Efficiency of crop yield forecasting depending on the moment of prediction based on large remote sensing data set

Alexander Murynin[1]**, Konstantin Gorokhovskiy**[2] **and Vladimir Ignatiev**[3]
[1] Dorodnicyn Computing Centre of RAS, Moscow, Russia
[2] Institute for Scientific Research of Aerospace Monitoring “AEROCOSMOS”, Moscow, Russia
[3] Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Abstract—*Agricultural yields can be predicted from detailed multi-year remote sensing image sequences using measured features of vegetation conditions. In this paper, the dependency between the moment of prediction and the accuracy of the forecast is studied. The linear model is selected as a basic approach of yield forecasting. Then, the model is extended with non-linear components (factors) in order to improve the accuracy of the forecasts. The extensions take into consideration long-term technological advances in agricultural productivity as well as regional variations in yields (fertility of the lands). The accuracy of the model has been estimated based on the time period between the moment of the forecast formation and the harvest time.*

Keywords: Image mining, crop yield forecasting, nonlinear regression.

1 Introduction

Effective and efficient yield forecasting is an important area of the research which helps in ensuring food security all around the world. Nowadays yield forecasting based on multi-year observations of the land surface from space is a subject of intensive research based on data mining techniques.

The principal idea of the approach is the following. Having two years with similar observations of informative features of vegetation condition one should expect similar yields. However, the complexity of vegetation models and incompleteness of observations provides a challenge in verification of any yield forecasting method. The level of noise makes it difficult to extract a useful signal. Only by analyzing a large dataset which contains several regions and spans over many years it is possible to estimate the accuracy of a yield forecast model and reliably compare it with any alternatives.

For these reasons it is required to use a source of data that can provide reliable and accurate spatial-temporal measurements of vegetation conditions. This data can be obtained from remote sensing using satellite imaging. Various sources of remote sensing information can be used for the purposes of the crop yields forecasting as complimentary to weather measurements as well as a sole source of data [3], [4], [5], [6].

There were attempts to develop a computational algorithm which uses different channels from the multispectral radiometers [4]. As an intermediate step the multispectral images were transformed into vegetation indices. These indices were used for droughts detection as well as the crop yields forecasting. The technique has shown promising results [7], [8], [9].

Rather than studying a general accuracy of the forecasts the authors of this study concentrated on finding a dependency between the moment of prediction and efficiency of yield forecasting for the selected model.

2 Forecasting model

The proposed model can be described as follows. Crop yield of a particular culture at a given region should be fairly reliably predicted by function whose parameters are averaged (by this region) values of vegetation indices during growth and ripening period of the crop. The better the historical track record of the indices is known, the better the forecast of crop yields can be made.

The model for forecasting crop yields is based on the history of vegetation indices, accumulated over a fixed period of the year but not earlier than the start of the growing season.

The model for crop yields forecasting in general looks like:

$$y_{kr} = f_{kr}\left(v(t), v(t+1), v(t+2), \ldots\right) \qquad (1)$$

where

y_{kr} - predicted value of the yield at the end of the season for territorial region r and crop type k,

f_{kr} - unknown function of the yield forecast for the region and crop type,

$v(t)$ - vegetation index value for a region,

t - time of the start of the measurements in the current growing season, with *t+1*, *t+2*, ... corresponds to a discrete points in time when the measurements carried out during this season.

According to the recent studies in the field of crop yield forecasting there is a close correlation between vegetation indices obtained from multispectral images and productivity of plants [10], [11], [12]. In order to forecast the yield most of the studies require so called crop masks [13]. A reliable extraction of crop masks is organizationally difficult task. It requires close collaboration with farmers. Not to mention that is it often financially unfeasible activity. The proposed in this study method extracts the information from the overall condition of vegetation in the given area instead of using crop masks.

Regional administrative divisions are selected as units of the area. This choice is made due to the structure of available statistical information on the crop yields for previous years, which are officially provided by the government and publicly available. For example, the State Statistics Service of the Russian Federation allows obtaining historical information about the crop yields for all regions of the country [14]. Availability of this information makes it possible to adjust free parameters of a model to a specific region and crop type through learning process (or optimization).

From the available statistical data one can make a conclusion that the variability of the yield is small relative to its magnitude. Hence, after expansion of a yield model function in equation (1) into the Taylor polynomial the main contribution to the accuracy of the forecast will be made by the linear terms of the polynomial. As a simplification the non-linear terms of higher orders can be ignored. In this case, the model becomes linear, i.e. f_{kr} is a linear combination of $v(t)$.

2.1 Basic approach

As was mentioned earlier the model can be transformed into the linear one assuming that the soil and climate characteristics have a small variation for within (but ton between) the studied regions. The simplified linear model can be written as:

$$y_{rk} = \sum_{t=1}^{T} \alpha_{rk}(t) \cdot \langle v(t) \rangle_r \qquad (2)$$

where

k - index indicating the crop type,

r - index pointing to a territorial region of the Russian Federation,

y_{rk} - crop yield estimate for a given area r, and crop type k,

$\langle v(t) \rangle_r$ - average value of the vegetation condition index for a given territorial region, $\langle \cdot \rangle_r$ is averaging operator by region r,

$\alpha_{rk}(t)$ - adjustable parameters of the model for individual time intervals of the vegetation period (or calendar year).

The insufficient amount of statistical information available for one region makes it difficult to adjust this simple model. Indeed, only a decade of yields data is available.

Thus, the model needs to be extended in order to be used in practical applications.

2.2 Resultant model with factor adjustment for regions and temporal trend

In the case when the amount of statistical data available for the adjustment of the individual models for each of the region is not sufficient it is required to reduce the number of adjustable parameters. Thus, in particular, one can assume that the main contributions to the difference in crop yields are made by the following factors:

- fertility of soils in a region,
- climatic differences between regions,
- amount of solar radiation, depending on the latitude of a region.

At the same time to build the model, we deliberately ignore the temporary displacement of growing season for various regions, for example, for the western part of the Russian Federation taken for this study. Using the above assumptions, the following formula can be suggested:

$$y_{rk} = C_{rk} \cdot \sum_{t=1}^{T} \alpha_k(t) \cdot \langle v(t) \rangle_r \qquad (3)$$

where

k, r, y_{rk}, $\langle v(t) \rangle_r$ - were defined for equation (2),

$\alpha_k(t)$ - adjustable parameters of the model for crop type k but are now independent from the region

$\langle \cdot \rangle_r$ - averaging operator by region r,

C_{rk} - coefficient of performance of the region r for specific crop type k.

During the validation of the model described by the equation (3) was found that there are regular errors which depend from the year of the forecast. This observation was used to make a hypothesis about existence of a long-term trend in the yields. This trend hypothesis needed to be validated. In order to do that the original forecasting model has been modified to take into account the assumed trend as described further in the text.

Indeed, in the past few decades, there has been a stable growth of crop yields per unit of cultivated area [15] all over the globe. This is due to several factors. First of all, it is worth noting the progress in genetic engineering for crops improvement. Improved seeds are more resistant to drought, temperature changes and parasites. Another factor is the more efficient use of fertilizers. Progress in the field of agricultural technology has allowed to harvest with fewer losses. Improved methods of chemical treatment resulted in better control of pest populations.

Such improvements can be referred as a trend in crop yields. It is likely required to take it into account in order to improve the accuracy of the forecast. This trend may not continue but it is essential to (at least) remove this regular error from the training data.

Making an assumption that the yields changes are linearly dependent on time within the studied historic period it is possible to modify the previous model to predict the long-term increase in yields.

The average yield for the current year can be expressed from the yield of previous year by the following equation:

$$\frac{\langle y_{current}\rangle - \langle y_{start}\rangle}{\langle y_{start}\rangle} = \beta \cdot (Y_{current} - Y_{start})$$

where

$\langle y_{current}\rangle$ - average crop yield for the current year $Y_{current}$,

$\langle y_{start}\rangle$ - average crop yield in year of the beginning of observations Y_{start},

$\langle \cdot \rangle$ - averaging operator,

β - relative annual increase in productivity due to long-term trend.

Let us express $\langle y_{current}\rangle$ in terms of the other variables:

$$\langle y_{current}\rangle = [1 + \beta \cdot (Y_{current} - Y_{start})] \cdot \langle y_{start}\rangle$$

The following nonlinear regression formula for the refined model of crop yields is obtained:

$$y_{rk} = [1 + \beta \cdot (Y - Y_{start})] \cdot C_{rk} \cdot \sum_{t=1}^{T} \alpha_k(t) \cdot \langle v(t)\rangle_r$$

where

k, r, y_{rk}, $\langle v(t)\rangle_r$, $\alpha_k(t)$, C_{rk} - were defined for equations (2) and (3),

Y - current year for which the crop yields are evaluation,

Y_{start} - the year of the beginning of observations,

β - relative annual increase in productivity due to long-term trend.

Unlike initial linear approach this model can no longer be qualified as a linear but rather a factor model due to multipliers describing productivity of a region and the trend.

Authors made attempts to reduce this model back to linear one by adding coefficients C_{rk} and $[1 + \beta \cdot (Y - Y_{start})]$ but the accuracy of the model has been reduced drastically in this case. It can be explained by significant variation of the above mentioned multipliers (which can be also called factors). For example the productivity (fertility) C_{rk} can differ by the factor of 2 between the regions.

On the other hand insufficient data per region makes it impossible to build separate linear model per individual region.

2.3 Forecast accuracy and the moment of the prediction

One can assume that the earlier in time we are making the forecast the less accurate it will be. In the contrary the closer we get to a harvest the more reliable forecasts we can achieve. Usually, it is required to know how reliable the forecast is depending on the date of the prediction. This study tries to provide the answer to this question for the described above model.

3 Results

The accuracy of the model was assessed using K-fold cross-validation method. The whole set of the input data has been partitioned several times into two subsets: the training subset and the testing subset. Each time the testing subset was different. In total 10 unique testing subsets were used so that the data for each year available were used as a testing subset at least once.

In addition, the dependency between the accuracy of the forecast and the moment of the forecast was studied. In each case it was assumed that the remote sensing data was available up to the moment of the forecast. That is: if, for example, a prediction takes place in August 13 one can assume that all the remote sensing data (for this year) prior to this date is already available for the analysis.

Cross-validation is used to evaluate the performance of the forecasting model in a manner similar to that which is commonly used for classifiers.

Due to insufficient amount of statistical data during the validation the chronological order of training data and validation data was not preserved. This does not jeopardize the validation for the following two reasons:

1) the forecasting scenario for each year is based on processing of the current year data and does not depend of the data from other (including previous) years.

2) the forecasting algorithm uses only data that strictly precede the forecasting time within the giving vegetation period (within the current year). In other words, the model uses only past observations for each forecasting moment and does not involve any future data within the considered year.

Remote sensing data for 14 regions of Russian Federation over span of 10 years (from 2000 to 2009) were used for training and validation of the model. Total data set used for training and validation consisted of more than 1500 images with dimensions 2400 x 2400 pixels each. The size of the images set was more than 54 GB. After the process of model training was complete the smaller set of images was used in the forecast for a given year. The images used in the forecast represent 7 separate moments in time with 16 days distance from each other. These images are 16 days cloudless composites snapshots with resolution of 500 meters captured by MODIS TERRA satellite.

For example, figure 1 shows the image with vegetation condition index (NDVI) for 3 regions of the Russian Federation: Ivanovo, Vladimir and Nizhny Novgorod regions. The image represents values of the index for 9 May 2007.

In order to simulate the change in prediction date the snapshots used in the model were selected in a “sliding window” manner. This is to maintain the number of snapshots constant and equal to 7. The constant number of observations was required to avoid model overfitting and preserve the ratio of the amount of training cases versus the number of coefficients in the model.

The resultant accuracies of prediction for two groups of cultures are shown in Table 1. Forecasting errors of crop yields is evaluated in the form standard deviation of forecasted values from the yield data available through the official statistics.

As can be seen from Table 1 the worst result is generated in late spring / early summer. This is due to the fact that information about vegetation condition in early stages of growth is less informative than in final stages. The visual representation of the forecasting errors is shown in Figure 2.

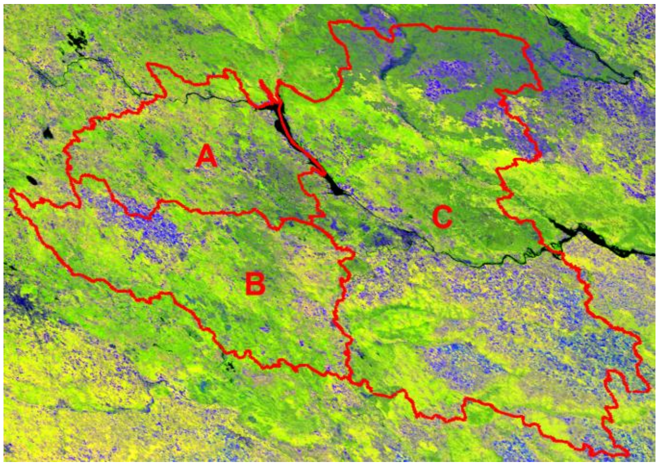

Fig. 1. The area in study: for Ivanovo (A), Vladimir (B) and Nizhny Novgorod (C) regions for 9 May 2007 (Vegetation index map).

It is worth noting that the proposed model does not require crop masks which are usually used in similar studies [13]. Our method extracts the required information from the overall condition of vegetation in the given area rather than condition of a given crop. The lack of crop mask may reduce the accuracy of the forecasts. Nevertheless, the comparison of our results with the results from other studies [13] shows that our model demonstrate competitive accuracy even without the crop mask or other information about cultivated areas such as soil types and weather conditions.

TABLE 1
STANDARD DEVIATION OF THE FORECASTS CROP YIELDS FOR DIFFERENT CULTURES USING CROSS-VALIDATION FOR THE MODEL WITH FACTOR ADJUSTMENT FOR REGIONS AND TEMPORAL TREND FOR THE PERIOD 2000-2009. BEST ACCURACIES ARE MARKED WITH BOLD ITALIC FONT.

	Date of the forecast					
	June 10	June 26	July 12	July 28	August 13	August 29
Grain	16,1%	15,2%	13,7%	12,7%	***12,5%***	13,5%
Potato	19,8%	22,1%	20,4%	18,7%	18,0%	***16,9%***

4 Conclusion

This study introduces an approach to develop an efficient model for crop yield forecasting via extracting information from the large set of satellite images.

Also, the dependence between the moment of the forecast and its accuracy has been studied. It is shown the closer to the harvest the prediction is performed the better accuracy can be achieved. However, the useful forecast can be done even several months before the harvest.

The dependency of forecasting errors from the date of the forecast is shown in Figure 2 for the yields of grain and potatoes.

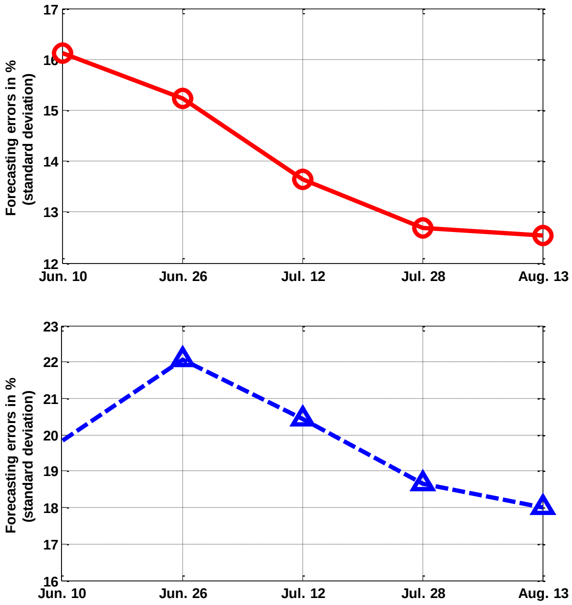

Fig. 2. Standard deviation of yield predictions for grain (top image) and potato (bottom image) cultures. As can be seen the accuracy of predictions improves gradually as we get closer to the harvest.

The main advantage of the suggested approach is the possibility to use free to access information, including satellite multispectral images and official statistical data.

It is shown that by finding out the appropriate form of forecasting function on the basis of remote sensing images and official government statistics data is possible to obtain fairly accurate results of yields forecasting.

Other advantage is that the proposed approach does not require any specific information about the cultivated areas. It minimizes the amount of the input data for practical implementation of the models. Specifically, this approach does not require crop masks. In other words the method uses overall condition of the vegetation in the given area rather than the condition of specific culture.

The analysis of the accuracy of forecasting crop yields using cross-validation method demonstrates the advantages and disadvantages of the proposed approach. The model with factor adjustment for regions and temporal trend allows obtaining forecasting errors from 12% to 22% depending on the culture, and the moment in time of the forecast. The closer we get to the harvest the better accuracy we can expect for such kind of forecasts.

We plan to continue this study with enhanced forecasting models in order to improve the accuracy and generality of the crop yield prediction as well as extend the forecasts to cover the more territorial regions.

References

[1] J. D. McQuigg, "Economic Impacts of Weather Variability," Atmospheric Science Dept University of Missouri, Columbia, 1975

[2] T. Hodges, D. Botner, C. Sakamoto and J. Hays Haug, "Using the CERES-Maize model to estimate production for the U.S." *Cornbelt. Agricultural and Forest Meteorology*, vol. 40, iss. 4, pp. 293-303, 1987.

[3] C. J. Tucker and P. J. Sellers, "Satellite remote sensing of primary production," *International Journal of Remote Sensing*, vol. 7, iss. 11, 1986.

[4] F. N. Kogan, "Global Drought Watch from Space," *Bulletin of the American Meteorological Society*, no. 78, pp. 621-636, 1997.

[5] R. Benedetti and P. Rossini, "On the use of NDVI profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forecast in Emilia Romagna," *Remote Sensing of Environment*, vol. 45, pp. 311–326, 1993.

[6] M. S. Rasmussen, "Operational yield forecasting using AVHRR NDVI data: prediction of environmental and inter-annual variability," *International Journal of Remote Sensing*, vol. 18, pp. 1059–1077, 1997.

[7] L. S. Ungana and F. N. Kogan, "Drought monitoring and corn yield estimation in Southern Africa from AVHRR data." *Remote Sensing of Environment*, vol. 63, pp. 219–232, 1998.

[8] E. Aigner, I. Coppa and F. Wieneke, "Crop Yield Estimation Using NOAA − AVHRR Data and Meteorological Data in the Eastern Wimmera (South Eastern Australia)," *International Archives of Photogrammetry and Remote Sensing*, vol. 33, part B7, Amsterdam, 2000.

[9] Cs. Ferencz, P. Bogna, R. J. Lichtenberger, D. Hamar, Gy. Tarcsai, G. Timar, G. Molnar, Sz. Pasztor, P. Steinbach, B. Szekely, O. E. Ferencz and I. Ferencz-Arkos, "Crop yield estimation by satellite remote sensing," *International Journal of Remote Sensing*, vol. 25, no. 20, pp. 4113–4149, 2004.

[10] L. B. Phillips, A. J. Hansen and C. H. Flather, "Evaluating the species energy relationship with the newest measures of ecosystem energy: NDVI versus MODIS primary production." *Remote Sensing of Environment*, vol. 112, iss. 9, pp. 3538-3549, 2008.

[11] M.P. Kale, Sarnam Singh and P.S. Roy, "Biomass and productivity estimation using aerospace data and Geographic Information System" *Tropical Ecology*, vol. 43 no. 1, pp. 123-136, 2002.

[12] G. Edward, H. Alfredo, N. Pamela and N. Stephen, "Relationship Between Remotely-sensed Vegetation Indices, Canopy Attributes and Plant Physiological Processes: What Vegetation Indices Can and Cannot Tell Us About the Landscape," *Sensors 8*, no. 4, pp. 2136-2160, Mar. 2008.

[13] A. S. Islam and S. K. Bala, "Estimation of yield of wheat in greater Dinajpur region using Modis data," presented at 3rd International Conference on Water & Flood Management, ICWFM-2011, 2011.

[14] Regions of Russia. Social and Economic Indicators. 2011. Available: http://www.statbook.ru/eng/catalog.html?page=info&id=306

[15] R. A. Fischer, D. Byerlee and G. O. Edmeades, "Can Technology Deliver on the Yield Challenge to 2050?," presented at the Expert Meeting on How to Feed the World, Food and Agriculture Organization of the United Nations, Rome, 2009.