Crop Yield Forecasting for Major Crops in Ukraine

Andrii Shelestov
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Space Research Institute NASU-SSAU
Kyiv, Ukraine
andrii.shelestov@gmail.com

Leonid Shumilo
The University of Maryland
Maryland, USA
shumilo.leonid@gmail.com

Hanna Yailymova
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Space Research Institute NASU-SSAU
Kyiv, Ukraine
anna.yailymova@gmail.com

Sophia Drozd
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine
sofi.drozd.13@gmail.com

***Abstract*—The annual harvest growth is very important for the development of the agricultural sector of each country. To ensure its growth, there is a need for modern monitoring of agricultural land indicators. These indicators are usually obtained directly from local agronomists, but this method of collecting information is too long and complicated. As a state-of-the-art alternative is the use of remote sensing data from satellites. Using remote sensing technologies, it is possible to predict the yield of major crops in the studied areas by classical methods of machine learning, in particular, regression analysis. In this paper, a regression model for yield forecasting for the territory of Kyiv region (Ukraine) will be create based on vegetation indices obtained from satellites data. The main purpose is to assess the reliability of the obtained model. The paper will refute or confirm the feasibility and effectiveness of the satellite data using in regression models for crop yield forecast for Kiev region.**

Keywords — regression, analysis, satellite, yield, agricultural, vegetation.

I. Introduction

The agricultural sector of each country has a direct impact on the economy. It is very important to increase the yield of crops for this industry development. The process of monitoring agricultural land indicators is important for improving yields. Using modern mathematical approaches, it is possible to predict the yield for the majority crops for the next year, in particular, having data on soil parameters. Thus, it is possible to make the most cost-effective decision on sowing specific crops in certain fields and minimize crop failure.

However, the collection of data on agricultural indicators of all fields in the country is a very complex and time-consuming process, if we obtain data directly from the records of local agronomists and farmers. The collection of this information requires the involvement of a large number of resources, the establishment of reliable two-way information channels for further communication. In addition, studies by local agronomists may not be reliable enough. To solve this problem, the world has been practicing data collection by remote sensing for many years. Until a few decades ago, this method allowed to assess agricultural soil indicators only in general, at the oblast or regional levels. Usually, it is the regression task and the regressors in it are: Vegetation indices such as NDVI, DVI and etc. [1], biophysical parameters such as LAI [2] obtained using remote sensing data, modeled date that reflects crop growth and state [3] and agroclimatic data such as air temperature, land surface temperature and precipitation [4]. Usually, the best way to build high quality yield forecasting model is combination of regressors with different types [5], [6]. However, the latest technology allows to get data even on the individual fields level. Using remote sensing technologies, it is possible to predict the yield of major crops in the studied areas by classical methods of machine learning, in particular, regression analysis.

The topic of this study is to create a regression model for yield prediction for the Kyiv region based on remote sensing data. The main purpose is to verify the reliability of the constructed model. The work is designed to assess the appropriateness of the use of remote sensing data in forecasting yields in the Kyiv region. The structure of the work consists of the following parts: the practice of using remote sensing data in the world, the choice of data and materials for research, a description of the method of research, results, conclusions

II. The Practice of Using Remote Sensing Data in the World

With the introduction of remote sensing data, science has taken a big step forward. In particular, in the crop yield forecasting. The best way to build the high-quality yield forecasting model is a combination of different remote sensing data sources of different types and physics with the addition of in-situ measurements. The world experience in the yield forecasting shows a good informativeness of local agroclimatic data. The weather stations on the fields are the best data sources, but it is also possible to use the nearest weather stations and geospatial data interpolation techniques. Such an approach was used for the strawberry yield forecasting at the filed level in California, USA [7]. The Fig 1 and Fig. 2 shows the weather parameters and adjusted R^2 score for respective linear models.

In this way the best regressors were fall average soil temperature, net radiation, solar radiation, cumulated chill hours, volumetric soil moisture, soil temperature, ambient temperature.

A great example of field level yield forecasting using local data shown in [8]. The research was made for maize

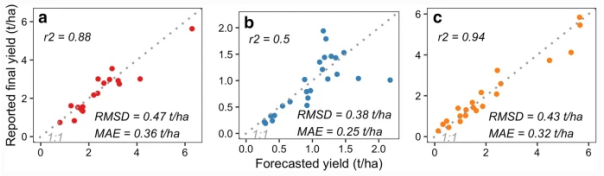

Fig. 1 Comparisons between the forecasted yield and reported final yield across (a) Morogoro, (b) Kagera and (c) Tanga (note that the ranges for both axes in a – c differ)

Id	Parameters	Units	Notation	Daily	Moving We
1	Average leaf wetness minutes	minutes	*LWM*	0.004	0.041
2	Average leaf wetness count		*LWC*	0.120	0.274
3	Average leaf wetness duration		*LWD*	0.092	0.148
4	Ambient temperature	°C	*ECT1*	0.460	0.505
5	Canopy temperature	°C	*ECT2*	0.030	0.116
6	Soil temperature	°C	*SMTa*	0.407	0.495
7	Volumetric soil moisture	m^3/m^3	*SM*	0.417	0.547
8	Daily chill hours	hours	*CHDaily*	0.189	0.292
9	Cumulated chill hours	hours	*cumChill*	0.431	0.462
10	Reference evapotranspiration	mm	*ETo*	0.338	0.585
11	Solar Radiation	Wm-2	*Rs*	0.421	0.667
12	Net Radiation	Wm-2	*Rn*	0.439	0.656
13	Average vapor pressure	kPa	*em*	0.134	0.210
14	Average relative humidity	%	*RHm*	0.000	0.002
15	Dew point	°C	*dP*	0.157	0.234
16	Average wind speed	ms-1	*uBar*	0.261	0.354
17	Penmann-Montieth Evapotranspiration	mm	*PMETo*	0.384	0.648
18	Fall reference evapotranspiration	mm	*ETo.F*	0.244	0.356
19	Fall solar radiation	Wm-2	*Rs.F*	0.129	0.196
20	Fall net radiation	Wm-2	*Rn.F*	0.242	0.300
21	Fall average vapor pressure	kPa	*em.F*	0.084	0.143
22	Fall average air temperature	°C	*aTm.F*	0.270	0.449
23	Fall average relative humidity	%	*RHm.F*	–0.005	–0.006
24	Fall average wind speed	ms-1	*u.F*	0.020	0.055
25	Fall dew point	°C	*dP.F*	0.071	0.116
26	Fall average soil temperature	°C	*STm.F*	0.739	0.748

Fig. 2 Weather parameters and adjusted R^2 score for respective linear models for yield forecasting fitted on it

yield in smallholder farmers' fields in Tanzania. The local data were obtained though the field questionnaire survey in the Survey Solutions. Survey Solutions is a computer-assisted personal interviewing software developed by the World Bank. The trained enumerators administered the field questionnaire survey using tablets with the questionnaire coded in the Survey Solutions application. The enumerators recorded the geographic location and surveyed the physical characteristics of the within-season plant (including planting density, stress level due to N, drought, weeds, pests and diseases) condition.

Other in-season information (including weather characteristics and maize cultivar, sowing time, irrigation and fertilization levels) were from enumerators' interviews with the farmers or farm workers. The complete survey was synchronized to the cloud storage. The authors processed the within-season information immediately after they received it through the cloud storage and provided the maize yield forecast for each of the sampling fields. They provided yield forecasts ranging from 14 to 77 days prior to harvest. The 25th and 75th percentile of the forecasting lead time was 30 and 55 days before harvest, respectively.

On Fig. 3 and Fig. 4 are shown the reported maize growth conditions on the located fields for three regions. The yield forecasting method is based on this reported data and SALUS crop model [9]. Yield forecasting models are different for these three regions, but the R^2 score for them is ranged from 0.5 to 0.94 (Fig. 4).

Data collection in the large scale is not an easy task, so it is also possible to combine field level and global datasets using different reanalysis and harmonization approaches. The weather data on filed level provide better performance rather than global products, but the number of fields with weather stations is not so big. Thus, it is possible to combine weather

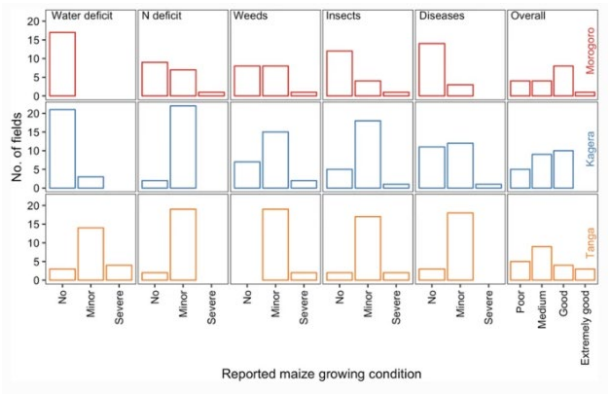

Fig. 3 Maize status, including water and N deficit, weed, insect and disease presence, and overall plant condition based on photos taken during in-season survey across the three districts.

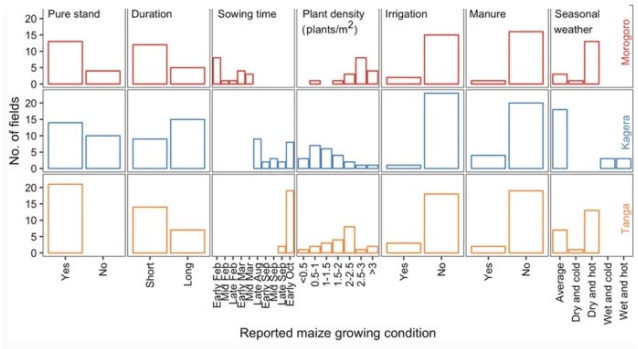

Fig. 4 Reported maize growing conditions, including pure crop stands versus intercropping, maize duration, sowing time, plant density, irrigation and manure use and growing season weather characteristics across the three districts

stations to conduct reanalysis of global products and obtained more accurate dataset for yield forecasting [10]. However, using field level data it is possible also run such advance biophysical models, such as WOFOST to obtain high quality crop growth data.

Thus, world practice has shown that the data of remote sensing and regression analysis with a reasonable choice of regressors based on them can be used to reliably predict the yield of different crops.

III. Data and Materials

Before selecting the data for the study, we made a number of input requirements for the correct yield forecast result. Therefore, the input data must meet the following conditions:

1. High variation of the data in the small range
2. Lack of useful historical information for the fields
3. Small scale of data
4. Limitation for the number of fields leads to the model's overfitting

According to these conditions, two data sets were prepared for yield forecasting. Both were based on the MODIS MOD13Q1 NDVI product collection created using Google Earth Engine. The first data set is the maximum NDVI between March 1 and June 1. It was used to predict the yield of winter wheat. The second data set is the maximum NDVI

between March 1 and September 1. It was used to predict the yield of summer crops.

As the yield forecasting strategy, we used the model fitting for each region based on the historical data and yield trend (from 2016 to 2021). In this case we are using one year from 2016 to 2020 as validation year, 2021-year prediction year and others as training data. To estimate more accurate model with more appropriate accuracy metric, we iteratively doing the model fitting and 2021 yield forecasting by changing validation year and averaging the final scores outputs. The model's fitting was conducted for the winter wheat and maize for Kyiv region.

Thus, for the study, we selected two data sets based on the MODIS MOD13Q1 NDVI product collection created using Google Earth Engine as the maximum NDVI of winter wheat from March 1 to June 1 and summer crops from March 1 to September 1. The main method used was to adjust the model for each region based on historical data and yield trends from 2016 to 2020. The harvest was projected for 2021.

IV. METHODS

During the experiment, the method of constructing linear regression was used. Linear regression is described by the following formula:

$$y = b_0 + b_1 * x_1 + .. + b_n * x_n,$$

where y is the predicted yield, x_i is the value of NDVI regressors, b_i - coefficients near the corresponding regressors.

A compliance model for each region was used as a yield forecasting strategy. The statistics were compared with those obtained by us during the experiment.

To evaluate a more accurate model with a more acceptable accuracy metric, an iterative fit of the model and yield forecasting for 2021 was performed, changing the year of validation and averaging the final scores. The following formula was used for averaging:

$$x_m = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Here x_m is the average value of the results of the final points, $x_1 \ldots x_n$ - results of final points, n is the number of results. An iterative trend model was established based on yield data with a loss of the test year.

The next step was to calculate the difference between the actual yield and the corresponding trend point for the respective year. This was done by the formula of absolute error:

$$\Delta y = y_f - y_t,$$

where Δy is the absolute error, the required difference between the actual yield and the corresponding trend point, y_f - true value, actual yield, y_t - the result of the experiment, the value of the yield on the trend line.

The average absolute error was also calculated. The formula is as follows:

$$\text{MAE} = \frac{\sqrt{\sum \Delta y_i}}{n},$$

where Δy_i is the difference between the actual yield and the corresponding i-th point on the trend line, n is the number of fixed points on the trend line.

The mean square error was also calculated to estimate the error of the experiment.

$$\text{MSE} = \frac{\sqrt{\sum \Delta y_t^2}}{n}.$$

To assess the quality of the regression model, the reliability coefficient R^2 was calculated

$$R^2 = 1 - \sum \frac{(y_i - y_f)^2}{(y_f - \Delta y_{fi})^2},$$

where R^2 is an estimate of the accuracy of the model, y_f - true value, actual yield, y_t - the result of the experiment, the value of yield on the trend line, Δy_{fi} is the average of all true yield values.

Thus, the main method we used in our study was to build a linear regression model with subsequent assessment of its reliability based on the measurement of errors.

V. RESULTS

The Fig. 5 and Fig. 6 show the averaged trend models for winter wheat and maize. Blue points – yield by statistics, dark blue line – yield trend, r^2 – r squared metrics, MSE1 – mean squared error between trend and merged validation & training data, MSE2 – mean square error between trend and validation data, MAE1 – mean absolute error between trend and merged validation & training data, MAE2 – mean absolute error between trend and validation data.

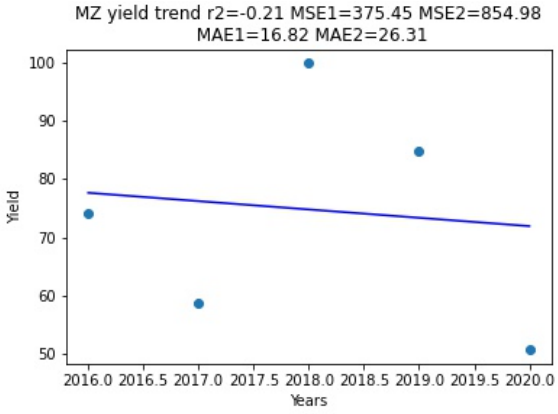

Fig. 5 Model of maize yield trend for Kyiv region

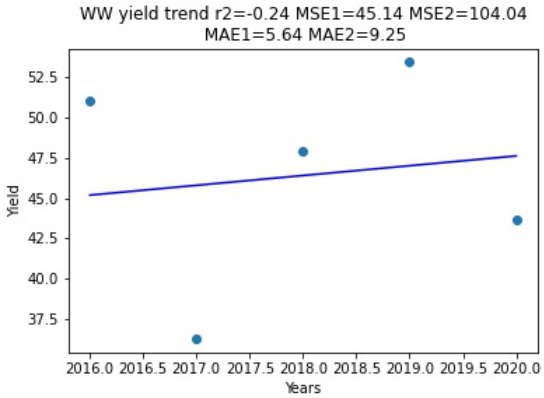

Fig. 6 Model of winter wheat yield trend for Kyiv region

The differences between the actual yield and the corresponding trend point for the respective year were used for the linear regression function, which is suitable for forecasting the yield. The Fig. 7 and Fig. 8 show the obtained models for forecasting the yield of winter wheat and maize. There green dots - profit according to statistics, red dots - projected yield, dark blue line - trend of profitability, r^2 - r - metric squared, MSE1 - mean square error between projected yield and combined validation + training data, MSE2 - average quadratic error between predicted yield and validation data, MAE1 is the mean absolute error between

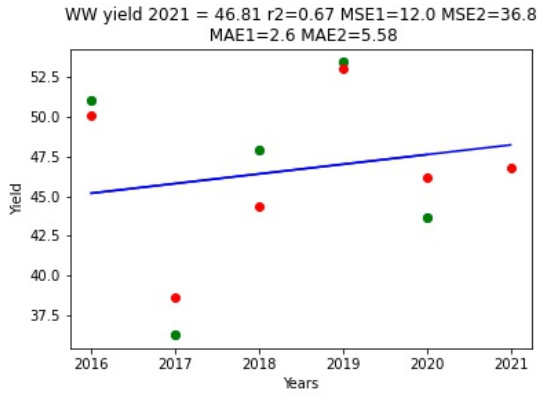

Fig. 7 Forecast of winter wheat yield

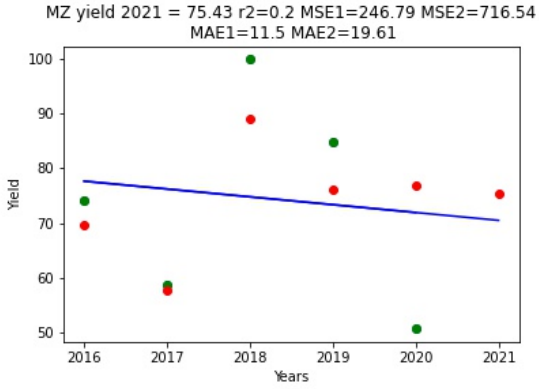

Fig. 8 Maize yield forecast

predicted yield and pooled validation + training data, MAE2 is the mean absolute error between predicted yield and validation data.

The forecast of winter wheat yield in 2021 is 46.81 t / ha, and for the labyrinth - 75.43 t / ha. Winter wheat has the highest score R^2, which is equal to 0.67, and low MAE - 2.6. Maize has a much lower R^2 of 0.2 and an MAE of 11.5. The main problem with these models is the significant drop in 2020 due to weather conditions and drought. This especially affects the result of maize, where the difference between the actual and projected yield is almost 29 t / ha.

Thus, the predicted regression model showed average reliability.

VI. Conclusions

During the experiment it was found that the data from the collection of MODIS MOD13Q1 products obtained by remote sensing can be used to predict the yield of individual crops in the Kiev region. In particular, winter wheat.

The main problem, due to which the error in forecasting the yield of other crops increased, was the inferiority of the regression model, which took into account only the vegetation indices for previous years. After all, the yield is significantly affected by weather conditions. They should be taken into account for better forecasting. However, this is not a problem with data obtained by remote sensing.

Thus, in conclusion, it can be argued that remote sensing in combination with other methods of obtaining agroclimatic indicators can be used to build reliable models for yield threat. Thus, remote sensing is a great new way of collecting data, which is easy to find a successful practical application.

Acknowledgment

The authors acknowledge the funding received by the National Research Foundation of Ukraine from the state budget 2020/01.0273 “Intelligent models and methods for determining land degradation indicators based on satellite data” (NRFU Competition “Science for human security and society”).

References

[1] B. Franch, E. F. Vermote, S. Skakun, J. C. Roger, I. Becker-Reshef, E. Murphy, and C.Justice, "Remote sensing based yield monitoring: Application to winter wheat in United States and Ukraine," International Journal of Applied Earth Observation and Geoinformation, vol. 76, pp. 112-127, 2019.

[2] A. Kolotii, N. Kussul, A. Shelestov, S. Skakun, B. Yailymov, R. Basarab, and V. Ostapenko, “Comparison of Biophysical and Satellite Predictors for Wheat Yield Forecasting in Ukraine,” International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. XL-7/W3, 2015, pp. 39-44.

[3] G. Ma, J. Huang, W. Wu, J. Fan, J. Zou, and S. Wu, “Assimilation of MODIS-LAI into the WOFOST model for forecasting regional winter wheat yield,” Mathematical and Computer Modelling, vol. 58, iss. 3-4, 2013, pp. 634-643.

[4] J. Mathieu, F. Aires, “Assessment of the agro-climatic indices to improve crop yield forecasting,” Agricultural and forest meteorology, vol. 253, 2018, pp. 15-30.

[5] S. Skakun, E. Vermote, B. Franch, J. Roger, N. Kussul, and J. Masek, “Winter Wheat Yield Assessment from Landsat 8 and Sentinel-2 Data: Incorporating Surface Reflectance, Through Phenological Fitting, into Regression Yield Models,” Remote Sensing., vol. 11, No. 15, 1768, 2019.

[6] F. Kogan, N. Kussul, T. Adamenko, S. Skakun, O. Kravchenko O. Kryvobok, A. Shelestov, A. Kolotii, O. Kussul, A. Lavrenyuk A., “Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models,” International Journal of Applied Earth Observation and Geoinformation, vol. 23, 2013, pp. 192-203.

[7] M. Maskey, T. Pathak, S. Dara, “Weather based strawberry yield forecasts at field scale using statistical and machine learning models,” Atmosphere, vol. 10, iss. 7, 2019, p. 18.

[8] L. Liu, L., B. Basso, “Linking field survey with crop modeling to forecast maize yield in smallholder farmers' fields in Tanzania,” Food Security, vol. 12, iss. 3, 2020, pp. 537-548.

[9] K. Dzotsi, B. Basso, J. Jones, “Development, uncertainty and sensitivity analysis of the simple SALUS crop model in DSSAT,” Ecological Modeling, vol. 260, 2013, pp. 62-76.

[10] A. Santamaria-Artigas, B. Franch, P. Guillevic, J. Roger, E. Vermote, S. Skakun, “Evaluation of Near-Surface Air Temperature from Reanalysis Over the United States and Ukraine: Application to Winter Wheat Yield Forecasting.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, iss. 7, 2019, 2260 - 2269.