A Feature Selection Method based on the Pearson’s Correlation and Transformed Divergence Analysis

Ying Zhang\textsuperscript{1,2}, Lin Wang\textsuperscript{1}, Danyang Geng\textsuperscript{1}, Yunfei Ai\textsuperscript{1}, Wei Xia\textsuperscript{1}, Xuejiao Bai\textsuperscript{1} and Shikai Sun\textsuperscript{1}

\textsuperscript{1} China Transport Telecommunications & Information Center, Beijing 100011, China
\textsuperscript{2} Beijing Forestry University, Beijing 100083, China

august12@163.com

Abstract. An improved feature selection method has been presented, which is based on Transformed Divergence (TD) considering weights of classes and Pearson’s correlation analysis. Using the improved method, this study evaluated several derived vegetation indices and texture measures based on Landsat-8 OLI data to determine their effect on improving the land cover classification separability in Jiangle county, Sanming city of Fujian province, China. The best vegetation indices combination was selected by the improved feature selection method and likewise, best textural combinations at different spatial resolution levels from multi-spectral bands or panchromatic band were obtained. The improved feature selection method found that a single feature could not maximize the separability of vegetation classes. When selecting four vegetation indices, the separability of vegetation classes can be maximized significantly; and two textural measures were suitable for maximizing the separability of vegetation classes. Overall, the result verifies that the feature selection method considering weights of classes and Pearson’s correlation coefficient can select optimal features to maximize class separability.

1. Introduction
Land-cover classification using remotely sensed image has been a hot study topic, which has attracted great attention [1, 2, 3, 4]. Landsat (TM, ETM+ and OLI) images are the most frequently used for classification and monitoring because of its free acquisition and high time phase [5]. The Landsat record has been extended because of the Landsat 8 Operational Land Imager (OLI). The OLI sensor has improved signal-to-noise performance and has been added new spectral bands in the blue and cirrus cloud-detection portion of the spectrum. Compared to the ETM+ panchromatic band, the OLI panchromatic band has a narrower wavelength range, which is more suitable to distinguish vegetation from non-vegetation area. Therefore, OLI data has greater potential for improving vegetation mapping.

Many vegetation indices and texture features based on remotely sensed images have been developed to improve classification performance in the area of thematic mapping [6, 7, 8, 9, 10, 11]. Using vegetation indices (VI) has many advantages relative to single original band, such as enhancing spectral differences for plants, attenuating atmospheric effect, soil background and shadow. Image textures are also an effective feature in improving the classification result. The common used texture measures are Gray Level Co-occurrence Matrix method (GLCM), which is extracted using different window sizes [12]. However, the application of a large volume of features inevitably increases the
computational time and the classification accuracy has little improvement. To select the suitable features, many researchers have developed a large number of feature selection methods to select optimal feature or features combination [13, 14, 15].

Selection of feature algorithm is an important hot topic in land-cover classification[12]. Many features have similar information and are not all useful for classification. To increase the separability of categories, it is necessary to use suitable method to identify the optimal features. Though most previous studies have used VI and texture measures for thematic mapping, the influence from sample amounts of different classes is not considered and the number of best features [13] used remains unclear for different features types. The objectives of this research are to develop a feature selection method for selecting optimal VIs and textural features in the study area. In addition, evaluate how many features are suitable for vegetation classification when using different feature types.

2. Data and processing
This study area is located in Jiangle county, FuJian province, Southeast China (see Figure 1). The study area 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.

Four vegetation classes including coniferous forest, broad-leaved forest, mixed forest, farmland were located and marked during ground survey. 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. Table 1 displays the amounts of training and validation samples for each class.

The sensor data, Landsat-8 OLI, was used in this research. The OLI sensor collects data using 9 shortwave spectral bands and the multispectral bands have a spatial resolution of 30 m. The panchromatic band has a spatial resolution of 15 m. The OLI image used in the study was produced on 17 October 2014. The sun elevation angle and sun azimuth angle are 50.309 and 149.441°, respectively. A image-to-image geometric correction was performed for the OLI image and the correction error is 0.351 pixels. The image was resampled to 30-meter of spatial resolution. The C correction model was used for reducing topographic effects. Subsequently, FLAASH atmospheric correction were performed and then the reflectance was obtained.

![Figure 1. Study area - Jiangle county, Fujian province, China.](image-url)
Table 1. Number of field samples, training and validation samples.

| Types             | Surveyed Fields | Training(Pixels) | Validation(Pixels) |
|-------------------|-----------------|------------------|--------------------|
| Coniferous Forest | 37              | 120              | 226                |
| Broad-Leaved Forest | 27              | 87               | 134                |
| Mixed Forest      | 38              | 112              | 265                |
| Farmland          | 43              | 126              | 298                |
| Felling Area      | 23              | 66               | 134                |
| Built-Up Area     | 36              | 132              | 187                |
| Bare Land         | 28              | 87               | 144                |
| Water             | 29              | 72               | 165                |
| Total             | 261             | 802              | 1553               |

3. Method

3.1. General methodology
First, the VIs and textural measures were obtained using the reflectance image. And then, transformed divergence algorithm (TD) considering weights of classes and correlation analysis were used to select suitable vegetation indices and texture measures at different spatial resolution levels.

3.2. The improved feature selection method
Transformed Divergence (TD) distance is an accurately quantitative indicator for classes separability, which was proved in many studies [16]. However, TD distance does not consider the weights of different classes influenced by sample amounts of different classes. So, accurate separability of multi-classes necessarily takes into account inconsistent sample number for all classes. In the study, the TD_{bh} distance considering weights of sample is used and it is calculated by Equation (2).

\[ TD_{ij} = 2(1-\exp(-\frac{D_{ij}}{8})) \]

\[ TD_{bh} = \sum_{i=1}^{N} \sum_{j=1}^{N} p(w_i) \times p(w_j) \times TD_{ij}^2 \]

\[ D_{ij} = \frac{1}{2} tr((C_i - C_j)(C_i^{-1} - C_j^{-1})) + \frac{1}{2} tr((C_i^{-1} - C_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T) \]

where \( i \) and \( j \) is the class, \( TD_{ij} \) is the transformed divergence distance of two classes, \( D_{ij} \) is the divergence distance of two classes, \( C_i \) and \( C_j \) are covariance matrixes of two classes, \( \mu_i \) and \( \mu_j \) are the mean vector, \( T \) is transpose of the matrices, \( N \) is the number of classes, \( p(w_i) \) and \( p(w_j) \) are prior probabilities of two classes \( i, j \).

The prior probabilities can be calculated using sample data in Table 1.

The features with greater TD_{bh} values were firstly selected using the equation 2. Then the correlation analysis was performed using Pearson’s correlation analysis. The features with high separability and low correlation were finally selected for classification. When two or more features were selected, the best feature combination is identified by the following equation (3).

\[ BFC = \sum_{i=1}^{m} TD_{bh_i} \times \sum_{j=1}^{m} R_{ij} \]

Where \( TD_{bh_i} \) is the TD value considering the prior probabilities of samples on the feature \( i \), \( R_{ij} \) is the correlation coefficient between two features, and \( m \) is feature number.
3.3. Selection of vegetation indices

Primary purpose of feature selection is enhancing separability of vegetation types in this paper. Therefore, the separability of features was evaluated only using training samples for four vegetation classes.

Some research found that vegetation indices containing near infrared and mid-infrared bands were correlated with forest vegetation parameters [17]. Therefore, this paper selects vegetation indices consist of OLI band 5, 6 and 7, as far as possible. The used VIs are included in Table 2 [17]. These VIs were divided into four groups for convenient analysis [18].

| Type                        | No. | Vegetation indices | Equation                                                                 |
|-----------------------------|-----|--------------------|--------------------------------------------------------------------------|
| Simple ratio                | 1   | OLI5/4             | OLI 5/ OLI 4                                                              |
|                             | 2   | OLI6/4             | OLI 6/ OLI 4                                                              |
|                             | 3   | OLI6/5             | OLI 6/ OLI 5                                                              |
|                             | 4   | OLI6/7             | OLI 6/ OLI 7                                                              |
|                             | 5   | NDVI               | (OLI5 – OLI4)/(OLI5 + OLI4)                                               |
|                             | 6   | NDWI               | (OLI5 – OLI6)/(OLI5 + OLI6)                                               |
|                             | 7   | ND6-4              | (OLI6 – OLI4)/(OLI6 + OLI4)                                               |
|                             | 8   | ND6-7              | (OLI6 – OLI7)/(OLI6 + OLI7)                                               |
|                             | 9   | ND4-3              | (OLI4 – OLI3)/(OLI4 + OLI3)                                               |
|                             | 10  | ND5-3              | (OLI5 – OLI3)/(OLI5 + OLI3)                                               |
|                             | 11  | ND5-36             | (OLI5 – OLI3 – OLI6)/(OLI5+ OLI3+ OLI6)                                   |
|                             | 12  | ND53-64            | (OLI5+ OLI3 – OLI6 – OLI4)/(OLI5+ OLI3+ OLI6+ OLI4)                        |
| Normalized vegetation indices | 13  | ND53-67            | (OLI5+ OLI3 – OLI6–OLI7)/(OLI5+ OLI3+ OLI6+ OLI7)                         |
|                             | 14  | ND5-46             | (OLI5 – OLI4 – OLI6)/(OLI5+ OLI4+ OLI6)                                   |
|                             | 15  | ND56-34            | (OLI5+ OLI6 – OLI3 – OLI4)/(OLI5+ OLI6+ OLI3+ OLI4)                        |
|                             | 16  | ND5-67             | (2 OLI5 – OLI6 – OLI7)/( OLI5+ OLI6+ OLI7)                                |
|                             | 17  | ARVI               | (NIR-2RED+BLUE)/(NIR+2RED-BLUE)                                          |
|                             | 18  | ASVI               | (((2NIR+1)−√((2NIR+1)−8(NIR-2RED+BLUE))/2                                  |
| Complex vegetation indices  | 19  | MSAVI              | (((2NIR+1)−((2NIR+1)−8(NIR-2RED))/2                                      |
|                             | 20  | GEMI               | ξ=[1–0.25ξ−(RED-0.125)/(1-RED)]                                           |
|                             |     |                   | ξ=2[(NIR²-RED²) + 1.5(NIR+0.5RED) ]/(NIR+RED+0.5)                          |
| Image transform             | 21  | VIS324             | OLI2+OLI3+OLI4                                                            |
|                             | 22  | MID67              | OLI6+OLI7                                                                 |
|                             | 23  | Albedo             | OLI2+OLI3+OLI4+OLI5+OLI6+OLI7                                            |
|                             | 24  | TC1                | 0.304OLI2+0.279 OL13+0.474 OL14+0.559 OL15+0.508 OLI6+0.186 OLI7           |
|                             | 25  | TC2                | -0.285 OLI2–0.244 OL13–0.544 OLI4+0.704 OL15+0.084 OLI6–0.180 OLI7        |
|                             | 26  | TC3                | 0.151 OLI2+0.197 OL13+0.328OL14+0.341 OLI5–0.711 OLI6–0.457 OLI7           |

3.4 Selection of texture measures

The texture measures consisting of mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment, and correlation are used in this paper, which is obtained based on OLI band 5, band 6,
band 7 and panchromatic band 8. The moving windows include 3 × 3, 5 × 5, 7 × 7, 9 × 9, 11 × 11, 15 × 15, 19 × 19, 25 × 25, and 31 × 31.

4. Result

4.1. Separability analysis

The improved TD distances considering weights of vegetation sample were used to analyze the separability between different vegetation classes for each feature. The separability analysis values for all VI and textural measures were showed in Figure 2 and Figure 3, respectively. Not all features have high separability values for vegetation types. The separability distances for VI (Figure 2) show that the simple ratios, OLI5/4 and OLI6/5, had the greatest potential to distinguish vegetation classes.

The separability distances for textural features varied greatly with bands and the size of moving window (Figure 3). In band 5, the textural feature based on entropy with 11×11 window size pixels had a better separability performance. In band 6, the entropy with 19×19 window size pixels had a higher $TD_{bh}$ distance. In band 7, the texture based on mean with 19×19 window size pixels provide best separability for vegetation classes. In panchromatic band, the texture based on second moment with 31×31 window size pixels provide best separability for vegetation classes. Different texture measures from different bands have various capabilities in separating vegetation classes. The entropy and second moment measures based on band 5 have high levels of separability and in band 6, the entropy, the second moment and mean measures have relatively high separability value. However, in the panchromatic band 8, the variance, second moment and correlation have a better performance in identifying the vegetation types. Therefore, there exists a best combination of texture measure and size of moving window in different spectral range suitable for the separation of vegetation.

![Figure 2. $TD_{bh}$ distance values for each vegetation indices based on selected training sample plots of different vegetation classes. Notes: the number in the x-axis is the same as the number of the Table 2.](image-url)
Figure 3. $TD_{0}$ distance values for each textural image from different bands based on selected training sample plots of different vegetation classes.

4.2. The optimal features combination

For determining the best number of feature combination, the single feature with the highest $TD_{0}$ distance was selected. Then, correlation coefficients analysis between vegetation indices and between the textural images identified least correction features. Finally, the optimal combination features were obtained by the feature selection method. The features with the best capability in vegetation separation but less correlation were identified using Best Feature Combination (BFC) analysis. Figure 4 shows the maximum $TD_{0}$ distances of different feature combination for VIs and textural measures from multi-spectral bands or panchromatic band. The identified four images of vegetation indices were from OLI6/5, ND6-4, ND53-67 and MSAVI. In the multi-spectral bands (OLI band5, 6 and 7), the selected two textural measures were from the entropy with $11 \times 11$ window size based on OLI band5 and second moment with $7 \times 7$ window based on OLI band7. And in the panchromatic band8, the selected two textural measures were from the variance with $5 \times 5$ window and correlation with $15 \times 15$ window. Four feature combinations for vegetation indices and two feature combinations for textural measures from
multi-spectral bands or panchromatic band were selected as optimal combinations for vegetation classification.

![Figure 4](image-url)  
**Figure 4.** $TD_{bh}$ distance values for VI and each textural image from different bands based on selected training sample plots of different vegetation classes.

5. Conclusion  
This research proposes an approach to select optimal feature by TD analysis considering the weights of different classes and band correlation coefficient. For a large number of features, the approach provides an easy method to obtain the optimal feature combinations to improve the class separability. Our study found that four vegetation indices combination are suitable for improving vegetation classification performance, while two features combination from textural images based on OLI multi-spectral bands or OLI panchromatic band are suitable for improving vegetation separability.

Acknowledgments  
This work was financially supported by The National Key Research and Development Program of China(2017YFC0803900). The author would like to express loyal thanks to reviewers and editors for their work.

References  
[1] D. Lu, et al., *Linear mixture model applied to Amazonian vegetation classification*. Remote Sensing of Environment, 2003. 87 (4): p. 456-469.
[2] D. Chen, et al., *Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case*. International Journal of Remote Sensing, 2004. 25 (11): p. 2177-2192.
[3] Y. Zhang, et al., *Coastal wetland vegetation classification with a Landsat Thematic Mapper image*. International Journal of Remote Sensing, 2011. 32 (2): p. 545-561.
[4] Z. Gong, et al., *Estimating wetland vegetation abundance based on spectral mixture analysis: a comparison between LSMA and FCM classification methods*. International Journal of Remote Sensing, 2013. 35 (1): p. 189-203.
[5] D. Lu, et al., *Fractional forest cover mapping in the Brazilian Amazon with a combination of MODIS and TM images*. International Journal of Remote Sensing, 2011. 32 (22): p. 7131-7149.
[6] J. Dash, et al., *Land cover classification using multi-temporal MERIS vegetation indices*. International Journal of Remote Sensing, 2007. 28 (6): p. 1137-1159.
[7] J. Oldeland, et al., Combining vegetation indices, constrained ordination and fuzzy classification for mapping semi-natural vegetation units from hyperspectral imagery. Remote Sensing of Environment, 2010. 114 (6): p. 1155-1166.

[8] L. Zhong, et al., Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. Remote Sensing of Environment, 2014. 140 (1): p. 1-13.

[9] R.S. Fletcher. Using Vegetation Indices as Input into Random Forest for Soybean and Weed Classification. 2016. 07 (15): p. 2186-2198.

[10] Z. Wentao, et al., Texture classification of vegetation cover in high altitude wetlands zone, in: IOP Conference Series: Earth and Environmental Science, 2014, pp. 682-691.

[11] V.K. Alilou, Texture Classification. International Journal of Systems Science, 2014. 28 (7): p. 669-682.

[12] G. Li, et al., A comparative analysis of ALOS PALSAR L-band and RADARSAT-2 C-band data for land-cover classification in a tropical moist region. ISPRS Journal of Photogrammetry and Remote Sensing, 2012. 70 (0): p. 26-38.

[13] D. Lu, et al., A Comparative Study of Landsat TM and SPOT HRG Images for Vegetation Classification in the Brazilian Amazon. Photogrammetric Engineering & Remote Sensing, 2008. 74 (3): p. 311-321.

[14] M. Pal, G.M. Foody, Feature Selection for Classification of Hyperspectral Data by SVM. Geoscience & Remote Sensing IEEE Transactions on, 2010. 48 (5): p. 2297-2307.

[15] Y. Zhang, et al., Feature selection for classification with class-separability strategy and data envelopment analysis. Neurocomputing, 2015. 166 (C): p. 172-184.

[16] P.W. Mausel, Optimum band selection for supervised classification of multispectral data. Photogramm Eng Remote Sens, 1990. 56 (1): p. 55-60.

[17] G. Li, et al., Land-cover classification in a moist tropical region of Brazil with Landsat Thematic Mapper imagery. International Journal of Remote Sensing, 2011. 32 (23): p. 8207-8230.

[18] D. Lu, et al., Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin. 2004. 198 (1): p. 149-167.