NDVI TIME SERIES MODELING IN THE PROBLEM OF CROP IDENTIFICATION BY SATELLITE IMAGES

N.S. Vorobiova, A.V. Chernov

Samara National Research University, Samara, Russia
Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, Samara, Russia

Abstract. The paper deals with the problem of NDVI time series modeling and application of simulated data in task of crop identification by satellite images. The simulation is performed for six types (classes) of crops in each agricultural zone, situated in the territory of the Samara region. Simulation parameters for each class are calculated from the coefficients of approximation which are obtained by approximating the time series of real agricultural fields by the function of a certain kind. The generated sets of simulated time series are used for crop recognition on real fields, located on the territory of the Samara region.

Keywords: time series, vegetation index, NDVI, satellite images, crops identification, crops recognition, algorithm for calculating estimates, time series approximation, time series modeling.

Citation: Vorobiova NS, Chernov AV. NDVI time series modeling in the problem of crop identification by satellite images. CEUR Workshop Proceedings, 2016; 1638: 428-436. DOI: 10.18287/1613-0073-2016-1638-428-436

Introduction

Nowadays crop identification using satellite images is an important task of remote sensing data application in agriculture [1]. A variety of methods for solving this task is big [2,3]. Let us consider the methods which use time series of vegetation indexes constructed by a set of satellite images [4]. Such methods most often use images of low spatial resolution but high temporal resolution (shooting at least once a day). This allows applying them for operational monitoring of a large area. Such methods are used in regional geoinformation systems of agroindustrial complex (hereinafter – GIS AIC) [5] for the control of agricultural land.

The quality of any recognition method depends on the probabilities of correct classification [6]. Obviously, specified quality criteria for the method of crop identification using time series will directly depend on the following factors: the date of classification (the longer the time series, the probability is higher), training sample size and the absence of errors in training sample.

It is a typical situation for regional GIS AIC when farmers declare information about crops sowed on the fields and some of declared data contain significant errors and false reporting; sometimes information may come too late – close to the end or after the end of the growing season. Therefore, such declared data cannot be used as a sample for evaluating the quality of classification, but only for general estimation of distribution in feature space. In this paper we propose a method for time series modeling and investigate the possibility of using simulated time series as a training sample for crop identification.

Boundaries of real fields and information about crops seeded on them for the years 2011-2015 are used in current work. All real fields are located in three agricultural zones (hereinafter - the zone) of Samara region: the northern, central and southern. Time series modeling and generating, crop recognition procedures are carried out for each agricultural zone which are homogeneous by climatic conditions.

In this paper we consider a partition of crops for the following classes: winter crops, early spring crops, late spring crops, fallows, perennial grasses, unused lands. The most popular vegetation index NDVI [7] is used for time series calculation.

1 Time series modeling

Represent time series as a mixture of useful (ideal) signal and additive noise having gaps in the daily observations due to cloud cover. In order to use the time series as a training set, we simulate only the useful signal, as noise and omissions lead to a deterioration of the classification quality. The question of choosing the shape of useful signal is raised.

Commonly used methods of time series reconstruction offer to approximate the time series by the functions like this: asymmetric Gaussian [8,9], the double logistic [10]. These functions and polynomials of different shape were compared by approximation quality, flexibility and sustainability. A detailed description of the comparison results is beyond the scope of this paper. In the end, the following function is encouraged to use for modeling a useful component of a time series (let's call it the ideal curve):

$$F(x) = (x-a)(x-b)P_n(x) + kx + e \tag{1}$$

where $P_n(x)$ – linear combination of Legendre polynomials up to the n degree, a – the earliest date when at least one object of the class in this zone has observation, b – the latest date when at least one object of the class in this zone still has observation. The coefficients k, e are obtained from the conditions:

$$\begin{aligned} k \cdot a + e &= y_a \\ k \cdot b + e &= y_b \end{aligned} \tag{2}$$

where y_a, y_b – the values of averaged class profile approximated by polynomial $P_n(x)$ in the points a, b respectively. Averaged class profile is calculated by using all time series for a class and consists of averaged values for each day of observation.

Thus, parameters k, e for a certain class are common for all curves of this class, but the coefficients of polynomial $P_n(x)$ are different. In other words, each ideal curve from some class is characterized by vector of coefficients $\overline{p} = (p_1, p_2, ..., p_{n+1})^T$. Next, we consider $n = 4$.
Next stage is calculation of modeling parameters for ideal curves.

2 Calculation of modeling parameters

Time series corresponded to the real fields and calculated by the data from Terra/MODIS satellite processed up to the level MOD09GQ will be used to calculate the modeling parameters. Total number of real time series is 20940. There is a correspondence "time series" - "class of crops". The following table shows the total number of real time series for each class.

Table 1. Total number of real time series for each class

Class of crops	Total number of real time series
perennial grasses	887
unused lands	2416
winter crops	2972
fallows	4650
late spring crops	5969
early spring crops	4026

Time series profiles vary considerably depending on the year, zone and class of crop, so a procedure of modeling parameters estimation is performed separately for each class of crop in a selected zone in a given year. For fixed triple "class of crop – zone – year" calculation of modeling parameters consists of the following steps:
1. Time series approximation by function of the form (1) according to the method of least squares. The result is a number of implementations of a coefficient vector $\overline{p}$ having multivariate normal distribution. Number of implementations is equal to the number of real time series for fixed triple "class of crop – zone – year".
2. Calculation of the vector of mathematical expectations $\overline{M}$ and the covariance matrix B, that characterize the law of vector $\overline{p}$ distribution.
90 different sets of parameters to simulate ideal curves defined by the formula (1) were calculated for the years 2011-2015 for Samara Region.
It should be noted that the residuals of approximation obtained in the first stage, are the values of noise subtracted from the useful ideal signal. Uncorrelated noise is obtained by time series approximation of the function (1) for each triple "class of crop-zone-year". Such conclusion can be drawn by analyzing the value of auto correlation function (hereinafter – ACF), built by a sequence of residuals for selected triple "class of crop – zone – year". All ACF values except corresponding to a zero lag do not exceed the 0.5 value, so it is possible to speak of uncorrelated ACF values. This

means that a sufficient degree of the polynomial $P_n(x)$ in the function (1) is selected, and the residuals do not contain useful signal remains unaccounted of the function (1). The average value of the mean square error of time series approximation for the triple "class of crop-zone-year" amounts to 0,047.

The figure below shows an example of time series approximation by function (1). The X-axis represents the time coordinate – the date of the time series observation. Time coordinate has been translated into the range [-1, 1] for the convenience of calculation. The Y-axis represents the values of NDVI index.

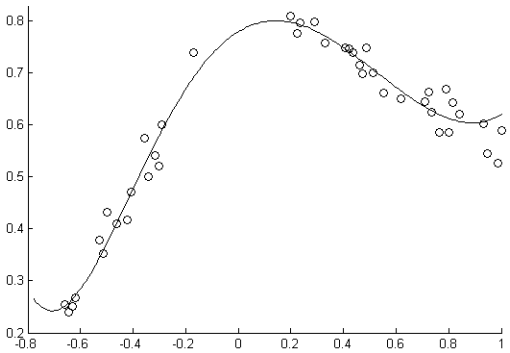

Fig. 1. Example of time series approximation

3 Ideal curves generating

Ideal curve is determined by the vector of coefficients $\overline{p}$, so to obtain ideal curve it is necessary to generate a vector $\overline{p}$ having multivariate normal distribution with given vector of mathematical expectations $\overline{M}$ and covariance matrix B as follows[11]:

1. Generate a vector $\overline{\xi} = (\xi_1, ..., \xi_{n+1})^T$ whose components are independent random variables having standard normal distribution.
2. Calculate the matrix A, which is a Cholesky decomposition of the covariance matrix B.
3. Calculate the vector of coefficients $\overline{p}$ through a linear transformation of the vector $\overline{\xi}: \overline{p} = A\overline{\xi} + \overline{M}$.

Next, daily values of ideal curve are calculated in an acceptable range $[a,b]$ by generated coefficient vector $\overline{p}$ and a set of parameters k,e that are common to the triple "class of crop-zone-year".

According to the specified algorithm ideal curves for six types of crops were generated for the three zones of the Samara region for years 2011-2015. For each triple "class of crop-zone-year" 4000 curves were generated.

4 Crop recognition algorithm

To detect crop types by using time series a method based on the algorithm for calculating estimates (abbreviated – ACE) is used [12]. The advantage of ACE is the ability to classify objects with gaps in the features. Gaps in the values of time series and, as a consequence, gaps in a set of features arise because of cloudiness. Let's concretize the ACE model to recognize crop types:
1. Features - the values of time series.
2. The system of reference feature sets consists of a single set including all features.
3. Proximity function for recognizable object a and reference object ω is calculated as follows:

$$\rho = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(\omega_n - a_n)^2}\ , \tag{3}$$

where $a_n, n = \overline{0, N-1}$ – features of object a (set of time series values); $\omega_n, n = \overline{0, N-1}$ – features of object ω. Proximity function is calculated only for the days n on which both the object ω and the object a have the values of time series.
4. The value of proximity function $f(\omega, a)$ for the reference object ω and the recognizable object a is calculated so:

$$f(\omega, a) = \begin{cases} 1, & \rho \le T \\ 0, & \rho > T \end{cases}, \text{ where } T \text{ is threshold of proximity.}$$

5. The estimation of proximity of the object a to a certain class Ω_j is calculated as follows:

$$\Gamma(\Omega_j) = \sum_{\omega \in \Omega_j} f(\omega, a).$$

6. Classification of recognizable object a will be done in class Ω_c according to the decision rule:

$$c = \arg \max_{m=0, M-1} \Gamma(\Omega_m)$$

where M – the number of classes.

5 Experiments

Results of experiments are given below. The experiments were carried out to analyze the practicability of using the simulated time series instead of time series built for real fields as training sample in the recognition methods. First three experiments used simulated time series as a reference sample in the algorithm for calculating estimates

and were carried out to identify the crops on real fields located in the Samara region in 2014 in the northern, central and southern zones, respectively. In the fourth experiment time series built for real fields were used as a training sample, and crop identification was carried out without separating the fields by zones. This is due to a small amount of real data, which is not enough in the case of the classification by zones to build both training sample, and a set of fields for identification. The probability matrixes of classification from class to class for each experiment are given below in Tables 2-5, respectively. The classification method for all experiments is ACE.

1. The northern zone. Total number of recognized objects is 1341. The overall probability of correct classification amounts to 0.58.

Table 2. Probability matrix of classification from class to class. Year 2014. The northern zone

Class	perennial grasses	unused lands	winter crops	fallows	early spring crops	late spring crops
perennial grasses	0.33	0.51	0.05	0.01	0.05	0.04
unused lands	0.13	0.73	0.02	0.07	0.05	0.01
winter crops	0.03	0.10	0.74	0.08	0.04	0.01
fallows	0.07	0.30	0.04	0.44	0.09	0.05
early spring crops	0.02	0.13	0.04	0.07	0.56	0.16
late spring crops	0.03	0.11	0.02	0.04	0.14	0.66

2. The central zone. Total number of recognized objects is 2324. The overall probability of correct classification amounts to 0.73.

Table 3. Probability matrix of classification from class to class. Year 2014. The central zone

Class	perennial grasses	unused lands	winter crops	fallows	early spring crops	late spring crops
perennial grasses	0.25	0.53	0.11	0.08	0.01	0.02
unused lands	0.03	0.88	0.02	0.04	0.01	0.02
winter crops	0.03	0.05	0.87	0.03	0.01	0.01
fallows	0.03	0.25	0.05	0.55	0.06	0.05
early spring crops	0.01	0.14	0.01	0.06	0.68	0.11
late spring crops	0.00	0.12	0.01	0.03	0.07	0.76

3. The southern zone. Total number of recognized objects is 2766. The overall probability of correct classification amounts to 0.70.

4. Classification of real data on real data. Evaluation of the classification quality was carried out by cross-validation. Sample of all real time series of year 2014 was divided five times into training sample (reference objects) and control sample (recognizable objects) in a ratio of 2:1. Total number of objects in sample is 6432, the number of

reference objects is 4288, and the number of recognizable objects is 2144. Average value of the overall probability of correct classification amounts to 0.69.

Table 4. Probability matrix of classification from class to class. Year 2014. The southern zone

Class	perennial grasses	unused lands	winter crops	fallows	early spring crops	late spring crops
perennial grasses	0.44	0.39	0.06	0.03	0.06	0.03
unused lands	0.12	0.83	0.01	0.02	0.00	0.01
winter crops	0.06	0.02	0.87	0.02	0.01	0.02
fallows	0.07	0.14	0.03	0.56	0.09	0.10
early spring crops	0.03	0.05	0.01	0.08	0.68	0.14
late spring crops	0.01	0.03	0.02	0.09	0.14	0.70

Table 5. Probability matrix of classification from class to class. Year 2014

Class	perennial grasses	unused lands	winter crops	fallows	early spring crops	late spring crops
perennial grasses	0.40	0.42	0.08	0.03	0.01	0.03
unused lands	0.17	0.77	0.01	0.02	0.02	0.01
winter crops	0.07	0.03	0.86	0.01	0.01	0.02
fallows	0.12	0.19	0.03	0.53	0.06	0.07
early spring crops	0.04	0.09	0.01	0.04	0.68	0.14
late spring crops	0.03	0.07	0.01	0.07	0.07	0.75

Thus, the results of the above experiments lead to the conclusion that it is possible to use the simulated time series as a learning sample in the methods of crops identification.

Conclusion

The idea of using the simulated time series as a training sample in the method of crop type identification was demonstrated in this paper. This idea has the following advantages:

1. Generating of training sample of any size by calculated modeling parameters.
2. Ability to use a set of simulated time series as a basis to assess the classification quality of any classification algorithm.
3. Ability to improve the classification quality by choosing different approximation functions.
4. The possibility of crop identification at the beginning of the season, when the training sample size is not sufficient for classification, but there is a set of fields with reli-

able information about the seeded crops (let's call them the support fields). This option will appear if there is a set of time series for a number of past years. In other words, if there is an accumulated historical statistics of the crops development on the real fields. Modeling parameters are calculated by this historical time series, and set of modeling time series will be generated by obtained modeling parameters. These modeling time series, built according to past years, will be called templates of crop development over the past years. Thus at the beginning of the season the closest template to development of crops in support fields will be selected from a set of historical patterns. Further, modeling time series of selected historical template can be used as a training sample in the method of crop identification in the current year.

Acknowledgements

This work was financially supported by RFBR, project № 16-37-00043_mol_a «Development of methods of using data from geoinformation systems in remote sensing data processing» and project № 16-29-09494_ofi_m «Methods of computer processing of multispectral remote sensing data for vegetation areas detection in special forensics».

References

1. Vorobiova NS, Timbay EI. Geoinformation system of agricultural lands inventory and control development. Computer Optics, 2009; 33(3): 340-344. [in Russian]
2. Denisova AYu, Sergeyev VV. Impulse response identification for remote sensing images using GIS data. Computer Optics, 2015; 39(4): 557-563 [in Russian]. DOI: 10.18287/0134-2452-2015-39-4-557-563.
3. Sergeyev VV, Denisova AYu. Iterational method for piecewise constant images restoration with an a priory knowledges of image objects boundaries. Computer Optics, 2013; 37(2): 239-243. [in Russian]
4. Bartalev SA, Egorov VA, Loupian EA, Plotnikov DE, Uvarov IA. Recognition of arable lands using multi-annual satellite data from spectroradiometer MODIS and locally adaptive supervised classification. Computer Optics, 2011; 35(1): 103-116. [in Russian]
5. Vorobiova NS, Denisova AYu, Kuznetsov AV, Belov AM, Chernov AV, Myasnikov VV. How to Use Geoinformation Technologies and Space Monitoring for Controlling the Agricultural Sector in Samara Region. Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications, 2015; 25(2): 347-353.
6. Kuznetsov AV, Myasnikov VV. A comparison of algorithms for supervised classification using hyperspectral data. Computer Optics, 2014; 38(3): 494-502. [in Russian]
7. Maiorova VI, Bannikov AM, Grishko DA, Jarenov IS, Leonov VV, Toporkov AG, Harlan AA. Monitoring condition of agricultural fields based on prediction of NDVI with the use of multi-spectral and hyper-spectral data from space imagery. Science & education: scientific edition of Bauman MSTU, 2013; 07: 199-228. [in Russian]
8. Ozdogan M. The spatial distribution of crop types from MODIS data: Temporal unmixing using Independent Component Analysis. Remote Sensing of Environment, 2010; 114(6): 1190-1204.

9. Fischer A. A Model for the Seasonal Variations of Vegetation Indices in Coarse Resolution Data and Its Inversion to Extract Crop Parameters. Remote Sensing of Environment, 1994; 48: 220-230.
10. Wei W, Wu W, Li Z, Yang P, Zhou Q. Selecting the optimal NDVI time-series reconstruction technique for crop phenology detection. Intelligent Automation & Soft Computing. Special Issue: Intelligent Automation with Applications to Agriculture, 2016; 2: 237-247.
11. Kolomiets EI, Myasnikov VV. Simulation of experimental data for pattern recognition tasks. Guidelines for the laboratory work No 1, 2010: 20 p. [in Russian]
12. Vorobiova NS. Crops identification by using satellite images and algorithm for calculating estimates. Proceedings of International conference and school for young scientists "Information technology and nanotechnology (ITNT-2015)", 2015: 83-88. [in Russian]