A comparison of the Bayesian classifier and the Estimates-based algorithm for crop identification by Terra/MODIS 250-m data

N S Vorobiova¹ and A V Chernov¹,²

¹Samara National Research University, Moskovskoye Shosse 34, Samara, Russia, 443086
²Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Molodogvardeyskaya str. 151, Samara, Russia, 443001

e-mail: natalyavorobiova@gmail.com

Abstract. We studied and compared the Bayesian classifier (BC) and the Estimates-based algorithm (EBA) in this research for crop identification by a set of Terra/MODIS 250-m data. Methods are based on time-series calculated by satellite images. EBA has an undoubted advantage. It uses time series even with missing values whereas BC requires reconstructing missing values of time series a priori. Experiments are based on data obtained for the Samara region in 2014-2016 years. The numerical results are given for the different length of time series to compare two methods throughout the growing season.

1. Introduction
Crop identification by remote sensing data is a crucial task of agricultural monitoring [1, 2]. Identification results have such fields of application: lands inventory, searching for unused lands, control of farmlands and subsidies allocated to agricultural producers [3]. Early identification conducted as close as possible to the start of growing season is the most important and allows inspecting the usage of agricultural lands in-situ [4]. In this paper, we consider early regional-scale (size of Russia regions) identification. Most large-areas crop identification is performed by time series of satellite images [5, 6]. Time series allow observing crop growth dynamics during the sowing season or part of it.

Information technology for early regional-scale crop identification has been proposed in [4]. It is based on time series calculated by a seasonal sequence of satellite images and has an important advantage. The technology provides crop identification when the data of the current year is not enough to set up recognition methods [2, 5, 6]. It uses previous-year data and crop knowledge of the current year on a small number of so-called reference fields. Information technology has two stages: setting by
previous-year data and current-year identification. Each stage includes a set of blocks. The technology uses the Estimates-based algorithm (from now on - EBA) for crop identification.

This work aims to compare EBA used in [4] for crop identification with some other recognition method and to evaluate its feasibility for early crop identification. The Bayesian classifier for normally distributed feature vectors is chosen for comparison.

2. Data
We used the data obtained for the Samara region in 2014-2016: satellite images and ground data (field boundaries and data on planted crops). Crop type of each field was from the following list: perennial grasses and unused land (PG and UL); winter crops; fallow; early spring crops (early SC); late spring crops (late SC).

According to natural and economic conditions, climate patterns, and soil characteristics the territory of the Samara region consists of three agroclimatic zones (further - zones): northern, central and southern [7]. A total number of fields in ground data are summarized in table 1 by crop type, year and zone.

| Year | Zone     | PG and UL | Winter Crops | Fallow | Early SC | Late SC |
|------|----------|-----------|--------------|--------|----------|---------|
| 2014 | Northern | 289       | 210          | 291    | 334      | 217     |
|      | Central  | 530       | 357          | 408    | 472      | 549     |
|      | Southern | 164       | 526          | 598    | 507      | 971     |
| 2015 | Northern | 304       | 243          | 220    | 296      | 225     |
|      | Central  | 498       | 368          | 486    | 436      | 525     |
|      | Southern | 111       | 265          | 620    | 626      | 982     |
| 2016 | Northern | 252       | 220          | 214    | 235      | 207     |
|      | Central  | 461       | 302          | 330    | 362      | 397     |

The NDVI time series was calculated for each field by satellite images of the corresponding year. NDVI (Normalized Difference Vegetation Index) is the most popular index used for crop identification and research of vegetation by satellite images [2, 5, 8 – 12]. It is calculated using the equation:

\[ \text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{RED}}}{\rho_{\text{NIR}} + \rho_{\text{RED}}}, \]

where \( \rho_{\text{NIR}} \), \( \rho_{\text{RED}} \) is the surface reflectance of the near infrared and red bands, respectively.

Terra/MODIS 250 m MOD09GQ data were used for time series calculation. The MOD09GQ product is an estimate of the surface spectral reflectance corrected for atmospheric conditions such as gasses, aerosols, and scattering. It has a geo-referencing, a radiometric and atmospheric correction and represents a minimum level of combining the daily 250 m data [13]. Each satellite image was used with a cloud mask, which allowed excluding non-informative pixels from calculations. Therefore missing values appeared in time series on some days.

Training/etalon data and validation data were generated by a set of time series and used for crop identification by the Estimates-based algorithm and the Bayesian classifier and comparison of results. Crop identification was performed separately in each zone because the differences in agroclimatic conditions could affect the quality in certain parts of the region [10]. Each zone in every year had own set of training/etalon data and validation data consisted of time series for five crop classes.

2.1. Training/etalon data
Time series calculated by satellite images are usually noisy. Authors of information technology suggest replacing noisy time series intended for use as training/etalon data by some mathematical
models in the form of smooth curves [4]. Such curves describing the time series undistorted by noise has been called ideal. The cubic spline is the best fitting function for describing ideal curves [14, 15]. The method of ideal curves generating is given in [14, 15].

In this paper, we applied the method described in [14, 15] to generate time series of training/etalon data. For a given year, zone and class of crops, we used individual generation parameters (mean vector and covariance matrix of the multivariate normal distribution law) and generated 2000 ideal curves.

2.2. Validation data

Generated time series formed validation data. Generation was performed using a mathematical model of the noisy time series. The model was the ideal curve with superimposed noise and excluded points due to the cloudiness.

The generation of the noisy time series included two stages:

- noise contamination of ideal values \( \tilde{y}_i, i = 1, n \) by an additive noise: \( y_i = \tilde{y}_i + v_i \), where \( v_i \) is noncorrelated noise with normal distribution law, \( y_i \) is noised values of time series, \( \tilde{y}_i, i = 1, n \) are ideal curve values with one-day time-step;

- excluding some of the points from \( y_i, i = 1, n \) in a random manner. The average number of excluded points corresponded to the average number of cloudy days.

For a given year, zone and class of crops, we used individual generation parameters and generated 1000 noised curves.

The examples of ideal curves and corresponding noisy time series generated by the considered models are shown in Figure 1. The x-axis represents the date, the y-axis represents NDVI values. Dates have been recalculated into a range [-1; 1].

![Figure 1. The examples of ideal curves (gray curve) and corresponding noisy time series (black curve). Winter crops, Southern zone, 2015.](image)

In total, we generated nine sets of reference/etalon and validation data for each year and the zone from table 1. Further, we called them as nine experimental sets.

3. Methods

The values of time series were the features for recognition. The number of features matched the number of days in growing season. The overall probability of correct classification \( Q \) calculated by the following equation was the criterion of recognition quality:

\[
Q = \frac{m}{M},
\]

where \( m \) – the number of correctly recognized objects, \( M \) – the total number of objects for recognition. Five classes \( \Omega_l, l = 1, 5 \) were used for the analysis (see table 1).

3.1. Estimates-based algorithm

Zhuravlev Yu. I. proposed a class of recognition algorithms based on the estimates [16]. The principle of the EBA is to calculate some similarity estimates that characterize the similarity of the investigated
(recognized) and etalon objects. Each class has its own set of etalon objects. A single estimate that characterizes the similarity of the investigated object and some class is formed by similarity estimates of the investigated and etalon objects from this class. Further, the similarity estimates to each class are compared, and the classification is performed.

A class of recognition algorithms based on the estimates is not a predetermined algorithm, but a model of the recognizing algorithm. Each recognition task requires the specification of this model. A detailed description of the EBA for crop identification by using time series is considered in [4]. In this paper, we give a summary of the method.

The method has one parameter \( T \) which is the threshold value of similarity. The system of reference sets of features consists of a single set and includes all the features. The proximity function \( \rho(x) \) of the investigated \( x \) and etalon \( x^* \) objects is calculated by the equation:

\[
\rho(x) = \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i^* - x_i)^2 \right]^{1/2},
\]

where \( x_i, x_i^*, i=1, n \) – the features of objects \( x \) and \( x^* \), respectively. The proximity function is calculated only for the days in which both objects have the values of time series.

The value of the proximity function is calculated by the equation:

\[
f(\rho(x)) = \begin{cases} 
1, & \rho(x) \leq T \\ 
0, & \rho(x) > T. 
\end{cases}
\]

The estimate \( \Gamma(\Omega_i) \) of the proximity of an object \( x \) to a class \( \Omega_i \) is calculated as follows:

\[
\Gamma(\Omega_i) = \sum_{x \in \Omega_i} f(\rho(x)).
\]

The object is classified into the class \( \Omega_k \) according to the decision rule: \( k = \arg \max_{i=1, T} \Gamma(\Omega_i) \).

We investigated the values of the parameter \( T \) that provided the best recognition quality for different lengths of time series. For each of the nine experimental sets, we performed the recognition of the validation data by the etalon data with \( T \) from the range \([0.45; 0.1]\) in increments of 0.005. The end date of the time series changed with ten days step. For each end date of the time series, we chose \( T \), which provided the best recognition quality. We denoted this quality as \( Q_{\max} \). It was found that \( T \) slightly varied when the length of the time series changed.

The values of \( T \) chosen for different lengths of time series for the northern, central and southern zones in 2015 (a) and 2016 (b) and providing the best quality of classification \( Q_{\max} \) are shown in Figure 2. The x-axis represents a relative date calculated from March 1 and corresponding to the end of time series. The y-axis represents \( T \). The graphs in Figure 2 show a weak change of \( T \) in time.

![Figure 2](attachment:image.png)

**Figure 2.** The values of \( T \) for different lengths of time series for the northern, central and southern zones in 2015 (a) and 2016 (b).
Therefore, for a particular zone and year, we suggested using the only average $T$ value. We denoted this value as $T_{av}$. For each of the nine experimental sets, we performed the recognition for the different lengths of time series using the corresponding value of $T_{av}$ and conducted the sequence of quality values $Q(T_{av})$. The comparison of $Q_{max}$ and $Q(T_{av})$ showed that, on average, the deviation of $Q(T_{av})$ from $Q_{max}$ was not above 0.5%. The following figures show some experimental results.

The values of $Q_{max}$ and $Q(T_{av})$ for the central zone in 2015 (a) and 2016 (b) are given in Figure 3. $Q_{max}$ and $Q(T_{av})$ are close indeed for almost any end date of the time series.

![Figure 3](image)

**Figure 3.** The comparison of $Q_{max}$ and $Q(T_{av})$ for the central zone in 2015 (a), 2016 (b).

Figure 4 shows the values of $Q(T_{av})$ for different lengths of time series for the northern, central and southern zones in 2015 (a) and 2016 (b). $Q(T_{av})$ varies significantly depending on the year and zone.

![Figure 4](image)

**Figure 4.** The quality of recognition $Q(T_{av})$ for different lengths of time series for the northern, central and southern zones in 2015 (a) and 2016 (b).

3.2. Bayesian classifier

Recognition was conducted by using the Bayesian classifier for normally distributed feature vectors [17]. We assumed that the feature vectors $x = (x_1, \ldots, x_n)^T$ in each class $\Omega_l, l = 1, 5$ had a multivariate normal distribution law with a probability density function:

$$
 f(x / \Omega_l) = \frac{1}{(2\pi)^{n/2} \sqrt{|B_l|}} \exp\left(-\frac{1}{2} (x - M_l)^T B_l^{-1} (x - M_l)\right), l = 1, L,
$$

where $M_l, B_l$ – mean vector and covariance matrix of feature vectors from class $\Omega_l$, respectively. The discriminate functions of the feature vector had the following form:
\[ d_l(x) = \ln P(\Omega_l) - \ln \sqrt{\det B_l} - \frac{1}{2} (x - M_l)^T B_l^{-1} (x - M_l), l = 1, L, \]

where \( P(\Omega_l) \) is a prior probability of objects from the class \( \Omega_l \). The feature vector \( x \) was classified into the class that corresponded to the largest value of the discriminate function.

We analyzed the values of the features for objects from training/etalon and validation data and concluded that the features were linearly dependent. The principal component analysis reduced the dimension of the feature space. The number of components was chosen such that percentage of the total variance explained by selected components was not less than 99. The number of selected components varied for different lengths of time series, so the number of features used in the recognition method also changed. Figure 5 shows the number of features used for recognition depending on the end date of the time series in 2016. The number of features did not change from zone to zone in 2016.

For each of nine experimental sets and given end date of the time series we reduced the feature space by the same operator for both training/etalon and validation objects and got a new feature subspace of smaller dimension.

Figure 6 shows the histograms of the first (a) and the second (b) features in the new reduced feature space and their scatter plot (c) for objects from the class "Perennial grasses and unused land" located in the southern zone of 2015. In this example, the relative end date of the time series equalled 54. Given histograms and scatter plot demonstrate the proximity of the feature distribution to the normal distribution law.

![Figure 5](image5.png)

**Figure 5.** The number of features for the different lengths of time series in 2016.

![Figure 6](image6.png)

**Figure 6.** The histograms of the first (a) and the second (b) features in the new reduced feature space and their scatter plot (c). Class "Perennial grasses and unused land", the southern zone, 2015.

We conducted the recognition using the Bayesian classifier for normally distributed feature vectors on these new subspaces for each of nine experimental sets with ten days time-step. \( Q_{BC} \) was the quality corresponding to the Bayesian classifier.

### 3.3 The comparison of recognition methods

The experimental results showed that the quality of the Estimates-based algorithm \( Q(T_{av}) \) was close to the quality of the Bayesian classifier \( Q_{BC} \). The average deviation of \( Q(T_{av}) \) from \( Q_{BC} \) was 1.6%. In
most cases in the first half of the growing season, $Q(T_w)$ was higher than $Q_{bc}$, and therefore we concluded that the EBA was appropriate for early crops identification.

The values of $Q_{bc}$ and $Q(T_w)$ for different lengths of time series for the southern zone of 2015 (a), northern (b) and central (c) zones of 2016 are shown in Figure 7.

4. Conclusion

In this paper, we have considered the method based on the estimates and used in the information technology of early crop identification by satellite images. We have compared this method with the Bayesian classifier for normally distributed feature vectors. The study has demonstrated that both classifiers give a close recognition quality, but in the first half of the growing season, the EBA gives a better recognition quality (in most experiments). The additional advantage of the EBA is much better adaptation and lower requirements for the initial data in comparison with the Bayesian classifier. According to the study, we have concluded that the method based on the EBA is suitable for usage in the information technology for early crops identification.

Figure 7. The recognition quality of the Bayesian classifier $Q_{bc}$ and the Estimates-based algorithm $Q(T_w)$ for different lengths of time series for the southern zone of 2015 (a), northern (b) and central (c) zones of 2016.

5. References

[1] Abade N A, De C, Guimarães R F and De O 2015 Comparative analysis of MODIS time-series classification using support vector machines and methods based upon distance and similarity measures in the Brazilian Cerrado-Caatinga boundary Remote Sensing 7 12160-12191
[2] Masialeti I 2008 Assessment of time-series MODIS data for cropland mapping in the U.S. Central Great Plains PH.D. (University of Kansas)
[3] Schmedtmann J and Campagnolo M L 2015 Reliable crop identification with satellite imagery in the context of Common Agriculture Policy subsidy control Remote Sensing 7 9325-9346
[4] Vorobiova N S, Sergeyev V V and Chernov A V 2016 Information technology of early crop identification by using satellite images Computer Optics 40 929-938
[5] Victoria D D C, da P, Coutinho A C, Kastens J and Brown J C 2012 Cropland area estimates
using Modis NDVI time series in the state of Mato Grosso, Brazil *Pesquisa Agropecuaria Brasileira* **47** 1270-1278

[6] Plotnikov D E, Bartalev S A, Zharko V O, Mihailov V V and Prosyannikova O I 2008 An experimental assessment of crop types recognisability using time-series of intra-seasonal spectral reflectance measurements by satellite sensor *Sovremennye Problemy Distantzionnogo Zondirovaniya Zemli iz Kosmosa* **8** 199-208

[7] Vasin A V 2006 *Formation of highly productive polyspecific agrophytocenosis of forage crops in the Middle Volga the doctor of agricultural sciences* (Kinel)

[8] Terekhin E A 2015 NDVI seasonal dynamics of perennial grasses and its use for classification of their crops in the Belgorod Region *Sovremennye Problemy Distantzionnogo Zondirovaniya Zemli iz Kosmosa* **12** 9-17

[9] Kalinín N A, Pyankov S V, Sviyazov E M and Smirnova A A 2010 Technology of the complex estimation of crop biomass according to the remote Earth sounding *Bulletin of the Udmurt University. Series Biology. Earth sciences* **6** 3-10

[10] Wardlow B D, Egbert S L and Kastens J H 2007 Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains *Remote Sensing of Environment* **108** 290-310

[11] Wardlow B D and Egbert S L 2008 Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains *Remote Sensing of Environment* **112** 1096-1116

[12] Liu M W, Ozdogan M and Zhu X 2014 Crop type classification by simultaneous use of satellite images of different resolutions *IEEE Transactions on Geoscience and Remote Sensing* **52** 3637-3649

[13] Justice C O, Townshend J R G, Vermote E F, Masuoka E, Wolfe R., Saleous N, Roy D P and Morisette J T 2002 An overview of MODIS Land data processing and product status *Remote Sensing of Environment* **83** 3-15

[14] Vorobióva N and Chernov A 2017 Curve fitting of MODIS NDVI time series in the task of early crops identification by satellite images *The proceedings IV International Conference on Information Technology and Nanotechnology (ITNT-2018)* (Russia, Samara) 390-399

[15] Vorobióva N and Chernov A 2017 Curve fitting of MODIS NDVI time series in the task of early crops identification by satellite images *Procedia Engineering* **201** 184-195

[16] Zhuravlev J I, Kamílov M M and Tulyaganov S E 1974 *Estimate-calculating algorithms and their application* (Tashkent: FAN Publisher)

[17] Duda R O and Hart P E 1976 *Image recognition and scene analysis* (Moscow: MIR Publisher)

**Acknowledgments**

This work was supported by the Federal Agency for Scientific Organizations (Agreement 007-G3/43363/26) and by RFBR according to the research projects № 16-29-09494_ofi_m «Methods of computer processing of multispectral remote sensing data for vegetation areas detection in special forensics» and № 18-07-00748 A «Data fusion and mutual calibration methods of images produced by miscellaneous remote sensing sensors for the purposes of land use mapping».