CHEMICAL ENGINEERING TRANSACTIONS
VOL. 88, 2021
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

ISBN 978-88-95608-86-0; **ISSN** 2283-9216
DOI: 10.3303/CET2188203

Mathematical Optimization as A Tool for the Development of "Smart" Agriculture in Kazakhstan

Nail Alikuly Beisekenov[a,*], Marzhan Anuarbekovna Sadenova[a], Petar Sabev Varbanov[b]

[a]Priority Department Centre «Veritas» D. Serikbayev East Kazakhstan technical university, 19 Serikbayev str. 070000, Ust-Kamenogorsk, Kazakhstan
[b]Sustainable process integration laboratory, researcher NETME CENTRE Faculty of Mechanical Engineering, Brno University of Technology, Technická 2896/2,616 00 Brno, Czech Republic
bnail@ektu.kz

This article uses methods for predicting plant performance indicators in Kazakhstan. In the work, using deep learning, visualization of predicted indicators (indicators and others), statistics from predicted values and identified changes, time series have been developed. Sentinel satellite data and statistical indicators for the last few years for the agricultural territories of Kazakhstan are used as primary data. It is found that the upward trend in wheat quality, however, increases the size of fertilizers, variables based on the NDVI also significantly contribute to the forecasting model. It has been shown that the amount of applied fertilizer has stabilized in the past few years due to economic and environmental constraints, so NDVI-based models will become increasingly important for enhancing forecasting models. Four machine learning algorithms have been evaluated and compared, namely boosted regression trees (BRT) and support vector machine (SVM), to map and predict the field yield of the Experimental Oil Farm in East Kazakhstan using readily available additional data. Based on the results of the work, a forecast of crop yields and general statistical recommendations for increasing yields were obtained.

1. Introduction

Linear Programming (LP) is successfully used in many sectors of the national economy. Particularly effective is the application of LP in smart farming. One of the problems of Linear programming is the limitations, and the impossibility of structuring the desired system. For this reason, the concept of "smart" agriculture was applied - this is a concept of agricultural activities based on the introduction of new technologies: IoT, sensors, UAVs, GPS and GLONASS, automation systems, etc., in the process of obtaining agricultural products, in order to increasing yields and reducing costs of expenses. The use of mathematical optimization for "smart" agriculture in Kazakhstan is a rapidly developing and promising area. An important advantage of satellite imagery is the efficiency of obtaining information on the spatial distribution of the used arable land, as well as the objectivity and independence of the information received. In recent decades, there has been an upward trend in crop yields around the world, mainly due to technological advances and increased use of fertilizers. (Mueller et al., 2012). Deviations from the trend are mainly related to weather conditions (Nagamani and Mariappan, 2017). To predict these deviations, agrometeorological models have been developed based mainly on two variables: air temperature and precipitation. This method features its mathematical model in the use of soil and vegetation information, as well as GIS interpolation tools to create spatial surfaces of variable environments and climate based on point observations (Rasmussen, 1998a). Adding environmental information and using multiple regression techniques improved the yield model further (Rasmussen, 1998b). The best combination of planting criterion and maize varieties analysis was then achieved by optimizing planting dates and maize varieties in the decision support system for agrotechnology transfer (DSSAT) environment (Mugiyo et al., 2021). Space observations of the Earth have gained importance for monitoring agricultural crops since the 1990s. Vegetation indices from remote sensing have shown a great potential as explanatory variables in yield models. The most commonly used vegetation index is the Normalized Difference Vegetation Index (NDVI) (Gautama et al., 2015).

Paper Received: 16 May 2021; Revised: 20 July 2021; Accepted: 9 October 2021
Please cite this article as: 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 DOI:10.3303/CET2188203

The production of grain and other crops in Kazakhstan is becoming more difficult, mainly due to the lack of water, while grain consumption continues to grow. To cope with unavoidable imports, forecasting production is becoming increasingly important, preferably as early as possible during the growing season. As elsewhere in the region, the long-term upward trend in winter wheat yields have been accompanied by increased fertilization. To explain the off-season deviation from this trend, models are constructed using remote sensing vegetation indices such as the NDVI. Since these indices usually do not show a clear linear relationship with yield, traditional Multiple Linear Regression (MLR) modeling will reach its limits. When a limited number of training samples are available (in this study, 16 observations (years) per prefecture) MLR will not be able to provide a good fit (Mountrakis and Ogole, 2011). Machine learning techniques such as Support Vector Machines or boosted regression trees perform better with a large set of predictors when only a few observations are available (Heremans at el., 2015). They rely on teaching examples to learn the empirical relationship between input variables and output (returns) without any statistical assumptions. Based on a literature review and previous experience of researchers, it was decided to develop an intelligent geographic information system for farmers. This article describes the application of two machine learning methods to simulate winter wheat yields in the Experimental Oilseed Farm (EOF) using predictors based on NDVIs, also called functions, with the help of which it turned out to improve the variance of the yield forecasts, which gave us 91% accuracy in yield forecast based on multivariate modeling. In addition to comparing the performance of the two methods and an improved machine learning yield prediction model, this article also aims to evaluate the use of accumulated NDVI values as predictors. The novelty and purpose of this work is the creation and integration of a geographic information system that predicts a number of factors that determine the effectiveness of crop production in Kazakhstan using two machine learning methods. also includes field management tools, recommendations based on multivariate data analysis (meteorological conditions, agrochemical analysis, NDVI vegetation data, etc.). To solve this problem and optimize data analysis, a multilayer neural network architecture was created on the open software library for machine learning Tensorflow and Keras, which allows predicting and determining the type and characteristics of plant growth changes in a specific research area.

2. Data and Methods

Multiple datasets, such as satellite imagery, agrochemical analyzes and vegetation index, each have a set of quantified properties that play a key role in soil classification. However, they are difficult to predict and few attempts have been made to map their spatial distribution. Four machine learning algorithms have been evaluated and compared, namely boosted regression trees (BRT) and support vector machine (SVM), to map and predict the field yield of the Experimental Oil Farm in East Kazakhstan using readily available additional data. Resampling mechanisms and ensemble methods significantly improved prediction accuracy (measured based on the area under the receiver operator characteristic curve (AUC)) and gave more stable results for BRT (AUC 0.921 ± 0.012, mean ± standard deviation) which is the most commonly used machine learning method. Although the SVM algorithm gave a comparable AUC value (0.906 ± 0.006) with the RF and BRT algorithms, it is sensitive to parameter settings that are extremely time consuming.

SVM is a machine learning method based on statistical learning theory. The SVM uses kernel functions to project data into a multidimensional space where partitioning is performed. BRT - Combines the benefits of two algorithms (i.e., Regression Trees and Boost Trees) to improve the performance of a single model. Boosting is a numerical optimization algorithm that minimizes the loss function by adding a new tree to the first regression tree model at each stage. A BRT model was developed using the "gbm" package in R.T. The following three estimated indices were also calculated: coefficient of determination (R2), root mean square error (RMSE) and mean absolute error (MAE) (equations (1) - (3)).

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \overline{O_i})^2}{\sum_{i=1}^{n}(O_i - \overline{O_i})^2} \tag{3}$$

where n represents the number of samples; Pi and Oi represent the predicted and observed values at site i, respectively. For every soil property, each model was run a hundred times and their average was used as the final prediction. The standard deviation (SD) of each cell in the raster was calculated from the generated 100 soil maps and the spatial variation of these SDs was used to represent the forecast uncertainty.

In the linear regression context, subsetting means choosing a subset from available variables to include in the model, thus reducing its dimensionality. Shrinkage, on the other hand, means reducing the size of the coefficient estimates (shrinking them towards zero). Note that if a coefficient gets shrunk to exactly zero, the corresponding variable drops out of the model. Consequently, such a case can also be seen as a kind of subsetting. Shrinkage and selection aim at improving upon the simple linear regression. There are two main reasons why improvements may be needed, the accuracy of the forecasts and the interpretability of the model.

Primary data are data from the Sentinel satellite, the characteristics of the observed surface (height, rotation, tilt), the position of the satellite device and the sun, surface temperature, humidity, precipitation, cloudiness, soil types and statistics for the last 5-10 years for selected agricultural areas. destination in Kazakhstan. To monitor the state of grain crops in East Kazakhstan, information from the Sentinel-2 satellite (MSI Sentinel-2 scanner; coverage area 2,330 $(km)^2$, spatial resolution 250 m, in two spectral ranges - 0.62-0.67 μm (in the red regions of the spectrum) and 0.84-0.87 μm (in the infrared range) and French satellites SPOT-2 and SPOT-4 (spatial resolution 20 m, bandwidth 60 km). Data from satellites are promptly received and processed 2-3 times a day at the East Kazakhstan Technical University (VKTU) named after D. Serikbayev, located in the city of Ust-Kamenogorsk. The NDVI was used as a quantitative characteristic of the state of crops. It was assumed that at a certain point in the image, normalized difference vegetation index is the ratio of the difference between the intensities of the reflected light in the infrared and red ranges of the spectrum to their sum. It is known that in the red region of the spectrum there is the maximum absorption of solar radiation by chlorophyll, and in the infrared region of the spectrum - the maximum reflection by the cellular structures of the leaf (Ustin and Middleton, 2021). Statistical data on the yield of winter wheat (T ha-1) and the application of chemical fertilizers (T) are kept in the East Kazakhstan region. For this study, data were available from 2010 to 2021. In the same period, Sentinel-2 satellites received NDVI composites with a spatial resolution of 1 km. The time series were additionally smoothed per pixel. This procedure identifies cloud-affected measurements and replaces them with interpolated values. Each month of the winter wheat season from January to December is coded with the capital letter of the given month (Example: October - O, December - D, January - J, February - F, March - M, April - A, May - Y, June - U) for followed by a number denoting a decade of days to shorten the month names, and the variables were entered into the systems in the same way. A month consists of three decades, the first lasts from 1 to 10 days, the second from 11 to 20 days, and the third consists of the remaining days of the month.

The predictors in this study consisted primarily of NDVI variables, but one non-NDVI variable was included, chemical fertilization (CFI). Three types of NDVI variables are evaluated. First, there are Single NDVI variables corresponding to the value for one decade. There are 25 of them in the winter wheat season, from sowing in the last decade of October (O3) to harvesting in the last decade of June (U3). In addition, there are two types of cumulative NDVI variables: incremental and target NDVI, where individual values are accumulated over a specified number of decades. Incremental NDVI corresponds to the sum of NDVI values from the first decade (O3) to each decade during the season. For the target NDVI, specific periods that are expected to be phenologically significant were selected by calculating the absolute slope changes for the cumulative NDVI time series. Decades with the greatest variation were selected for each year in each prefecture Ultimately, the decades that were repeatedly selected, were used as start and/or end points for specific NDVI periods. The selected decades were: M1, M3, A1, A3, Y2, Y3, U1 and U3. These decades can be combined to span 28 periods, giving rise to the same number of Targeted NDVI variables. Thus, for each prefecture, there is an available dataset of 78 input variables (CFI, 25 Single NDVI values, 24 Incremental NDVI values and 28 Targeted NDVI values), covering 16 y of observations.

2.1 Machine learning

Machine learning techniques are so-called data-driven techniques and they build a model based on evidence from a set of sampled data (input, output). The learning phase results in a function that can be applied to new inputs to predict the corresponding outputs. Algorithms can detect complex patterns by combining simple components. (Heremans et al., 2015) reports that ensemble tree methods, a subset of machine learning, have great potential for predicting yield. Boosted regression trees (BRT) is part of the subgroup of ensemble tree methods. Its modeling approach is based on splitting a complex decision into several simple binary decisions.

2.2 Boosted regression trees

BRT models were tuned based on the methods described earlier. They need to specify two meta-parameters: the number of separation levels (interaction depth) and the compression ratio, which affects the convergence rate. Since the dataset only contained 16 observations, it was decided not to use a separate validation dataset,

but to use a k-fold cross validation. Thus, 16-fold cross-validation was applied to optimize meta-parameters (two per model) in feature selection. The following meta-parameter values have been tested. The number of division levels in each individual tree: 1; 2; 3. Shrinkage ratio: 0.01; 0.05; 0.10; 0.25; 0.50.

2.3 Support Vector Machines

When the statistical learning theory of SVM is applied to continuous outputs, this approach is often referred to as support vector regression. The model input is mapped into a multidimensional feature space using a kernel function, and then a model is built in this new function space. linear regression SVM models for each prefecture were created using the 'caret' package with 'R' version 3.1.2. The linear kernel was chosen as the meta-parameter.

For each prefecture, two BRT models and two SVM models were trained. The first was based on all the features selected as relevant in the previous section, including CFI. The second model was based solely on the selected characteristics of the NDVI.

3. Results and discussion

Observations of the dynamics of the development of agricultural crops according to remote sensing data showed that in the space of spectral characteristics, each type of crop at a certain time and at a certain stage of development forms a compact cluster (a set of homogeneous photometric points). The CFI predictor was of great importance in predicting yield. BRT models have chosen this feature in all prefectures. Within the SVM models, CFI was relevant for 15 prefectures. Based on the trends of increasing yields and CFI over the years, also in Experimental Oilseed Farm (EOF). Obtaining derived images from satellite data by processing according to special algorithms in selected spectral regions makes it possible to study plant productivity, biomass and the intensity of photosynthesis. Figure 1 shows the statistics of changes in the NDVI indicator of agricultural crops at the experimental site in the period from 2017 to 2021, which shows that, depending on various conditions, the state of the soil, crops and other crop production indicators differ markedly.

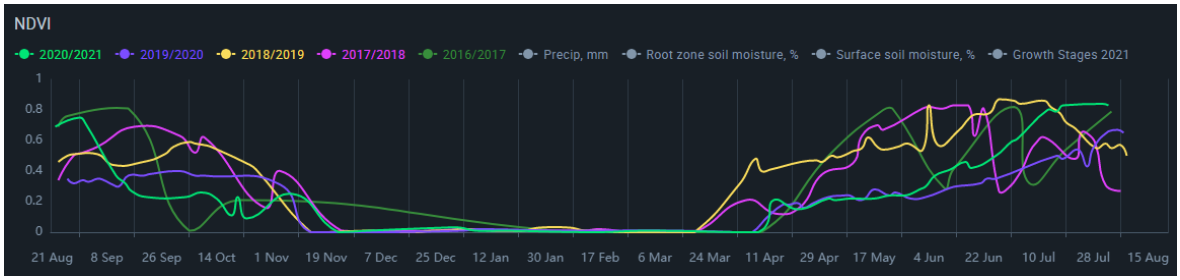

Figure 1 Statistics of changes in the NDVI indicator of agricultural crops at the experimental site in the period from 2017 to 2021.

Feature relevance patterns differ between prefectures as well as between the two machine learning methods applied, as shown in Figure. 2. The Figure consists of three different parts, each of which summarizes the different types of NDVI variables. Table 1. compares the selection rate (number of times of selection) for each type of NDVI variable by method (BRT or SVM). The total number of times a feature was marked as relevant is the same for both methods, BRT 69, SVM 65. 71 % of relevant BRT features belonged to the NDVI single variable group. This group has the lowest proportion of relevant samples for SVM, for which more suitable functions were found for the two types of cumulative NDVI variables. The largest share of relevant features for SVM (43 %) belonged to the NDVI target variables group. For BRT there were only four cases when a function from the incremental NDVI variable group was relevant, for SVM - nineteen cases. In Figure. 2a, which shows the results for a single NDVI, a period of high significance can be distinguished, namely April-May. Looking at Figure. 2c, which shows the results for the target NDVI, shows that A3 and Y2 (the target cumulative NDVI from late April to May) is also an important feature. The feature will be included in the final models if it is selected as valid in at least 5 of the 17 prefectures. This threshold is indicated by the dotted line in Figure. 2. Based on this, O3 and Y2 are the only Single NDVI features that should be included in the final models when using BRT (Figure. 2a). The increase in yield is mainly attributable to CFI over the years, however some variation can be explained by NDVI variables. NDVI variables selected for the final step of the model, belonging to the same group of Single NDVI numbers, or to target NDVI variables. This confirms that when using cumulative NDVIs, it is better to sum the values over a period compared to using incremental NDVIs for each possible period. When reviewing the results for the incremental NDVI (Figure. 3), it becomes clear that the actual dates for performing

functions that were more than once (O3Y2, O3Y3, O3U1, O3U2, O3U3) basically correspond to the decades chosen as boundaries for the target periods. (Y2, Y3, U1 and U3). It is clear from both Single NDVI and Target NDVI that the period from March to June has the greatest impact on winter wheat yields. Phenologically, this period includes docking, earing, and maturation. RMSE cross validation was examined to compare the performance of BRT and SVM. The results are visualized as rectangular diagrams in Figure 3. The RMSE is expressed in the same units as the model outputs, ie. harvest (T ha-1). The RMSE values for BRT are consistently less than those for SVMs. Models that include CFI have a lower RMSE than models based solely on NDVI. The machine learning methods used in this study, mainly BRT, is included in the geographic information system of smart agriculture.

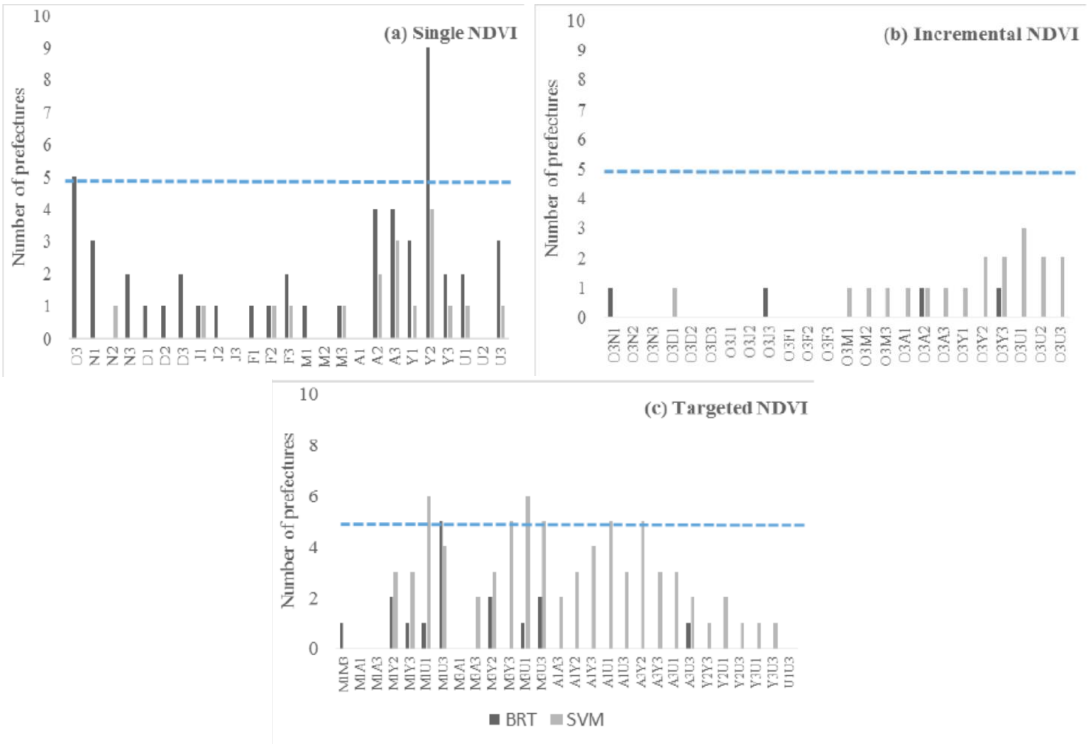

Figure 2: These three graphs show in how many prefectures the projected NDVI was up to date. The dotted line indicates the threshold of five that was used to select the functions for the final simulation

Table 1: Comparison of the importance of different types of NDVI by method

	BRT	SVM
Single NDVI	49 (71 %)	18 (28 %)
Incremental NDVI	4 (6 %)	19 (29 %)
Targeted NDVI	16 (23 %)	28 (43 %)

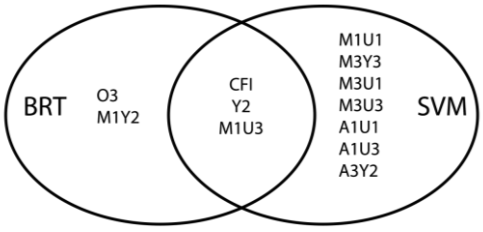

Figure 3: Venn diagram of selected functions

4. Conclusions

The created system is self-learning and has the function of analyzing the economic efficiency of agricultural production by the farmer based on the data sets described above. Both machine learning methods had low rms, but BRT seems to outperform SVM for predictions of prefectural yields at EOF. Previous studies have produced conflicting results; sometimes BRT is superior to SVM, while other studies show that SVM is more efficient than BRT. SVM is better suited for multidimensional processing. Models including CFI have outperformed models based solely on NDVI. This was to be expected, since the constant intensification of fertilizer use led to an increase in yields from year to year. Although this trend hides the predictive power of NDVI functions, the difference in RMSE between models with and without CFI is very small (0.09 for BRT and 0.07 for SVM). For the region and the EOF, fertilizer production growth is expected to stabilize from year to year due to economic and environmental constraints, so NDVI-based models will become increasingly important. Models must include both single NDVI variables and target NDVI variables. For the latter, more research on critical periods during the growing season is recommended.

Acknowledgements

This research has been supported by the Project IRN 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 and the EU project "Sustainable Process Integration Laboratory - SPIL", project No. CZ.02.1.01/0.0/0.0/15_003/0000456 funded by EU as "CZ Operational Programme Research, Development and Education", Priority 1: Strengthening capacity for quality research under a collaboration agreement with D. Serikbayev East Kazakhstan Technical University.

References

Gautama V.K., Gaurava P.K., Murugana P., Annaduraia M., 2015. Assessment of Surface Water Dynamics in Bangalore Using WRI, NDWI, MNDWI, Supervised Classification and K-T Transformation. Aquatic Procedia. 4, 739-746.

Heremans S., Dong Q., Zhang B., Bydekerke L., Van Orshoven J., 2015. Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data. Journal of Applied Remote Sensing, 9(1), 097095.

Mountrakis G., Im J., Ogole C., 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247 –259.

Mueller N.D., Gerber J.S., Johnston M., Ray D.K., Ramankutty N., Foley J.A., 2012. Closing yield gaps through nutrient and water management. Nature, 490, 254-257.

Mugiyo H., Mhizha T., Chimonyo V.G.P., Mabhaudhi T., 2021. Investigation of the optimum planting dates for maize varieties using a hybrid approach: A case of Hwedza, Zimbabwe. Heliyon, 7(2), e06109.

Nagamani K., Mariappan V., 2017. Remote Sensing, GIS and Crop Simulation Models - A Review. International Journal of Current Research in Biosciences and Plant Biology, 4(8), 80-92.

Rasmussen M.S., 1998a. Developing simple, operational, consistent NDVI-vegetation models by applying environmental and climatic information: Part I. Assessment of net primary production, International Journal of Remote Sensing, 19(1), 97-117.

Rasmussen M.S., 1998b. Developing simple, operational, consistent NDVI-vegetation models by applying environmental and climatic information. Part II: Crop yield assessment, International Journal of Remote Sensing, 19(1), 119-139.

Ustin S.L., Middleton E.M., 2021. Current and near-term advances in Earth observation for ecological applications. Ecological Processes, 10, 1.