Analysis of features importance based on different seasonal OLI images by Random Forest

Ying Zhang¹²*, Lin Wang¹, Yunfei Ai¹, Xuejiao Bai¹

¹ China Transport Telecommunications & Information Center, Beijing 100011, China
² School of Forestry, Beijing Forestry University, Beijing 100083, China
Corresponding author: august12@163.com

Abstract. The paper explores the importance of phenological information of vegetation for vegetation classification. The time series of OLI images in different season are used to obtain the samples of different vegetation types. The importance of single OLI image and the multi-temporal OLI image combinations are separately calculated using Random Forest (RF). The results show that near infrared band (band5) or short-wave infrared band (band6) for single OLI image are important bands to distinguish vegetation types, and Coastal band (band1) also has potential information to distinguish vegetation types. For different time series of OLI image combinations, OLI images in May and July were optimal for vegetation types differentiation. Therefore, the phenological information of vegetation is important to improve the separability of vegetation, and the use of multi-sequence images can improve the classification accuracy of vegetation and achieve rapid and accurate remote sensing forest mapping.

1. Introduction

In the Vegetation classification study, spectrum between vegetation types is easy to confuse due to similar reflection characteristics[1]. Therefore, many features have been considered comprehensively to improve classification accuracy, such as spectral features, vegetation index[2-4], geometric features, image transformation and texture information[5, 6]. Forest vegetation growth is affected by climate, hydrology and soil. Therefore, the vegetation signal characteristics captured by satellite images in different season are different, which correspond to the seasonal variation characteristics of vegetation within a year [7, 8]. Compared with single OLI image, different types of features extracted from different time series of images will provide richer information of vegetation classification.

Due to the advantages of Landsat ETM+/OLI high timeliness, it is used to evaluate the importance of vegetation phenological information. In this paper, the time series Landsat-8 OLI remote sensing images in one year were used for exploring the importance of vegetation spectrum in different season within one year. The temporal sequence data were precisely preprocessed to obtain the reflectance data, and then the original spectral of different vegetation types were extracted. By the feature analysis of random forest (RF), the importance scores for all features were calculated for mono-temporal image and multi-temporal image combinations. According to the sorting results of the features, the importance of mono-temporal image or multi-temporal image combinations were analyzed. Further, phenological information of vegetation was evaluated.

2. Materials and methods
2.1. The study area
The study area is located in Jiangle County, Fujian province, China (Figure 1), which has a central subtropical climate, which is characterized by abundant rainfall and sufficient heat. The annual average temperature is 19.47 °C and the average rainfall 1613.96 mm. The forest species are dominated by *Pinus massoniana*, cedarwood, *Schima superba* and *Castanopsis sclerophylla*. Six vegetation classes including *Cunninghima lanceolate*, *Pinus massoniana*, *Phyllostachys heterocycle*, Broad-leaved forest, needle broad-leaved mixed forest and Crop were located and marked during ground survey.

Figure 1. Study area: Jiangle County, Fujian province, China.

2.2. Data acquisition
In the front of fieldwork, we used Worldview-2 image obtained in December 2015 to select the candidate sample locations in the laboratory. A total of 261 field samples were surveyed and 2355 sample pixels were obtained. Digital elevation model (DEM) data with 30 m spatial resolution and Landsat-8 Operational Land Imager (OLI) images were used in this research. The OLI images used were listed in the Table 1, respectively. The OLI images were radiometric calibrated, which the OLI DN (Digital Numbers) were converted to at-sensor reflectance with an apparent reflectance model by ENVI tools. Then, the OLI images were rectified with 23 control points which were collected from the corrected Worldview-2 image with high spatial resolution. The root mean square error (RMSE) was controlled within a pixel.

| Product ID | abbreviation | acquisition time | sun elevation | sun azimuth |
|------------|--------------|-----------------|---------------|-------------|
| LC81200422014034LGN00 | OLI-034 | 20140203 | 40.567 | 146.285 |
| LC81200422015133LGN00 | OLI-133 | 20150513 | 68.029 | 106.030 |
| LC81200422013191LGN00 | OLI-191 | 20130710 | 68.055 | 94.451 |
| LC81200412014290LGN00 | OLI-290 | 20141017 | 49.183 | 150.639 |
| LC81200412014354LGN00 | OLI-354 | 20141220 | 35.056 | 155.239 |

2.3. The method for feature importance analysis
RF model based on multiple decision trees has high application efficiency for classification and regression[9, 10]. It is high efficiency for processing a large number of features. By the feature
analysis of random forest (RF), the importance scores for all features were calculated for mono-temporal image and multi-temporal image combinations.

3. Results
The importance score of all features of each time phase was calculated by random forest. Figure 2 shows the importance sequence of spectral features of the original band of each time phase. It shows that the most important features of OLI-034, OLI-290 and OLI-354 are short-wave infrared band (band6). The most important characteristics for OLI-133 and OLI-191 are the near infrared wave band (band5), which has the importance score of 106 and 111. The near infrared band 5 for OLI-034, OLI-290 and OLI-354 is the second important characteristic. Therefore, the near infrared band (band5) or short wave infrared wavelengths (band6) contain abundant vegetation discrimination information. In the time-series images of OLI-034, OLI-290 and OLI-354, the importance of band1 is also relatively high. The importance scores are 57, 66 and 54, indicating that the Coastal band also has potential information to distinguish vegetation types.

![Figure 2. Mean relative feature importance of the 10 most important original spectral features for each image](image)

Figure 2 shows the sequence of the top 10 important features for different OLI image combinations. In different time sequence lengths, it can be seen that spectral features of near-infrared band (band5) and short-wave infrared band (band6) still have relatively richer vegetation discrimination information compared with other bands. For example, the single OLI-034 image has the most important feature of band5 and band6. For the combination of OLI-034-133 images, the relatively important features are
band5 and band7 of OLI-133 images. The most important features for 3 time series combinations (034-133-191), 4 time series combinations (034-133-191-290) and 5 time series combinations (034-133-191-290-354) were band5 of OLI-133 and OLI-191 images. Relative to other OLI images, OLI-133 and OLI-191 images have more abundant vegetation information. With the increase of the length of the sequence, the OLI-133-191 spectral characteristics are always in the most important place. Especially for the OLI-191 image, the score of its near infrared band (band5) is significantly higher than that of any other images band spectrum.

Figure 3. Mean relative feature importance of the 10 most important original spectral features for different image combinations

4. Conclusions
The importance of the original band spectrum was separately analyzed for each images and multi-image combinations. The analysis of band spectrum shows that near-infrared band (band5) or short-wave infrared band (band6) for each image has rich information for distinguishing vegetation types. For analysis of multi-image combinations, the spectral features in the OLI-133(May 13, 2015) image and OLI-191(July 10, 2013) image have more abundant vegetation information. When the time series is longer than three, the top ten features only contain band spectra features of OLI-133 (May 13, 2015) and OLI-191. Therefore, the phenological information of vegetation is important to improve vegetation classification.
Acknowledgment
This work was financially supported by the China Transport Telecommunication & Information Center reserve project in 2018 (2018CB09) and the National Key Research and Development Program of China (2017YFC0803900). The author would like to express thanks to anonymous reviewers and the editors for their remarks.

References
[1] Fu A, Sun G, Guo Z, Wang D 2010 IEEE Journal of Selected Topics in Applied Earth Observations 3(2) 178-189
[2] Chen C, Zhao B, Li S, Ping T, Wu H 2015 International Journal of Remote Sensing 36(14) 3701-3713
[3] Hao P, Wang L, Niu Z 2015 Plos One 10(9) e0137748
[4] Huete A R 1988 Remote Sensing of Environment 25(3) 295-309
[5] Lu D, Weng Q 2007 International Journal of Remote Sensing 28(5) 823-870
[6] Chen D, Stow D A, Gong P 2004 International Journal of Remote Sensing 25(11) 2177-2192
[7] Li Z G, Tang H J, Yang P, Wu W B, Chen Z X, Zhou Q B, Zhang L, Zou J Q 2012 Chinese Journal of Agricultural Resources and Regional Planning 33(5) 20-28
[8] Gamon J A, Peñuelas J, Field C B 1992 Remote Sensing of Environment 41(1) 35-44
[9] Sandri M, Zucolotto P 2012 Pattern Recognition Letters 31(14) 2225-2236
[10] Cutler A, Cutler D R, Stevens J R 2004 Machine Learning 45(1) 157-176