

Analysis of the Influence of Vegetation Index Choice on the Classification of Satellite Images for Monitoring Forest Pathology*

Evgeniy Trubakov ¹[0000-0002-8381-9737] and Olga Trubakova ¹[0000-0003-4057-5362]

¹ Bryansk State Technical University, Bryansk, Russia
trubakoveo@gmail.com, trubakovaor@gmail.com

Abstract. Rational use of natural resources and control over their recovery, as well as over destruction due to natural and technogenic causes, is currently one of the most urgent problems of the humanity. Forests are no exception. Multi-spectral images from Earth's satellites are most often used for monitoring changes in forest planting. This is due to the fact that merging images taken in certain spectra makes it possible to recognize vegetation containing chlorophyll quite well. It also allows to detect changes in the level of chlorophyll, which shows the differences between healthy and damaged plants. Large areas of planted forests create the need to process huge amounts of data, which is difficult to do manually. One of the most important stages of image processing is the classification of objects in these images. This paper deals with various classification methods used to solve the problem of classifying images of remote sensing of the Earth. As a result, it was decided to evaluate the accuracy of classification methods on various vegetation indices. In the course of the study, the evaluation algorithm was determined, as well as one of the options for analyzing the results obtained. Conclusions were made about the work of classification methods on different vegetation indices.

Keywords: Remote Sensing of the Earth, Forest Pathology Monitoring, Vegetation Indices, Image Processing, Methods of Image Classification.

1 Introduction

Today, wood remains a very valuable material in many industries, so deforestation has become a profitable business. This often happens illegally, without control, without taking into account the damage to forest plantings and the environment. Also, major damage to the forest is caused by natural phenomena, such as droughts or windfalls, forest pathologies such as tree diseases or insect pests, which is a bigger problem. In addition, forest fires also cause great damage to forests, destroying more than a million

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

* Publication supported by RFBR grant № 19-07-00844

hectares of forest per year. For this reason, it is necessary to monitor the state of the forest constantly [1].

The main source of data for monitoring the state of forests is digital images obtained by artificial earth satellites. Because of vast forest territories, it is necessary to track dozens of images for one region, and taking into account their updating (for example, for Sentinel-2 satellite system every 2-3 days), the volume of processed information increases tenfold [2].

At the moment, the monitoring system operation can be divided into three parts. The first part consists of selecting a suitable satellite image, which will be a reference. The essence of this stage is to search for an image in which the region of interest will not be blocked by interference, such as clouds, cloud shadows, and so on [3]. The next stage is processing of space images. The problem while working with satellite images is that the image is taken in different spectra that are difficult to be processed by humans, so it is necessary to pre-process the image, i.e. to construct a vegetation index. Then, in order to search for objects of interest in the image, we need to make a training sample, classify the satellite image, and vectorize the classification results. After, the described actions should be performed for another image obtained after a period of time for the same territory. The final step is to compare the results of work on the reference image and the new one. This algorithm is iterative and repeats throughout the vegetation season [4]. The complexity of the algorithm is that it is necessary to involve experts to process images, since monitoring systems are not able to identify problem regions automatically.

The relevance of this topic is due to the fact that forest monitoring involves checking large amounts of data received from satellites. At the same time, most of the work is performed manually and takes a long time, so it is necessary to execute some stages of data processing semi-automatically or automatically. For example, the region for monitoring may be blocked by clouds or other interference, but the operator will spend time performing this step. Therefore, it is necessary to investigate methods for identifying space images suitable for monitoring and automate this stage. And given that the detection of forest pathologies by remote means is based on the fact that the stressed tree is vegetatively dries out, a big problem is the shortest possible time to identify pathologies and eliminate them. Therefore, it is necessary to reduce the operator's inefficient working time as much as possible.

2 Vegetation indices

Almost all satellite systems provide medium and high-resolution images in the form of multispectral images. This feature of such images allows to select channels that provide more information about the typical objects under study, i.e. cut off information about extraneous objects from the image and emphasize the data for the task being solved. The selected channels are combined according to certain rules, forming a single image. This procedure is the first in space image processing, so it is performed on all images

used in monitoring [5]. Since it is necessary to recognize vegetation in images to monitor the forest, specialized methods for merging image parameters – vegetation indexes are used.

Vegetation index is an indicator calculated as a result of operations with different spectral data ranges (channels) of remote sensing, and it is related to vegetation parameters in a given pixel of the image.

Let us consider the most common and well-established indices that are used in research.

2.1 Normalized Difference Vegetation Index

Normalized Difference Vegetation Index (NDVI) is the most popular and frequently used vegetation index, which takes positive values for vegetation, and the larger the green phytomass, the higher the index is [6]. The index values are also affected by the species composition of vegetation, its closeness, state, exposure, the angle of the surface, and the color of the soil under thinned vegetation. NDVI is often used as one of the tools for conducting complex types of analysis, which can result in maps of forest and agricultural productivity, maps of landscapes and natural zones, soil, arid, phytohydrological, phenological and other ecological and climatic maps.

The index is calculated using the formula:

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}, \quad (1)$$

where NIR is the pixel value in the near-infrared region; RED stands for the pixel value in the red region. The NDVI itself varies between -1.0 and +1.0.

2.2 Infrared Percentage Vegetation Index

Infrared Percentage Vegetation Index (IPVI) in contrast to NDVI does not require subtracting the red component from the numerator, which makes this index faster regarding calculations [7].

The index is calculated using the formula:

$$\text{IPVI} = \frac{\text{NIR}}{\text{NIR} + \text{RED}}, \quad (2)$$

where NIR is the pixel value in the near-infrared region; RED stands for the pixel value in the red region. The index varies between 0 and 1.

2.3 Atmospherically Resistant Vegetation Index

Atmospherically Resistant Vegetation Index (ARVI) was developed by Kaufman and Tanre [8]. This index is an improved NDVI, used to correct the influence of the atmosphere. It is most useful in regions with high atmospheric aerosol content, including tropical areas contaminated with soot.

The index is calculated using the formula:

$$ARVI = \frac{NIR - Rb}{NIR + Rb'} \quad (3)$$

where $Rb = RED - \alpha * (RED - BLUE)$, as a rule, $\alpha = 1$ (if there is small vegetation covering and unknown type of atmosphere $\alpha = 0.5$); NIR is the pixel value in the near-infrared region; RED stands for the pixel value in the red region; BLUE is the pixel value in the blue region. The index varies between -1 and 1.

2.4 Enhanced Vegetation Index

Enhanced Vegetation Index (EVI) is an optimized vegetation index NDVI, when assessing the state of plants, it has advantages, since the influence of soil and atmosphere in the values of this index is minimized [9]. The index allows to assess the state of plants, both in the conditions of dense and thinned vegetation covering.

The index is computed following this equation:

$$EVI = \frac{NIR - RED}{NIR + C1 * RED - C2 * BLUE + L} * (1 + L), \quad (4)$$

where BLUE stands for the pixel value in the blue region; RED is the pixel value in the red region; NIR is the pixel value in the near-infrared region; coefficients C1, C2 and L empirically defined as equal to 6.0, 7.5 and 1.0 respectively. The index varies between -1 and 1.

2.5 Soil-Adjusted Vegetation Index

Soil-Adjusted Vegetation Index (SAVI) is a vegetation index that tries to minimize the impact of soil brightness by using a soil brightness correction factor [10].

The index is calculated using the formula:

$$SAVI = \frac{NIR - RED}{NIR + RED + L} * (1 + L), \quad (5)$$

where NIR is the pixel value in the near-infrared region; RED stands for the pixel value in the red region; L is a canopy background adjustment factor. The index varies between -1 and 1.

3 Image classification

The next stage of monitoring, after creating the vegetation index, is the search for objects in the image - classification of the image. Currently, the most commonly used approach for topical processing is relative classification, based on widely used multi-spectral images and additionally collected data, which are necessary to establish a correspondence between groups of pixels with similar characteristic values and classes of

the Earth's surface. This data can be collected as a result of field studies, and more limited in comparison with classical field methods, since classes must be identified only for a small number of pixels [11].

There are two types of relative classification: supervised classification (with training) and unsupervised classification (without training).

The essence of the supervised classification is to assign each of the image pixels to a specific class of objects on the ground, which corresponds to a certain area in the characteristics space.

Supervised classification includes several stages. The first step is to determine which object classes will be allocated as a result of the entire procedure. These may include vegetation types, agricultural crops, forest species, hydrographic objects, and so on. At the second stage, typical pixels are selected for each of the object classes, i.e. a training sample is formed. The third stage is the calculation of parameters, the "spectral image" of each of the classes formed as a result of a set of reference pixels. The set of parameters depends on the algorithm that is supposed to be used for classification. The fourth stage of the classification procedure is to view the entire image and assign each pixel to a particular class. The result of this stage is an image (classification map), as well as a table that gives the coordinates of the pixel and the name of the class it belongs to.

Unsupervised classification is based on a fully automatic distribution of pixels into classes based on statistics of pixel brightness distribution. This type of classification is used if it is initially unknown which objects are present in the image, or if the number of objects is large. As a result, the machine itself gives the resulting classes.

Let us consider the most common classification methods used in researches.

3.1 Minimum distance method

This method is used when spectral characteristics of different classes are similar, and the ranges of their brightness overlap. In the classification the method of minimum brightness of pixels is used to consider a vector in the space of spectral characteristics. Spectral distance between the reference vectors and vectors of brightness of all image pixels is calculated, then pixels are distributed into classes, if the distance from this vector to the reference one is less than a predetermined value (which is set in advance), then this vector is referred to this class. If the distance is greater than the specified value, it is referred to another class, or it does not belong to any of the classes.

Minimum distance calculates the spectral distance between the pixel vector and the average vector for each signature.

Euclidean distance. Euclidean distance is a common distance function. It represents a geometric distance in a multidimensional space:

$$E = \sqrt{\sum_{i=1}^n |t_i - x_i|^2}, \quad (6)$$

where n is the number of ranges; i is a certain range; t is an unknown spectrum; x is a reference spectrum; E is Euclidean distance.

Manhattan distance. Manhattan distance is the distance which is the average of the differences in coordinates. In most cases, this measure of distance leads to the same results as for the usual Euclidean distance. However, for this measure, the impact of individual large differences is reduced (because they are not squared). Formula for calculating Manhattan distance is the following:

$$M = \sum_{i=1}^n |t_i - x_i|, \quad (7)$$

where n is the number of ranges; i is a certain range; t is an unknown spectrum; x is a reference spectrum; M is Manhattan distance.

The disadvantage of this method is that it does not take into account the distribution (dispersion) of the pixel brightness in the reference areas. This can lead to errors during classification.

3.2 Method of spectral angle

Classification by the method of spectral angle is used to compare the spectral characteristics of an image with the spectral characteristics of references. The algorithm determines the proximity between these two characteristics by calculating the spectral angle between them. To do this, they are represented as vectors in n -dimensional space, where n is the number of spectral channels.

Since the method of spectral angle uses only the direction of vectors, it is not sensitive to the absolute brightness of pixels, since it is the length of the vector that determines the measure of their brightness. All possible brightness levels are treated in the same way, since pixels with lower brightness are simply located closer to the origin of coordinates of the scatterplot. The color of pixels corresponding to their class in the n -dimensional characteristics space is determined by the direction of their radius vectors.

The following formula is used to calculate the spectral angle:

$$\alpha = \cos^{-1} \left(\frac{\vec{t} * \vec{x}}{\|\vec{t}\| * \|\vec{x}\|} \right), \quad (8)$$

where α is the spectral angle between vectors x and t ; t is an unknown spectrum; x is a reference spectrum.

The expression can also be represented as:

$$\alpha = \cos^{-1} \left(\frac{\sum_{i=1}^{nb} t_i * x_i}{\left(\sum_{i=1}^{nb} t_i^2 \right)^{\frac{1}{2}} * \left(\sum_{i=1}^{nb} x_i^2 \right)^{\frac{1}{2}}} \right), \quad (9)$$

where nb is the number of image spectral channels.

4 Assessment of classification accuracy

An important step of the classification is to assess the accuracy of the results obtained. This assessment is performed by comparing the image resulting from the classification with field measurement data and other data, such as data of relevant thematic maps. These materials are called reference data. This comparison is possible because each pixel in the resulting image has geographical coordinates, and it is possible to compare the type of surface that the pixel belongs to as a result of classification with the actual surface type known from other sources. The accuracy of classification is assessed by comparing the classification result with reference data, which are thematic maps, a set of points studied in the field, etc. Points are selected on the resulting classification, and the corresponding points on the reference data are considered. The comparison results are recorded into a table called the matrix of errors (table 1). It contains the number of right (located on the diagonal) and wrongly classified points [12].

The reliability of the obtained assessments of classification accuracy is achieved by selecting a sufficient number of points for each of the classes obtained during classification. In the best case, each point of the classification result is compared with the reference data.

If we add the diagonal elements (correctly recognized image points) and divide this number by the total number of points involved in the assessment, we get the overall classification accuracy. For each class, there are two values: the ratio of correctly recognized pixels either to the line sum (the number of points in this class) or to the column sum (the number of points in the reference data). A user error is a value that indicates the probability that a point marked as class 2 on the classification result is actually class 2 point. Kappa parameter is also calculated based on the matrix of errors. This parameter compares the number of pixels in each of the matrix cells with the possibility of distributing pixels as a random variable.

Table 1. Matrix of errors

Classes	Classes according to reference data		Number of reference pixels	
	Class 1	Class 1		
Classes in classification results	Class 1	a	b	e
	Class 2	c	d	f
Total		a+c	b+d	e+f

Kappa parameter is defined as follows:

$$\kappa = \frac{N * \sum_{i=j=1}^m D_{ij} - \sum_{i=1}^m R_i * C_j}{N^2 - \sum_{i=1}^m R_i * C_j}, \quad (10)$$

where κ is Kappa parameter, N stands for the number of image pixels, m is the total number of classes, $\sum D_{ij}$ stands for the sum of diagonal elements of the error matrix (the sum of correctly classified pixels of the whole image), R_i is the total number of

pixels in i-line (pixel sum in i-line), C_j is the total number of pixels in j-column (pixel sum in j-column).

Kappa statistics can be calculated for each selected class. For a qualitative assessment of map matching based on Kappa statistics the following ratios are used: poor and very poor matching if $\kappa < 0.4$, satisfactory if $0.4 < \kappa < 0.55$, good if $0.55 < \kappa < 0.7$, very good if $0.7 < \kappa < 0.85$, and excellent if $\kappa > 0.85$.

5 Results

At the initial stage of the classification with training of the satellite image, it is necessary to identify all classes of the underlying surface that are present in this territory. The task of classification research was to identify deforestation.

The classification was performed using three methods: the minimum distance method, which uses Euclidean distance, the minimum distance method, which uses Manhattan distance, and the spectral angle method. NDVI, IPVI, ARVI, EVI, SAVI indices were used as vegetation indices for preprocessing of satellite images.

As a result of the classification of the image fragment, four types of underlying surface (classes) are defined: deforestations (red), coniferous forests (dark green), deciduous forests (light green), lakes (blue).

The result of the classification methods on the selected vegetation indices is shown in table 2.

After receiving the results, the classification accuracy was assessed. Accuracy was evaluated using the matrix of errors and Kappa statistics.

An image provided by experts was used as reference data. A matrix of classification errors was formed for deforestation class (Table 3). Coniferous forests, deciduous forests, and lake classes were combined into one class-background. Deforestations were defined into a separate class.

The following conclusions were made for a qualitative assessment of map matching based on the results of Kappa statistics:

- To detect deforestation, the minimum distance method (Euclidean distance), the minimum distance method (Manhattan distance), and the spectral angle method showed excellent classification results, using the following indices as the vegetation ones: NDVI, ARVI, EVI.
- For IPVI and SAVI indices, only two methods showed excellent results: the minimum distance method (Euclidean distance), and the minimum distance method (Manhattan distance).
- The spectral angle method performed poorly for IPVI vegetation index. And very good, but not excellent it performed for SAVI.

Table 2. Result of classification methods on various vegetation indices

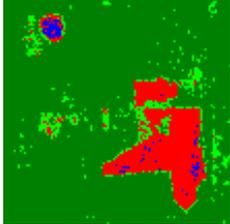
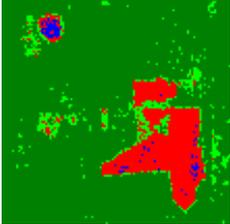
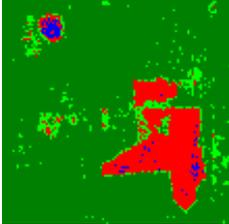
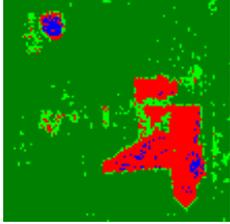
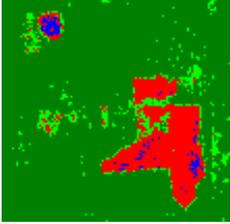
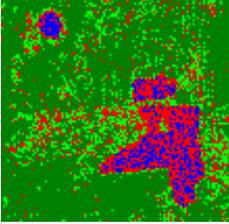
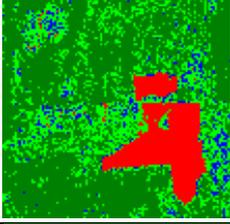
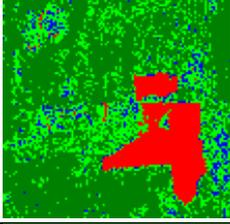
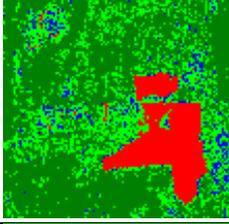
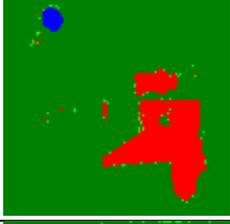
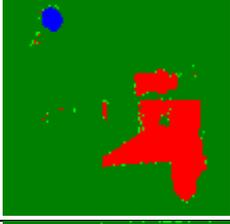
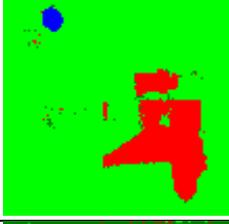
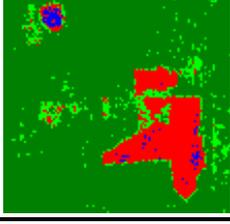
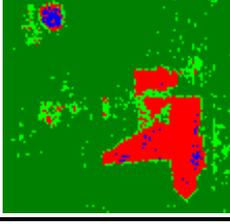
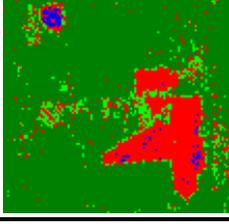
VI	Minimum distance method (Euclidean distance)	Minimum distance method (Manhattan distance)	Spectral angle method
NDVI			
IPVI			
ARVI			
EVI			
SAVI			

Table 3. Matrix of classification errors

Classes	Minimum distance method (Euclidean distance)		Minimum distance method (Manhattan distance)		Spectral angle method		Number of reference pixels
	Reference data classes		Reference data classes		Reference data classes		
	Background	Deforestation	Background	Deforestation	Background	Deforestation	
NDVI							
Background	10486	111	10486	111	10481	116	10597
Deforestation	133	1370	138	1365	118	1385	1503
Σ	10619	1481	10624	1476	10599	1501	12100
IPVI							
Background	10485	112	10468	129	9458	1139	10597
Deforestation	196	1307	161	1342	640	863	1503
Σ	10681	1419	10629	1471	10098	2002	12100
ARVI							
Background	10526	71	10521	76	10522	75	10597
Deforestation	104	1399	94	1409	96	1407	1503
Σ	10630	1470	10615	1485	10618	1482	12100
EVI							
Background	10547	50	10547	50	10545	52	10597
Deforestation	78	1425	78	1425	78	1425	1503
Σ	10625	1475	10625	1475	10623	1477	12100

By the results of Table 4 Kappa statistics was calculated. Table 3 gives the calculation results.

Table 4. Kappa statistics

Index name	Kappa statistics		
	Minimum distance method (Euclidean distance)	Minimum distance method (Manhattan distance)	Spectral angle method
NDVI	0.9	0.9	0.91
IPVI	0.88	0.88	0.4
ARVI	0.93	0.94	0.93
EVI	0.95	0.95	0.95
SAVI	0.9	0.89	0.75

6 Conclusion

The paper analyzes monitoring of forest pathologies. The necessity to automate some stages of the forest monitoring algorithm was identified. Empirical research was conducted for using vegetation indices and methods of classification of forests on space images.

The research reveals the relationship between the choice of vegetation index and the classification method. Depending on the area under study, it is offered to use the necessary index (for example, in areas with tropical climate, it is better to use an index that takes into account high air humidity (ARVI), etc.) and the proposed appropriate classification method to improve the effectiveness of the results.

References

1. Forest code of the Russian Federation as amended on December 27, 2018 (part 4, article 60.5).
2. Earth Observing System. Sentinel-2 Homepage, <https://eos.com/sentinel-2/c>, last accessed 2020/05/15.
3. Trubakov, E., Trubakov, A., Korostelyov, D., Titarev, D. Selection of Satellite Image Series for the Determination of Forest Pathology Dynamics Taking Into Account Cloud Coverage and Image Distortions Based on the Data Obtained from the Key Point Detector. Proceedings of the 29th International Conference on Computer Graphics and Vision, Moscow, pp. 159-163 (2019). DOI: 10.30987/graphicon-2019-2-159-163
4. The order of April 5, 2017 N 156 «On approval of the state forest pathology monitoring procedure».
5. Showengerdt, R. Remote sensing. Models and methods of image processing. M., 2010. 560 p.
6. Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J.-M., Tucker, C. J., Stenseth, N. C. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in Ecology and Evolution. 2005. Vol. 20. P. 503–510. DOI: 10.1016/j.tree.2005.05.011
7. Crippen, R. E., Calculating the Vegetation Index Faster. Remote Sensing of Environment. vol 34. pp. 71-73 (1990).
8. Kaufman, Y. J., Tanre D. Atmospherically resistant vegetation index (ARVI). Proc. IEEE Int. Geosci. and Remote Sensing Symp, IEEE, New York, pp. 261-270 (1992).

12 E.Trubakov, O. Trubakova

9. Skakun, R.S., Wulder, M.A., Franklin, S.E. Sensitivity of the thematic mapper enhanced wetness difference index to detect mountain pine beetle red-attack damage. *Remote Sensing of Environment*. vol. 86. pp. 433-443 (2003).
10. Mozgovoy, D.K., Kravets, O.V. Using multispectral images for classification of agricultural crops. *Ekologiya I Noosfera* (1-2), - 54-58 (2019).
11. Oreshkina, LV, Shidlovsky, Comparison, AV, Kovalenok, V.G. Comparison of classification methods for multi-zone satellite images. *Proceedings of the Second Belorussia Space Congress. 25-27 October, Minsk, Belarus. OIPI NAS of Belarus. 205-208 s* (2015).
12. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sensing of Environment* (80), pp. 185-201 (2002).