CHEMICAL ENGINEERING TRANSACTIONS

VOL. 94, 2022

A publication of AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Petar S. Varbanov, Yee Van Fan, Jiří J. Klemeš, Sandro Nižetić
Copyright © 2022, AIDIC Servizi S.r.l.
ISBN 978-88-95608-93-8; **ISSN** 2283-9216

DOI: 10.3303/CET2294116

Modelling of Alfalfa Yield Forecasting Based on Earth Remote Sensing (ERS) Data and Remote Sensing Methods

Marzhan Anuarbekovna Sadenova[a*], Nail Alikuly Beisekenov[a], Baitak Apshikur[a], Sergey Sergeevich Khrapov[b], Azamat Kaisarovich Kapasov[a], Asel Mukhtarkanovna Mamysheva[a], Jiří Jaromír Klemeš[a]

[a]Priority Department Centre «Veritas» D. Serikbayev East Kazakhstan technical university, 19 Serikbayev str. 070000 Ust-Kamenogorsk, Kazakhstan
[b]Volgograd State University, 100 Prospekt Universitetskiy. 400062 Volgograd, Russian Federation
MSadenova@ektu.kz

This study aims to develop a method for modelling early forecasting of alfalfa yield on a farm scale located in East Kazakhstan. The authors evaluated the correlation coefficient between forage crop yield and different data sets, including weather data, climate indices, spectral indices from drones and satellite observations. An ensemble machine learning model was developed by combining three commonly used basic training modules: random forest (RF), support vector method (SVM), and multiple linear regression (MLR). It is found that the best yield prediction algorithm in this study is the Random Forest (RF) algorithm, which predicts yields with $R^2 = 0.94$ and RMSE = 0.25 t/ha. The results of this study showed that combining remote sensing drought indices with climatic and weather variables from UAV and satellite imagery using machine learning is a promising approach for alfalfa yield prediction.

1. Introduction

World and domestic experience show that high and sustainable productivity of farming is possible only when all agrochemical and environmental factors necessary for normal growth and development of plants, formation of yields and their quality, and prevention of land degradation are taken into account in the complex. Rational use of agricultural lands and soil protection under market conditions requires adequate application of new scientific and methodological approaches. One of such system-analytical ways for organisations is a combination of traditional ground methods with geoinformation systems (GIS) technologies on the basis of the wide use of aerospace images of different resolutions. Remote sensing data has become vital for mapping features of terrestrial landscapes and infrastructures, managing natural resources, and studying environmental change (Habarov et al., 2019). Crop mapping and crop evaluation are the simplest but most important issues in agriculture. Sentinel-2 satellite data have been used extensively for these tasks over the past few decades (Nihar et al., 2019). In Guo et al. (2021), boundary line analysis shows that relative yield increases of 8-10% can be obtained by optimising yield-limiting factors. Timely yield estimation can help in making accurate management decisions, but traditional yield estimation approaches are labour-intensive and time-consuming, which hinders timely information in the field. Recently, unmanned aerial vehicles (UAVs) have attracted considerable attention in precision agriculture because of their efficiency in data collection. In addition, compared to other imaging methods, hyperspectral data can provide higher spectral accuracy for constructing narrow-band vegetation indices, which are important in yield modelling. Accurate seasonal forecasting of grain yields is an important decision support tool (Bouras et al., 2021). Yang et al. (2020) estimated land productivity as the potential for agricultural production by considering biophysical properties, including climate, soil, and land slope. Land productivity is approximated by the potential yield of six major crops: corn, soybeans, winter wheat, spring wheat, cotton, and alfalfa. In Yadav et al. (2021), the combination of climate and normalised difference vegetation index (NDVI) variables produced more accurate predictions compared to using NDVI alone to predict wheat, sorghum, and corn yields. The reason for the limitations of existing methods for yield forecasting is that

Paper Received: 19 May 2022; Revised: 15 June 2022; Accepted: 17 June 2022
Please cite this article as: Sadenova M.A., Beisekenov N.A., Apshikur B., Khrapov S.S., Kapasov A.K., Mamysheva A.M., Klemeš J.J., 2022, Modelling of Alfalfa Yield Forecasting Based on Earth Remote Sensing (ERS) Data and Remote Sensing Methods, Chemical Engineering Transactions, 94, 697-702 DOI:10.3303/CET2294116

the models are not adapted to the soil and climatic conditions of Kazakhstan. In this study, a seasonal forecast of alfalfa yield was performed using a combination of satellite and drone spectral imagery. The goal of our study is to develop an ensemble machine learning model by combining three widely used basic training modules to evaluate and determine the best algorithm.

2. Materials and methods

In this study, a large number of spectral indices were extracted from an array of satellite images and aerial photography results from unmanned aerial vehicles (UAVs) in the first phase of the study. Two Geoscan 201 Agro and DJI Phantom 4 Multispectral UAVs were used. These are specialised aircraft used in agricultural enterprises, the forestry industry, and all other industries where plant condition monitoring is required. Unlike a conventional optical camera, the DJI model provides maximum information by capturing different spectra and allows you to visually detect abnormalities and quickly make decisions to correct them. The additional spectral channels (blue, green) provide a unique opportunity to calculate not only the NDVI index (+NDRE) but also the Enhanced Vegetation Index (EVI), which provides enhanced baseline data. Geoscan 201 Agro is used for aerial surveys with a range of 30 km and a flight time of up to 3 hours. Geoscan 201 Agro allows to survey up to 8,000 ha/d and get orthophotos with georeferencing accuracy, corresponding to 1:500 scale requirements.

Necessary parameters were selected to decrease the data dimensionality. Data on weather, crop yield, and spectral data were collected from Sentinel-2, Landsat-8, and TERRA (MODIS scanner) satellites. In the next step, an ensemble machine learning model was developed by combining three commonly used basic training modules: random forest (RF) (Belgiu et al., 2016), support vector method (SVM) (Cihlar et al., 1991), and multiple linear regression (MLR) (Eberly, 2007). A schematic diagram of the proposed research methodology with an overview of the main input data is presented in Figure 1.

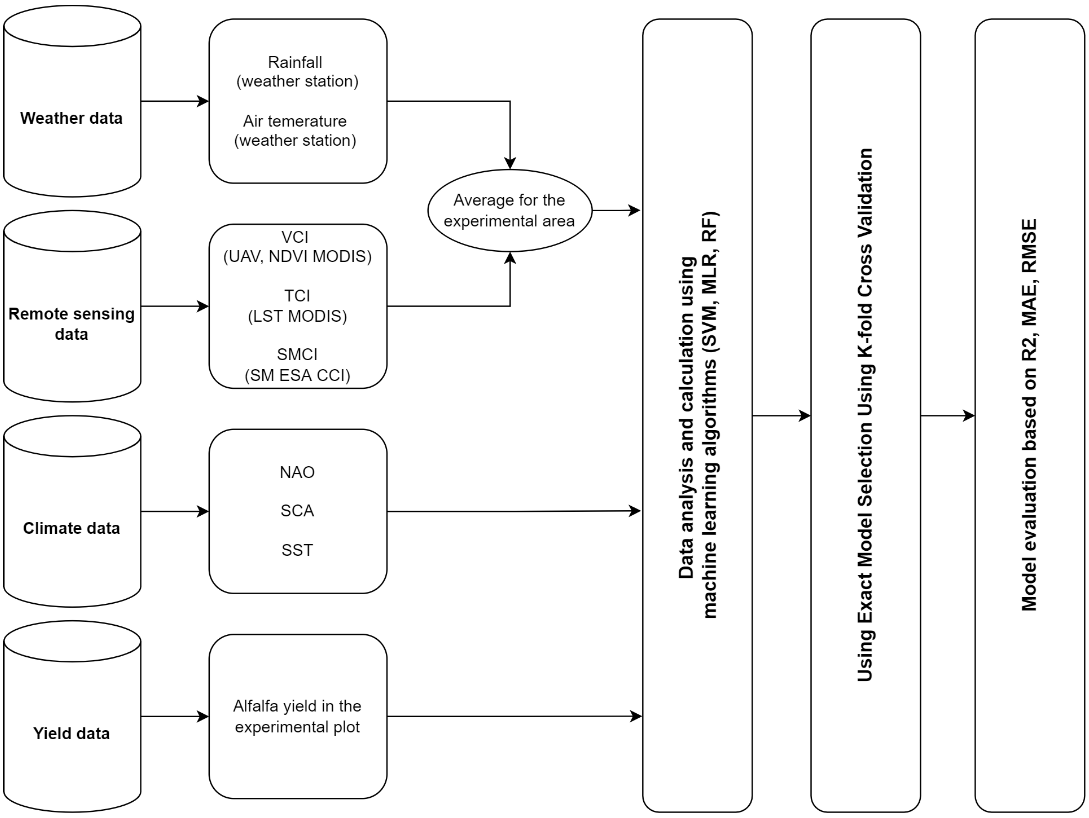

Figure 1: Schematic diagram presenting an overview of the main inputs data and the methodology proposed in this study

Widely used statistical measures were used to evaluate the performance of the developed models in this study (Feng et al., 2020). The coefficient of determination (R2) reflects the degree of a linear relationship between observed and predicted alfalfa yields Eq(1). The mean absolute error (MAE) indicates the percentage of the mean deviation of the predicted yield from the observation Eq(2). The root means square error (RMSE) measures the discrepancy between the predicted yield and observations Eq(3).

$$R^2 = \frac{(\sum_{i=1}^{n}(O_i - \bar{O})\ (F_i - \bar{F}))^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2 \sum_{i=1}^{n}(F_i - \bar{F})^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|F_i - O_i| \tag{2}$$

$$RMSE = \frac{1}{n}\sqrt{\sum\nolimits_{i=1}^{n}(F_i - O_i)^2} \tag{3}$$

where O_i is the observed return, F_i is the predicted return using the machine learning algorithm, $\bar{O}$ and $\bar{F}$ are the average values of the observed and predicted returns, and n is the number of samples used for the machine learning model.

The parameters of the selected models were adjusted using NDVI composite indices for the period from 2017 to 2022. Sixteen-day composite images with a resolution of 250 m obtained from MODIS (TERRA satellite) as well as Sentinel-2 and Landsat-8 were used to prepare the initial data. Yield data for the specified time interval were obtained from the farm "Experimental farm of oilseed crops" (EFoOC), located in Eastern Kazakhstan. An adjustment was made for the alfalfa crop. The total area of the experimental plot was 327 ha and is shown in Figure 2.

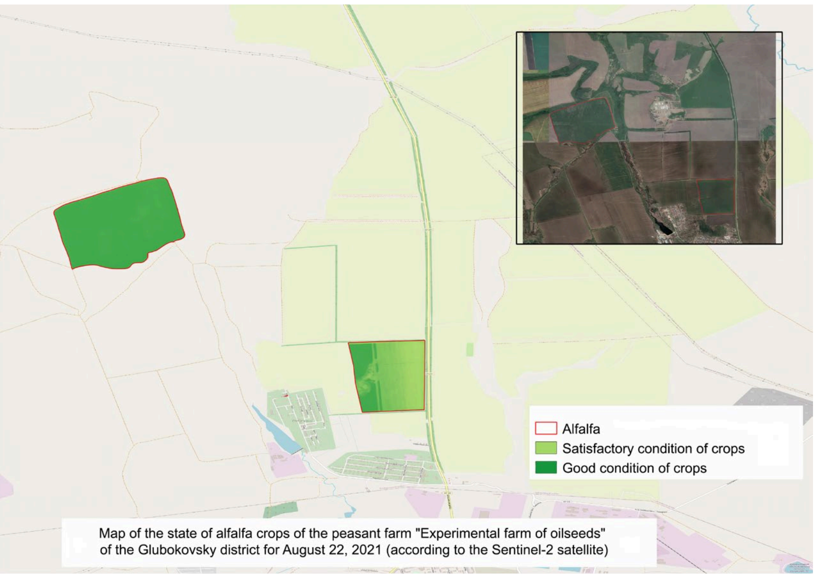
□ Alfalfa
🟩 Satisfactory condition of crops
🟩 Good condition of crops
Map of the state of alfalfa crops of the peasant farm "Experimental farm of oilseeds"
of the Glubokovsky district for August 22, 2021 (according to the Sentinel-2 satellite)

Figure 2: Study area and experimental field

Alfalfa is one of the most valuable and intensively grown fodder crops worldwide. The use of remotely sensed data (ERS) is the most technologically advanced and progressive method of vegetation monitoring, but it has a number of disadvantages compared to UAV imagery. Due to the high altitude of satellite images, the detail of the objects has limitations. The main problems connected with the space image use are related to the pixel resolution (30 m^2 per pixel for Landsat and 500 m^2 for MODIS) and circulation period (16 d for Landsat and 26 d for SPOT). This problem was solved when new satellites were introduced: WorldView 2-3 (DigitalGlobe, Longmont, Colorado, USA). WorldView-2 is the first commercial high-resolution satellite with eight spectral sensors ranging from visible to near-infrared. A feature of the satellite is that each sensor is narrowly focused on a specific range of the electromagnetic spectrum, which is sensitive to a particular feature of the earth or property of the atmosphere. However, images from this platform are very expensive. The re-visit time, which averages 16 d, also complicates agricultural applications, especially those related to water and nutrient management (Xue et al., 2017). On-board and/or unmanned platforms serve these two major challenges. Carrying out simultaneously with space monitoring and aerial surveys of land, followed by summarising the results of statistical processing of the dataset of climatic, agrochemical and others in the form of a mathematical model, seems an important and necessary task for crop yield forecasting.

3 Results and discussion

Analysis of seasonal dynamics of NDVI revealed that index values are higher in summer months (June, August) than in autumn months (October). This is explained by the seasonal dynamics of the vegetation index. The timing of the phases of development varies depending on the weather conditions of the year. As the phases of vegetative development change, the composition and content of pigments in the leaves of plants change, biomass increases and the amount of chlorophyll in the green leaves of plants increases. As chlorophyll accumulates, plant brightness decreases in the visible part of the spectrum, especially in the red zone, and increases in the infrared. Consequently, the NDVI value increases. Shortwave infrared (SWIR) measurements can estimate the amount of water in plants and soil because water absorbs SWIR wavelengths. Shortwave infrared bands (band - region of the electromagnetic spectrum; a satellite sensor can image the Earth in different bands) are also useful for distinguishing between cloud types (water clouds and ice clouds), snow and ice that appear white in visible light. In this composite image, vegetation appears in shades of green, soils and built-up areas have different shades of brown, and water appears black. The recently burned ground is strongly reflected in the SWIR bands, making them valuable for mapping fire damage. Each type of rock reflects shortwave infrared light differently, allowing you to map the geology by comparing the reflected SWIR light.

Aerial photography from a drone gives larger and more detailed data in high resolution (including digital images) and allows to work in any weather (except wind) and surveys up to 5,000 ha of crops. Visualisation of comparison of space imagery data with the results of aerial photos processing is presented in Figures 3 and 4.

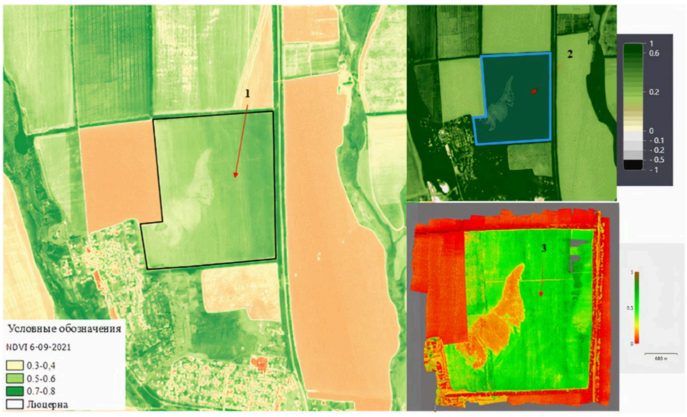

Figure 3: Calculation of NDVI index. Alfalfa crop: 1 - based on Sentinel satellite imagery, 2 - EO Browser web-application, 3 - based on aerial imagery from GEOScan 201 Agro UAV

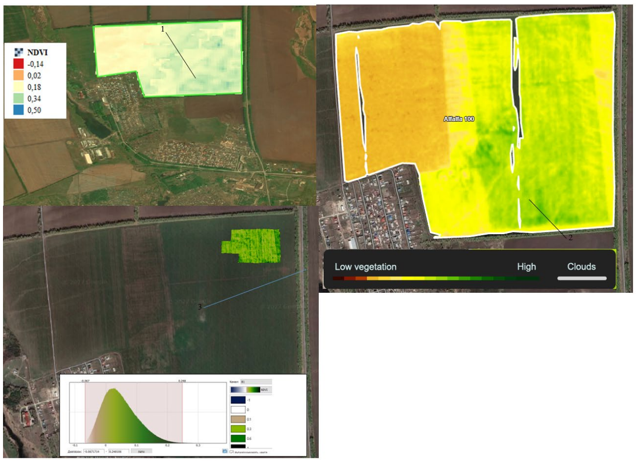

Figure 4: Calculation of NDVI index. Alfalfa crop: 1 - based on Landsat 8 satellite imagery, 2 - one soil web application, 3 - based on DJI Phantom 4 Multispectral aerial imagery

Figure 3 shows the calculation of the NDVI index of the experimental plot in the period from September 6-12, 2021. NDVI index values allow not only for comparing the performance of two levels of information acquisition but also for using accumulated values of NDVI indices as predictors in the yield forecasting model. Figure 4 shows the results of calculating the NDVI index of the experimental plot from 6 – 10 May 2022.

Three machine learning methods were developed (1 y before harvest) to determine the best combination of input data for modelling and forecasting alfalfa yield among satellite indices, weather data, and climate indices. To select the best hyperparameters of the machine learning algorithms, this study used complex grid search (GS) to explore all possible combinations of hyperparameters and also used cross-validation (CV) to evaluate the performance of the algorithms. In GS, a set of values was assigned to each hyperparameter, and a set of tests was generated by assembling all possible combinations of values. The evaluation was performed using k-fold cross-validation. CV is the most commonly used method for algorithm selection and evaluation because of its simplicity and ability to avoid overtraining. In k-fold cross-validation, the training data are randomly divided into k subsets, and the delay method is repeated k times so that each time one of the k subsets is used as the validation set of the model built using (k - 1) subsets.

Statistical measures for different combinations of input datasets and for different methods are presented in Table 1. Cross-validation is used to avoid over-training of the neural network. All data on which the model is built are divided into k blocks of equal size. Training is done on k-1 blocks, and testing is done on the k^{th} block. The procedure is repeated k times, and each time a different block is chosen for testing. As a result, all blocks turn out to be used both as training and testing blocks. This prevents and guarantees that the model is not be retrained in the future.

Table 1. Statistical performance of prediction models for several combinations of raw data and three machine learning methods (1 y before harvest)

Input data	Models	RMSE (t/ha)	MAE (t/ha)	R^2
Satellite drought indices only	SVM	0.58	0.41	0.75
	MLR	0.64	0.52	0.63
	RF	0.51	0.40	0.78
Hyperspectral indices from UAV and weather data	SVM	0.47	0.36	0.76
	MLR	0.38	0.41	0.86
	RF	0.32	0.30	0.89
Satellite drought indices, UAV data, weather data, and climate indices	SVM	0.33	0.25	0.87
	MLR	0.45	0.32	0.78
	RF	0.25	0.20	0.94

The results presented in Table 1 show that the satellite drought indices, the UAV data, weather data, and climate indices show better results than the other data. The results showed that the statistical performance of the model improves as the number of data sets used for prediction increases. All statistical performance improves with the addition of datasets for all methods tested. Results showed that yield variability correlated with satellite drought index values R^2 ranging from 0.63 (for MLR) to 0.78 (for RF) and RMSE from 0.58/ha (for MLR) to 0.51/ha (for RF).

By combining satellite drought indices and weather data, the performance of all models improves by 3 – 8 % for R^2 and 25 – 32 % for RMSE. The best statistical performance is obtained by combining the three data sets with a further statistical improvement of about 14 – 43 % for RMSE and 3 – 9 % for R^2, depending on the method used. This means that climate indices such as the NAO, SCA, and SST models, as well as the use of UAV data, contribute to improved model performance. In addition, nonlinear machine learning (RF, SVM) approaches outperformed linear approaches (MLR) when comparing different methods. This indicates that most of the relationships between returns and the predictors in question are nonlinear, and those nonlinear methods are obviously better at capturing these relationships than linear methods. Finally, the best yield prediction algorithm in our study is RF, which predicts yield with R^2 = 0.94 and RMSE = 0.25 t/ha. This result was confirmed by several studies on seasonal yield prediction, which showed better performance of the RF method compared to other nonlinear machine learning approaches such as SVM and MLR. This is consistent with the results of Adam et al. (2021), where the authors state that regression models using a coefficient of determination (R^2), standard error of estimate (SEE), and root mean square error (RMSE) will allow farmers to properly develop soil management plans and prevent acidification problems when combined with other soil property data.

4. Conclusion

When testing the existing models, shortcomings in predicting various cereals, legumes, and oilseeds were identified. Alfalfa was selected to adapt the existing algorithms since it is harvested three times during one calendar year, so it is possible to make adjustments when there is a discrepancy in the yield forecast. Yield forecasting provides critical and timely information that allows farmers to make quick decisions to improve yields by improving farming practices during the growing season. The main objective of this study was to develop an approach to forage crop yield forecasting in East Kazakhstan based on data from multiple sources and machine learning techniques. To this end, this study presents a methodology based on different machine learning approaches (MLR, SVM, RF) to predict alfalfa yield in the year before harvest using freely available datasets, including satellite drought indices, weather data, and climate indices. The results show that the combination of satellite drought indices, weather, and climate data as predictors of forage crop yields provides higher predictive accuracy than using any single data source. The results revealed that the RF method is superior to other machine learning methods. The RF method predicts yield with R^2 = 0.94 and RMSE = 0.25 t/ha. which is one of the closest to the actual data. In addition, the proposed approach provides a source of timely information for decision-making during the growing season. This work can be used to map yield gains and analyse yield gaps nationwide. The identified hotspot areas in terms of yield gaps are suggested for practice improvement and further research work.

Acknowledgements

This research has been supported by Project BR10865102 "Development of technologies for remote sensing of the earth (RSE) to improve agricultural management", funded by the Ministry of Agriculture of the Republic of Kazakhstan

References

Adam M., Ibrahim I., Sulieman M., Zeraatpisheh M., Mishra G., Brevik E. C. 2021, Predicting Soil Cation Exchange Capacity in Entisols with Divergent Textural Classes, The Case of Northern Sudan Soils. Air, Soil and Water Research, 14, 11786221211042381.

Beisekenov N.A., Sadenova M.A., Varbanov P.S., 2021. Mathematical Optimization as A Tool for the Development of "Smart" Agriculture in Kazakhstan, Chemical Engineering Transactions, 88, 1219-1224.

Belgiu M., Drăguţ L. 2016, Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.

Bouras E.h., Jarlan L., Er-Raki, S., Balaghi, R., Amazirh A., Richard B., Khabba S. 2021, Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco. Remote Sens. 13, 3101. doi: 10.3390/rs13163101.

Cihlar J., Laurent L. S., Dyer J. A. 1991, Relation between the normalized difference vegetation index and ecological variables. Remote sensing of Environment, 35(2-3), 279-298.

Eberly, L. E. 2007, Multiple linear regression. Topics in Biostatistics, 165-187.

Feng L., Zhang Z., Ma Y., Du Q., Williams P., Drewry J., Luck B. 2020. Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning. Remote Sensing, 12(12), 2028.

Guo X., Shukla M. K., Wu D., Chen S., Li D., Du T. 2021. Plant density, irrigation and nitrogen management: three major practices in closing yield gaps for agricultural sustainability in North-West China. Frontiers of Agricultural Science and Engineering, 8(4), 525-544.

Habarov D.A., Adiev T.S., Popova O.O., Chugunov V.A., Kozhevnikov V.A. 2019, Analysis of modern technologies for remote sensing of the Earth, Moskovskij ekonomicheskij zhurnal. 181-190. doi 10.24411/2413-046X-2019-11068

Nihar A., Patel N. R., Pokhariyal S., Danodia A. 2022, Sugarcane Crop Type Discrimination and Area Mapping at Field Scale Using Sentinel Images and Machine Learning Methods, Journal of the Indian Society of Remote Sensing, 1-9.

Xue J., Su B. 2017, Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications, Journal of Sensors, 2017, ID 1353691, doi: 10.1155/2017/1353691

Yadav K., Geli H. M. 2021. Prediction of Crop Yield for New Mexico Based on Climate and Remote Sensing Data for the 1920–2019 Period. Land, 10(12), 1389.

Yang P., Zhao Q., Cai X. 2020. Machine learning based estimation of land productivity in the contiguous US using biophysical predictors. Environmental Research Letters, 15(7), 074013.