Soil characteristics influence on the winter wheat yield regression modeling using NDVI vegetation index

A Y Denisova[1] and A V Evstiforova[1]

[1]Samara National Research University, Moskovskoye shosse 34, Samara, Russia, 443086

Abstract. The general way of a yield estimation using remote sensing data is a regression modeling based on cumulative NDVI vegetation index. Usually, the other additional factors are also included in the model to enhance a yield estimation accuracy, for example, weather conditions during the vegetation season or particular crop type information. Nevertheless, such factors as soil characteristics were not properly investigated and it is still an open question whether they should be included in modeling process or not. This article reveals the regression modeling accuracy dependency on the additional soil characteristics for winter wheat yield estimation using cumulative NDVI. We analyzed several groups of factors obtained from different data sources such as remote sensing images, geoinformation systems and meteorological data. These factors were used to construct primarily yield models without soil factors. The best two models were chosen as the basic models and then they were modified to take into account the additional soil characteristics. Finally, the comparison of the basic models with the modified models including soil parameters was made. We investigated such soil parameters as soil fertility score derived from a ten-year-old soil map and the results of the current agrochemical inspections. The comparison of basic and modified models was made for the winter wheat fields located in Samara region, Russia. The experiments showed that regression models give better determination coefficient and root mean square error if the additional soil characteristics are used for modeling. Moreover, the older data obtained from soil maps and reference book of the soil fertility score delivers better prediction than the current agrochemical inspection data. Therefore, there is no need to know lots of soil parameters for successfully winter wheat yield modeling. The results of this research can be applied to forecasting winter wheat yield in Samara region.

1. Introduction

Traditional yield prediction agricultural methods require a vast amount of ground truth measurements that makes these methods very expansive in use. To provide automated yield prediction without ground truth measurements the geoinformation systems (GIS) and remote sensing data (RS data) can be used. RS data provide objective information about agricultural crop growth whereas GIS contains effective instruments for additional parameters calculation based on graphic and semantic information about the field. Therefore, the development of simultaneous GIS and RS data for agricultural purposes remains a very promising research area.

Existing GIS and RS data yield prediction approaches are highly connected with the traditional methods used in agriculture that mainly utilize linear regression modeling and simulation modeling to find the relation between the observed factor values and the yield [1]. Traditionally only one critical factor with the greatest impact on the crop is used in regression modeling. It can be air temperature, moisture availability or sun radiance. This critical factor is generally determined by the experts as the

parameter which is probably the main reason for the crop stress in the current situation. Traditional linear regression models usually give yield forecast at the beginning of the vegetation season and define the potential crop yield. Simulation modeling is more precise than regression modeling, but it is more computationally intensive and requires lots of very specific parameters to be known. The main idea behind this approach is to simulate the crop phenology by means of several equations linking bioclimatic parameters and changes in vegetation biomass. The result of simulation modeling is an assessment of dry vegetation biomass for each date of the vegetation season. The final yield is determined as a cumulative biomass estimated using simulating modeling. The RS and GIS data-based yield models works in a similar way to traditional techniques, but the main variable reflecting biomass changes is estimated by means of RS data. Usually, the vegetation indexes (NDVI, LAI and others) are used to substitute the biomass variables in traditional simulating and regression models [2], wherein the regression models use cumulative vegetation index values. In both traditional and RS and GIS data-based cases, the regression modeling has more simple mathematical form and calibration than simulating modeling. Moreover, simulation models based on RS and GIS data (EPIC [3], WOFOST [4], SAFY [2] and etc.) have a huge number of parameters that still require ground truth measurements for each particular field which cannot be obtained from the RS or GIS data. Thus, the regression models using cumulative vegetation indexes are more preferable for yield prediction than simulation models.

In our research, we apply cumulative NDVI [5] vegetation index in winter wheat yield linear regression modeling. We selected a winter wheat as the most widespread agricultural crop in Samara region. Our research aim was to study the model behavior when the additional factors reflecting soil fertility are included in the model. The motivation of our research came from the fact that the soil fertility score and agrochemical survey data are very helpful in forecasting crop yields in agriculture. However, this fact is not obvious in regression modeling based on NDVI vegetation index. To obtain the basic model without soil factors, we analyzed three groups of factors extracted from different data sources and two approaches of the cumulative NDVI estimation. Several regression models were obtained and the best models with the optimal factor set were used to investigate the influence of soil factors. As a result, the recommendations on soil factors and cumulative NDVI estimation were given.

The article is structured as follows. Section 2 describes the factor groups used to construct the basic soil independent models and the general model equation. Section 3 reveals the details of two approaches to the cumulative NDVI estimation and the process of optimal factor set selection. Section 4 contains the results of the experimental evaluation of soil fertility impact on yield regression modeling.

2. Problem statement

Generally, regression modeling is used to find the relation between the expectation of the dependent random variable y and the set of independent variables $x_1, x_2, ..., x_n$ called factors. The type of regression function is determined according to the heuristic assumptions about the relationship between the observed independent parameters and dependent variable. In the present article, we regard a linear regression model in which the dependent variable y corresponds to the winter wheat yield:

$$y = \sum_{i=1}^{n} a_i x_i + a_0 , \quad (1)$$

where $a_i, i = 0,...,n$ are the coefficients of the model.

We consider several sets of factors for the model (1). The first set (basic set of factors (BSF)) contains factors extracted from the farmer data in GIS, meteorological services and RS data. BSF includes such subgroups of factors as:

1) the remote sensing factors. These factors are calculated using NDVI time series obtained by RS data. They include cumulative and maximum NDVI values for the period of the most active vegetation as well as the vegetation period duration and the vegetation start date. The following equation explains the process of factor extraction:

$$\xi_r = \sum_{t=r}^{T+r} \xi(t), \quad (2)$$

where ξ_r is the cumulative NDVI value, $\xi(t)$ is the average NDVI value for the particular field at the date t, r is the relative date of the vegetation period beginning, T is the vegetation period duration;

2) the bioclimatic factors. These factors are obtained from meteorological services. This subgroup contains such factors as maximum and minimum temperature for the vegetation period, maximum and minimum humidity during the vegetation period and climate zone in which the field is located;

3) the data of agricultural producers join such factors as crop cultivar, seeds reproduction, farmer identifier, date of seeding and geographical coordinates of the field object centroid. The factors of these group are obtained from the GIS database.

The second set of factors (SFS) contains soil fertility factors determined by the GIS soil maps. These factors are the soil fertility score and the humus content. The soil fertility score was determined according to the reference study [6] and the soil type given from the soil map in GIS.

The third factor set (TSF) is constructed as a combination of the soil fertility parameters according to the agrochemical survey data. TSF incorporates such factors as average humus, phosphorus, sulfur and potassium contents in the field soil. The agrochemical survey data were obtained from the semantic information of the agrochemical surveys map in GIS.

We conducted our research in several stages. First of all, we selected an optimal subset of the basic factor set (OBSF). OBSF was determined as the subset of significant factors of the model (1) which delivers the highest value of the determination coefficient (R^2). We further refer the regression model estimated for the OBSF as Model 1. In other words, the Model 1 is a name of the general soil-independent regression model estimated using OBSF. Then, in order to get the soil-dependent models, we combined OBSF with two other factor sets SSF and TSF and estimated the regression models for the resulting factor sets. The regression models obtained for OBSF and SSF are further referred as Model 2, while the regression models estimated using OFBSS and TSF are referred as Model 3. Model 2 and Model 3 are the names of the received soil dependent regression models. Finally, the comparison of soil-dependent and soil-independent models was made.

We applied the least square method [7] to estimate the regression coefficients. The quality of the particular regression model was determined by means of R^2 value [7] and root mean square error ε (RMSE):

$$\varepsilon = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(y_j - \hat{y}_j\right)^2}, \quad (3)$$

where $\hat{y}_j$ is a predicted yield value for the field j, y_j is an actual yield value for the field j and M is the number of fields.

3. Optimal factor set selection

The remote sensing subgroup of the basic factor set can be determined in different ways. We propose two approaches to estimate remote sensing factors that differ in estimating the vegetation beginning date r and the vegetation period duration T. Therefore, they lead to different cumulative NDVI values.

The first approach requires the NDVI threshold ξ_T to be given. This threshold is used to achieve the minimum $\tau_{\min}$ and maximum $\tau_{\max}$ relative dates for which the NDVI value is greater than the threshold ξ_T. Therefore, the vegetation beginning date r and the vegetation period duration T are defined as follows:

$$r = \tau_{\min}, \quad T = \tau_{\max} - \tau_{\min}. \quad (4)$$

Figure 1 illustrates the first approach to estimating remote sensing factors.

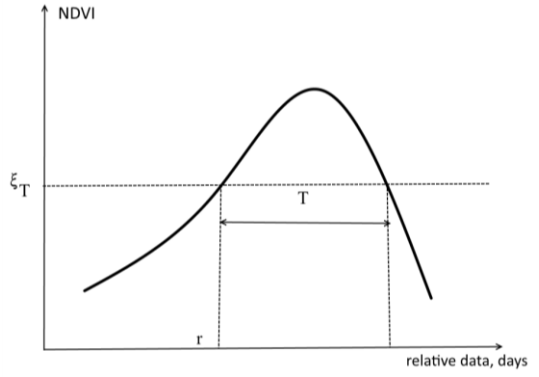

Figure 1. The first approach to estimating the remote sensing factors.

The second approach defines the vegetation period symmetrically around the maximum of NDVI value. Thus, for given vegetation period duration T, we obtain the vegetation beginning parameter in the following way

$$r = t_{\max} - T/2, \tag{5}$$

where $t_{\max} = \arg\max_t \xi(t)$. Figure 2 shows the scheme of the second approach.

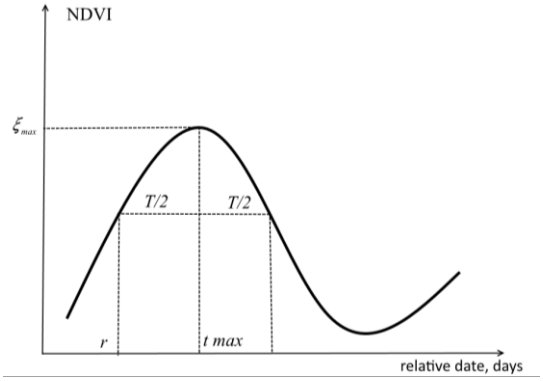

Figure 2. The second approach to estimating the remote sensing factors.

The proposed approaches determine two different subgroups of remote sensing factors. The factors corresponding to the first approach are cumulative NDVI value ξ_r, vegetation beginning date r and vegetation period duration T. The factors formed as a result of the second approach include cumulative NDVI value ξ_r, the date of maximum NDVI value $t_{\max}$ and half-width of the vegetation period $T/2$.

We tested several values of the NDVI threshold ξ_T for the first approach and several values of $T/2$ for the second approach. Nine BSFs corresponding to different remote sensing factors were formed. A complete list of the BSFs considered and their names are given in Table 1.

According to the different BSFs, nine regression models were estimated and the optimal BSF was chosen. The whole process of the optimal BSF selection is described as follows:

1. The independent categorical variables such as farmers identifier, seeds reproduction and others are previously converted in a set of binary variables taking values 0 or 1 depending on the chosen category.
2. Then all factors are normalized to get zero expectation and unit variance.
3. The normalized factors and the dependent variable values are processed by the least square method (LSM) to define the unknown regression coefficients in the expression (1).
4. For each of the BSFs, an adequacy of the constructed regression models is checked by testing the standard LSM hypotheses [7]:

Table 1. The list BSFs achieved for the different remote sensing factor extraction approaches with different parameters.

The basic factor set name	remote sensing factor extraction approach	Parameters
NDVI03	1	$\xi_T = 0.3$
NDVI04	1	$\xi_T = 0.4$
NDVI05	1	$\xi_T = 0.5$
NDVI06	1	$\xi_T = 0.6$
NDVI20	2	$T/2 = 20$
NDVI30	2	$T/2 = 30$
NDVI40	2	$T/2 = 40$
NDVI50	2	$T/2 = 50$
NDVI60	2	$T/2 = 60$

- the hypothesis about the statistical significance of the regression coefficients. This hypothesis is checked using the t-test,
- the hypothesis about the statistical significance of the regression equation. This hypothesis is checked using the F-test,
- the determination coefficient R^2 is estimated. For the proper modeling, the R^2 value has to be greater than 0.8,
- the regression residues analysis is performed. The residues have to be independent, without trend, distributed normally with zero expectation and satisfy the criteria of homoscedasticity.

5. The significant factors of regression models tested for adequacy are taken as optimal basic sets of factors if corresponding regression model has the R^2 value greater or close to 0.8.

4. Experimental research

The experimental study was conducted using the remote sensing and GIS data obtained for the season from April 1 till August 31, 2015. The farmers provided the field boundary information and semantics that includes wheat cultivar, seed reproduction, farmer identifier and crop yield in the observation period in the Agricultural GIS of Samara Region. The total number of fields used in experiments was 127. Climate zones and geographical coordinates of the field center were extracted by means of spatial queries from the vector layer in GIS. The climate zone parameter was defined as the index taking values from 1 to 3 according to the following notation [8]: 1 is northern climate zone, 2 is central climate zone and 3 is southern climate zone.

The factors based on remote sensing data were calculated using NDVI vegetation index time series. We applied the algorithm described in [9-11] and Terra and Aqua MODIS imagery with resolution 250 meters to estimate the NDVI values for each field. The remote sensing data were processed for the whole observation period from April 1 to August 31, 2015.

Climate data such as air temperature, relative humidity and precipitation were extracted from the weather archives of the following weather stations: Kinel-Cherkassy village (No. 548621) with coordinates 53°0'0"N47°0'0"E, Sernovodsk settlement (No. 496568) with coordinates 53°0'0"N92°0'0"E, Krasnoarmeyskoye village (No. 824366) with the coordinates 52°0'0"N 7°0'0"E. To estimate the climate factors we applied the measurements of the nearest to the field meteorological station among the considered ones.

As for the soil fertility factors, the soil fertility score was defined corresponding to the soil type marked on the soil map and the reference book [6]. The agrochemical survey data were provided by Agrochemical Service Station "Samarskaya".

We used 127 fields to estimate the coefficients of the Model 1 that describes the soil independent regression model. 20 fields were used to estimate the coefficients of the Model 2 that describes the soil

map-based model. And 32 fields were used to estimate the coefficients of the Model 3 that describes the model based on the agrochemical inspection data. The difference between the sample set size is explained by the various number of available data for each of the information sources used to extract the factors. The RMSE error was estimated for the same sample sets. The multiple linear regression coefficients estimation and the model adequacy verification was made using MATLAB software.

We obtained 9 models for the different BSFs listed in Table 1 and checked their adequacy. The significance level used to test adequacy was equal to $\alpha = 0.05$ for all tests. Table 2 shows the determination coefficients evaluated for each of these models. According to table 2, the significant factors of the models based on NDVI06 and NDVI02 BSFs should be taken as the optimal basic sets of factors. However, the model NDVI02 did not pass the Lilliefors test and, thus, it cannot be regarded as the statistically adequate model. The other models passed adequacy verification successfully. Therefore, we selected the model with NDVI04 BSF as an alternative to NDVI06 model.

Table 2. Determination coefficients of the models for BSFs listed in Table 1.

BSF name	Determination coefficient
NDVI03	0.7110
NDVI04	0.7162
NDVI05	0.6970
NDVI06	0.8035
NDVI20	0.7972
NDVI30	0.6999
NDVI40	0.6768
NDVI50	0.6405
NDVI60	0.7110

The further results were obtained for two OBSFs determined as the significant factors of the models based on NDVI04 BSF and NDVI06 BSF. The lists of factors included in the OBSF are given below:

1) for the model NDVI06: cumulative and maximum NDVI values during the vegetation period, the vegetation period duration and vegetation beginning date, maximum and minimum temperature during the vegetation period, maximum and minimum humidity during the vegetation period, seeds reproduction, sowing date and geographical coordinates of the field center;

2) for the model NDVI04: cumulative and maximum NDVI values during the vegetation period, the vegetation period duration and vegetation beginning date, maximum and minimum temperature during the vegetation period, maximum and minimum humidity during the vegetation period, the wheat cultivar, farmer identifier, sowing date and the geographical coordinates of the field center.

The other factors were excluded from the OBSF because they were insignificant. In further text, we will reference on OBSF according to the BSF name for which it was constructed.

To obtain soil-dependent models Model 2 and Model 3, we combined two OBSFs defined above with the SSF and TSF. Therefore, four different regression models were obtained. These models were passed the adequacy tests with the significance level $\alpha = 0.05$. The R^2 value and RMSE corresponding to the obtained soil-dependent models are given in Table 3.

Table 3. Determination coefficient R^2 and RMSE for soil-dependent models.

OBSF Name	Model 2 (OBSF+SSF)		Model 3 (OBSF+TSF)	
	R^2	RMSE	R^2	RMSE
NDVI06	0.9523	0.8866	0.8046	2.3931
NDVI04	0.9820	0.5447	0.8341	2.2051

Table 3 demonstrates that soil dependant models (Model 2 and Model 3) provide better R^2 value and RMSE than the soil-independent model (Model 1). In other words, including soil parameters

improves the crop yield forecasting quality. The best model quality corresponds to the Model 2 which applies SSF and OBSF based on NDVI04 remote sensing factors. This model has the minimum RMSE and maximum determination coefficient among the models considered.

Figures 3-8 illustrates actual and predicted yield values for the different types of regression models and OBSFs. The red color denotes the upper confidence interval for the predicted yield values. The green color denotes the lower confidence interval for the predicted yield values. The blue color denotes the yield value. The experimental results shows that the RMSE of modeling without soil parameters is in the range from 2.7 up to 3 centner per hectare and from 0.5 up to 2.4 centner per hectare with soil parameters. To summarize, we found that additional soil factors enhance the accuracy of regression modeling using NDVI data. We also recommend using soil fertility score and soil maps instead of agrochemical survey data. As for the remote sensing factors estimation, we propose to apply the first approach with the NDVI threshold equal to 0.4.

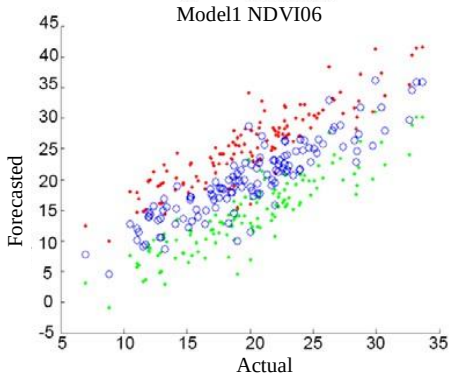

Figure 3. Diagram of actual and forecasted yield values for Model 1 with BSF NDVI06.

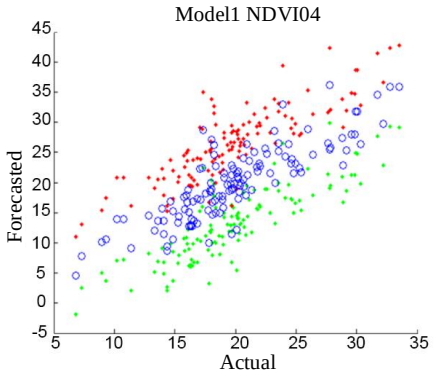

Figure 4. Diagram of actual and forecasted yield values for Model 1 with BSF NDVI04.

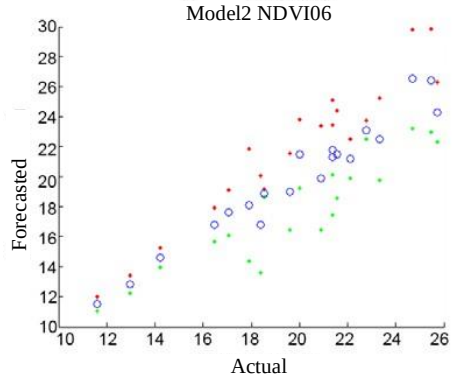

Figure 5. Diagram of actual and forecasted yield values for Model 2 with OBSF NDVI06.

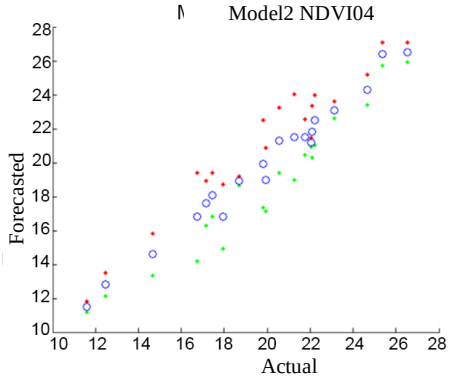

Figure 6. Diagram of actual and forecasted yield values for Model 2 with OBSF NDVI04.

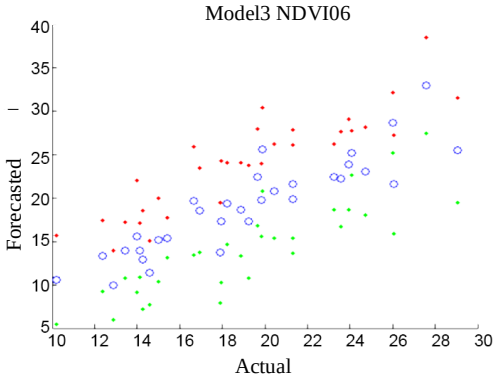

Figure 7. Diagram of actual and forecasted yield values for Model 3 with OBSF NDVI06.

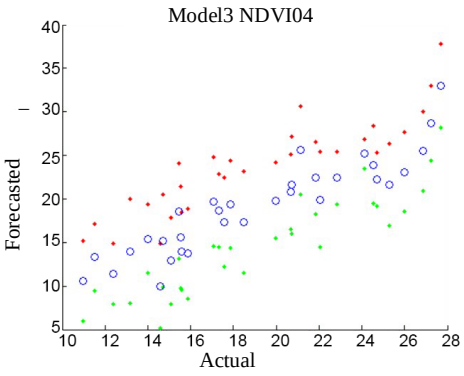

Figure 8. Diagram of actual and forecasted yield values for Model 3 with OBSF NDVI04.

5. Conclusion

The study presented in this article shows that the accuracy of winter wheat yield regression modeling using the NDVI vegetation index depends on the additional soil characteristics. We regarded several factor groups derived from the remote sensing data and the geoinformation systems to define the optimal basic factor set for the yield modeling without soil factors. Two types of additional soil factors were combined with the optimal basic factor sets, and comparative analysis of soil-dependent and soil -independent models was performed. The analysis was made for the sample fields in Samara region, Russia. As a result, we found that the soil parameters enhance the quality of winter wheat yield modeling and soil fertility score derived from a ten-year-old soil maps is provides more accurate results than the agrochemical inspection data. Therefore, only two parameters such as soil fertility score and soil type might for successfully winter wheat yield modeling instead of variety agrochemical characteristics. We also compared different approaches for the remote sensing factors computation using NDVI time series data. We concluded that the threshold 0.4 effectively determines the vegetation beginning date, vegetation period duration and cumulative NDVI in terms of the yield prediction accuracy. The results of this research can be applied to forecasting winter wheat yield in Samara region.

6. References

[1] Kayumov M K 1989 *Crop Yield Programming* (Moscow: Agropromizdat) p 368
[2] Chahbi A, Zribi M, Lili-Chabaane Z, Duchemin B, Shabou M, Mougenot B and Boulet G 2014 Estimation of the dynamics and yields of cereals in a semi-arid area using remote sensing and the SAFY growth model *International Journal of Remote Sensing* **35** 1004-1028
[3] Cabelguenne M, Debaeke P and Bouniols A 1999 EPICphase, a version of the EPIC model simulating the effects of water and nitrogen stress on biomass and yield, taking account of developmental stages: validation on maize, sunflower, sorghum, soybean and winter wheat *Agricultural Systems* **60** 175-196
[4] Ma G, Huang J, Wu W, Fan J, Zou J and Wu S 2013 Assimilation of MODIS-LAI into the WOFOST model for forecasting regional winter wheat yield *Mathematical and Computer Modelling* **58** 634-643
[5] Quarmby N A, Milnes M, Hindle T L and Silleos N 1993 The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction *International Journal of Remote Sensing* **14** 199-210
[6] Tregubov B A, Lobov G G and Holina M G 1976 *Soil fertility ranking in Kuybyshev region of Russia* (Kuybyshev: Knizhnoye isdatelstvo) p 111
[7] Kobzar A I 2006 *Applied mathematical statistics. For engineers and scientists* (Moscow: FIZMATLIT) p 816
[8] Vasin A V 2006 *Formation of highly productive multi-species agrophytocenoses of fodder crops in the Middle Volga region* (Kinel) p 513
[9] Vorobiova N S 2016 Crops identification by using satellite images and algorithm for calculating estimates *CEUR Workshop Proceedings* **1638** 419-427
[10] Vorobiova N S and Chernov A V 2016 NDVI time series modeling in the problem of crop identification by satellite images *CEUR Workshop Proceedings* **1638** 428-436
[11] Vorobiova N S, Sergeyev V V and Chernov A V 2016 Information technology of early crop identification by using satellite images *Computer Optics* **40(6)** 929-938 DOI: 10.18287/2412-6179-2016-40-6-929-938

Acknowledgements
This work was supported by the RFBR projects No. 16-29-09494 and 18-07-00748.