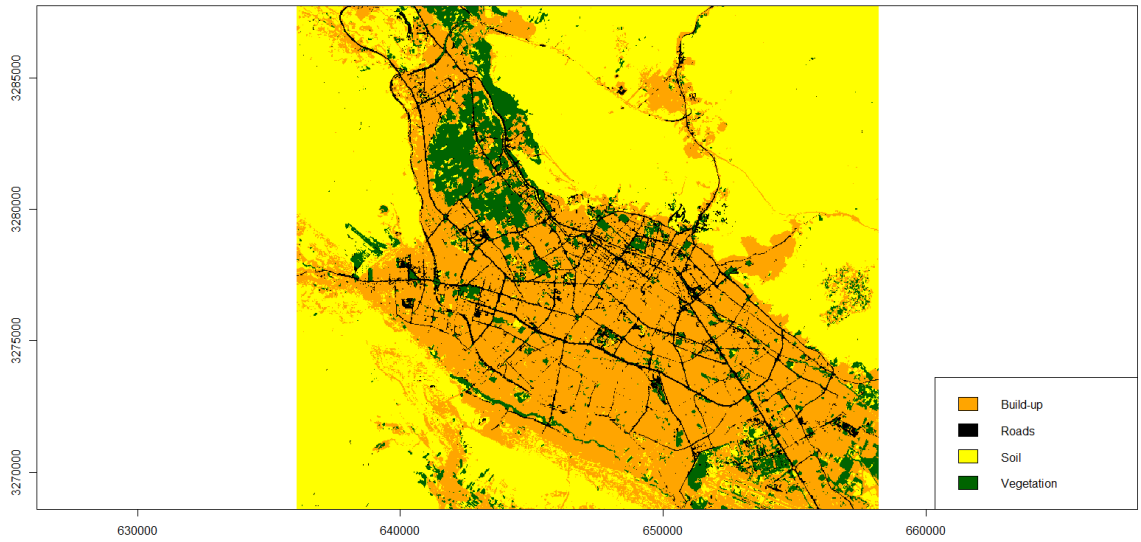

Figure 4. Image classification using SVM classification technique within R programming language.

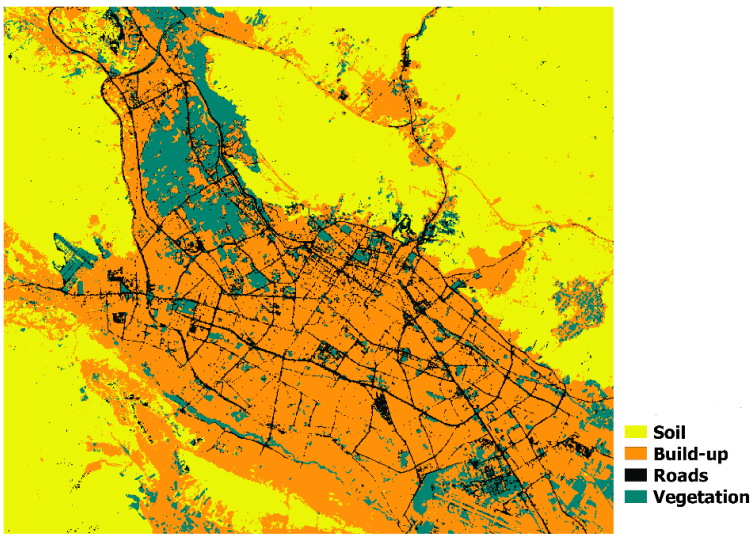

Figure 5. Image classification using ML classification technique with semi-automatic classification plug-in developed by Luca Congedo.

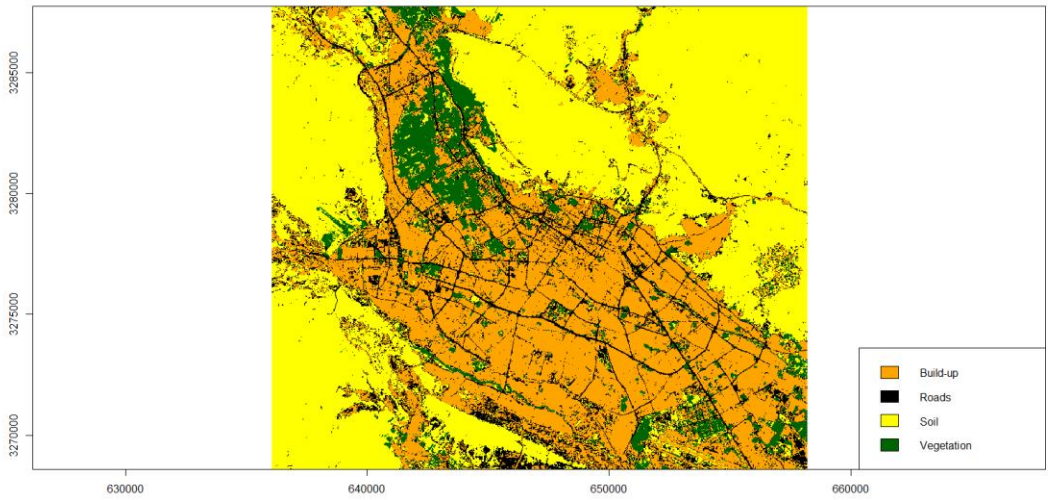

Figure 6. Image classification using a combination of SVM, ML and RF classification techniques within R programming language.

Overall accuracy of image classifications is presented in Tables 1 to 4. Images are classified by four materials including build-up, roads, soil and vegetation regions to evaluate performance of three image classification methods. RF, SVM and ML classification methods have overall accuracies of 99.83, 99.93 and 99.7 percent respectively for the evaluation dataset.

RF	Build-up	Roads	Soil	Vegetation
Build-up	2591	2	17	0
Roads	2	359	0	0
Soil	13	1	19485	0
Vegetation	0	7	0	1895
Overall accuracy	**99.83**			

Table 1. Accuracy of predicted materials by RF classification technique.

SVM	Build-up	Roads	Soil	Vegetation
Build-up	2603	2	4	0
Roads	2	360	0	0
Soil	1	0	19498	0
Vegetation	0	7	0	1895
Overall accuracy	**99.93**			

Table 2. Accuracy of predicted materials by SVM classification technique.

ML	Build-up	Roads	Soil	Vegetation
Build-up	2600	1	44	0
Roads	5	364	0	0
Soil	1	0	19440	0
Vegetation	1	4	18	1895
Overall accuracy	**99.70**			

Table 3. Accuracy of predicted materials by ML classification technique.

An advantage of R programming language is possibility of combining different classification methods considering a fit-for-purpose algorithm within an efficient, feasible and easy-to-use platform. Overall accuracy of the combined method is 99.71 percent for the evaluation dataset (see Table 4).

Combined	Build-up	Roads	Soil	Vegetation
Build-up	2600	3	8	0
Roads	6	360	48	0
Soil	0	2	19446	0
Vegetation	0	4	0	1895
Overall accuracy	**99.71**			

Table 4. Accuracy of predicted materials by a combination of SVM, ML and RF classification techniques.

As the main goal is the identification of garden regions in Shiraz city, the combined method for extraction of vegetation information shows better results compared to RF, ML and SVM methods for the evaluation dataset.

4. CONCLUSIONS

Free and commercial Earth Observation (EO) satellites sensor data are a key factor for large area environmental monitoring. Due to several climate change phenomena (e.g. increase of temperature due to green house gasses) in recent years and their impact on the land cover change and vice versa; the effect of land cover changes on earth climate, image classification for large area environments is a necessity.

R programming language is a powerful and efficient platform for EO researchers to implement their algorithm to monitor and model recent earth climate changes over land cover. A fit-for-purpose algorithms can be designed for certain applications such as vegetation extraction, Flood monitoring

and man-made (build-up) area estimation. Several classification algorithms can be combined to extract certain information to satisfy requirements of an end-user. In the case of the Shiraz city, to monitor preservation of garden regions, an algorithm just for identification of trees over other materials (e.g. water and man-made area) with a high accuracy is required to be constructed.

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. Proceedings of the National Academy of Sciences 104 (16), 6550–6555.

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.

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

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

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

Congalton, R. and Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton, FL: CRC Press.

DeFries, R.S., Chan, J.C.-W., 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment 74 (3), 503–515.

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.

Erdas Inc., 1999. Erdas Field Guide. Erdas Inc., Atlanta, Georgia.

Fisher, P. F. and 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.

Gevrey, M., Dimopoulos, I., & Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

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.

JARS, 1993. Remote Sensing Note. Japan Association on Remote Sensing. Available at http://www.jars1974.net/pdf/rsnote_e.html.

Han, S., Ren, F., Wu, C., Chen, Y., Du, Q., & Ye, X., 2018. Using the TensorFlow Deep Neural Network to Classify Mainland China Visitor Behaviours in Hong Kong from Check-in Data. *ISPRS International Journal of Geo-Information*, *7*(4), 158.

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.

Huang, C., Davis, L.S., Townshed, J.R.G., 2002. An assessment of support Vector Machines for Land cover classification. International Journal of Remote sensing 23, 725–749.

Leo, B., Friedman, J. H., Olshen, R. A., & Stone, C. J., 1984. Classification and regression trees. *Wadsworth International Group*.

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.

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.

NASA, 2013. Landsat 7 Science Data User's Handbook. Available at http://landsathandbook.gsfc.nasa.gov.

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.

Rogan, J., Miller, J., 2006. Integrating GIS and remotely sensed data for mapping forest disturbance and change. In: Franklin, M.W.A.S. (Ed.), Understanding Forest Disturbance and Spatial Pattern: Remote Sensing and GIS Approaches. CRC Press, Boca Raton, FL, pp. 133–172.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., Roberts, D., 2008. Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. Remote Sensing of Environment 112 (5), 2272–2283.

Skole, D.L., 1994. Data on global land cover change: acquisition assessment and analysis. In: Turner, II, W.B. (Ed.), Changes in Land Use and Land Cover: A Global Perspective. Cambridge University Press, Cambridge, pp. 437–471.

Tehrany MS, Pradhan B, Mansor S, Ahmad N., 2015. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena 125:91–101. doi:10.1016/j.catena.2014.10.017.

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

Young, N. E., Anderson, R. S., Chignell, S. M., Vorster, A. G., Lawrence, R., & Evangelista, P. H. (2017). A survival guide to Landsat preprocessing. *Ecology*, *98*(4), 920-932.