Machine learning prediction of maize yield using spatio-temporal data

A. Nyéki[1*], C. Kerepesi[2*], B. Daróczy[2], A. Benczúr[2], G. Milics[1], A.J. Kovács[1] M. Neményi[1],
[1]*Széchenyi István University, Faculty of Agricultural and Food Sciences, Department of Biosystems and Food Engineering, Mosonmagyaróvár, Vár square 2. 9200, Hungary*
[2]*Institute for Computer Science and Control, Hungarian Academy of Sciences, H-1111 Budapest, Kende Str. 13-17., Hungary*
corresponding author: nyeki.aniko@sze.hu
**Joint first authors*

Abstract

The accuracy of the yield prediction is crucial for the crop management. Site-specific research projects started in 2001 in order to investigate the site-specific technologies (VRA, VRS). Among others, soil properties, meteorological conditions, management characteristics, and yield are measured in the 15.3 ha research field, which is divided into 63 treatment units (management zones). The aim of this study is to predict maize yield using spatio-temporal training data. Counter-propagation artificial neural networks (CP-ANNs), XY-fused networks (XY-Fs), supervised Kohonen networks (SKNs), extreme gradient boosting (XGBoost) and support-vector machine (SVM) were used for predicting maize yield in 5 vegetation periods. Input variables for modeling were: soil parameters (pH, P_2O_5, K_2O, Zn, Clay content, *ECa*, draught force, Cone index), and microrelief averages for the 63 treatment units. For impact of meteorological parameters, precipitation, minimum and maximum temperature, global radiation, vaporization and aridity index were used. The best performed method (XGBoost) reached 92.1% and 95.3% of accuracy on the training and the test set.

Keywords: maize yield prediction, machine learning, XGBoost deep learning, soil variables

Introduction

Recently, several studies have drawn attention to the need for a paradigm shift. According to Longchamps et al. (2018), the current unfavourable effects of agriculture on the biosphere cannot be reduced by the knowledge of traditional experiments. Paradigm shift is also needed because the perception of the harmful phenomenon and the reaction time should be reduced. There has to be a change in scientific methods, and the potential of big data needs to be better exploited than before, resulting in artificial intelligence (in the following "AI"). The gap between environmental adverse impacts and our slowly expanding knowledge is growing, the imaginary scissors are continuously open, we cannot handle the challenges.
Nyéki et al. (2013, 2017) concluded that the increasing of the input data of plant physiological models can increase of accuracy of prediction to only a limited extent.
AI is used by more and more researchers to model a wide range of tasks in agriculture. There are large numbers of studies in which the relationship between soil properties and yield is analyzed (Irmak et al., 2006; Miao et al., 2006; Mike-Hegedűs, 2006; Dahikar

and Rode, 2014; Pantazi et al., 2016). A growing number of articles are presented in which climate effects are analyzed by AI (Elgaali and Garcia, 2004; Iizumi et al., 2018). More and more work is being done in which AI is used to test long term (20 to 30 years) observations (Chlingaryan et al., 2018; Folberth et al., 2019).
The gradient boosting modeling (GB) seems to be a promising method. Rice yield was forecasted by XGBoost prediction model from weather information, the cultivation data and the location information (Maeda et al., 2018). The weather data contains the daily maximum temperature, the daily minimum temperature and sunshine time. The crop management data includes: year of rice cultivation, seedling type, sowing style, fertilization level, date of fertilization and seeding, planting density…etc. Location characteristics are the latitude and longitude of each test site. It was concluded that the best production accuracy was observed at two integral intervals: one is from planting date to heading date and the other is from heading date to ripening date. Its prediction accuracy was 74.4 %.
Laacouri et al. (2018) assessed the corn N status comparing machine learning and vegetation indices. Hyperspectral images were collected by hexacopter UAV platform. Eight machine learning algorithms were compared for their accuracy, among other gradient boosting models. The authors concluded that the hyperspectral imagery combined with ML improved the assessment of corn N stress status at V5 growth stage, and achieved more than 90 % classification accuracy when the entire spectrum was mined. GB achieved an overall classification accuracy of 89%. SVM, logistic regression (LR), Multi-Layer Perceptron (MLP) and GB showed promising results.
Folberth et al. (2019) explored two machine learning approaches: extreme gradient boosting and random forests. ML was trained on global scale maize simulation of a GGCM (Global Gridded Crop Model) and exemplary applied to the extent of Mexico. The method provides very high accuracy (R^2 >0.96) for predictions of maize yields, hydrologic externalities evapotranspiration and crop available water.
Khanal et al. (2018) integrated the high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. High spatial resolution (<1m) was used for the capturing of soil properties (SOM, CEC, K, Mg, and pH) and corn yield. The accuracy of seven statistical methods (LM – linear regression, RF – random forest, SVM with linear and radial kernel function, GB, NN – neural network, and CUB - cubist) was compared for their ability to predict soil properties and corn yield. For pH and corn yield prediction Gradient Boosting Model and RF model performed better than other models.

Materials and methods

Study site
Field experiments were carried out in the 23.52 ha experimental field (Fig.1) belonging to Széchenyi István University in the vicinity of Mosonmagyaróvár, Hungary [N47°54’20.00”; E17°15’10.00”]. The experimental field is an alluvial plain of the Leitha River - on which precision agriculture has been applied since 2001.

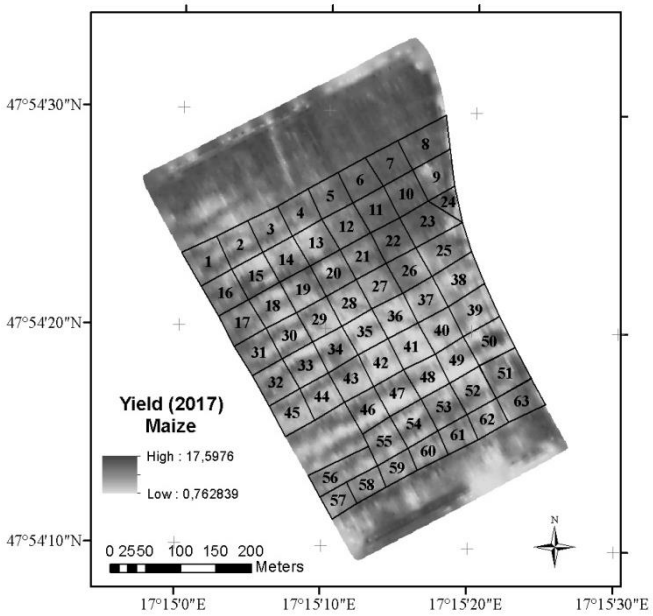

Figure 1. Site-specific maize yield (2017) in study site with 63 treatment units.

Crop and soil parameters affecting maize yield

The NDVI was calculated from satellite data by Landsat on July 26th, 2006; on July 5th; 2010; on July 29th, 2013 and on August 9th, 2017. The images were collected under high quality atmospheric conditions during the maize flowering period.

The soil textures according to USDA indicated the soil types of loam, silty loam and sandy loam.

Instrumentation

In order to find the correct position, a Trimble AgGPS system was used. The relative elevation (m) was determined by differential global positioning system (DGPS, Sokkia Radian-IS). Yield measurement was carried out by AGROCOM yield measurement system mounted on a Deutz Fahr grain harvester until 2010. From 2011, an AgLeader yield measurement system was used which was mounted on a CLAAS Medion 340 grain harvester. Continuous soil draft measurements were carried out by a self-developed system (Neményi et al., 2006). Site-specific penetrometer measurement was carried out by means of 3T system (Szőllősi, 2003). The soil EC_a was measured by a Veris Soil EC-3100 (Salina, KS, USA) instrument.

Feature table for supervised machine learning

Before supervised machine learning, a feature table was created from the measured parameters. The feature table consists of 315 (63*5) rows for each treatment unit and year, and also consists of 48 columns for each parameter (11 of soil parameters, 35 of meteorology parameters and one crop parameter) and the maize yield. In summary, four types of parameters were used: 1. Yearly changing soil parameters: measured in 5 years: pH, P_2O_5, K_2O, Zn; measured in 3 years: EC_a in two layers (Veris N3 and N4); measured in 1 year: Cone Index (MPa), pH (H_2O), draught force (kN). 2. Crop parameter (NDVI): measured in 4 years. 3. Non-yearly-changing soil parameters: measured in one year but used in all years: clay content (%), relative elevation (m). 4. 35 of meteorology features (not changing by treatment units): used five types in each

month from April (IV.) to October (X.), for maize vegetation seasons: sum of precipitation, average temperature, average relative precipitation, vaporization, aridity index. /Missing data were left blank./

Cross-validation and train-test split

Before training, we split the feature table into two parts, a training set of 50 and a testing set of 13 randomly selected treatment units. Since we use five years of measurements, we obtained 250 training and 65 testing samples. We selected the best parameters of the machine learning models by 5-fold cross-validation over the training set. The final models were trained on the entire training set and tested on the independent testing set.

Binary classification

Maize yield values were varied between 2.46 t/ha and 15.05 t/ha, and the upper limit of the first third of the lowest values of yield was 8.42 t/ha. The maize yield values were divided into two classes: low yield (<= 8.42 t/ha) and medium-high yield (> 8.42 t/ha). Several types of state-of-the-art supervised machine learning methods were used for classification: counter-propagation artificial neural networks (CP-ANNs), XY-fused networks (XY-Fs), supervised Kohonen networks (SKNs) extreme gradient boosting (XGBoost) and support-vector machine (SVM). CP-ANNs, SKNs and XY-Fs are supervised neural networks derived from hierarchical self-organizing maps (SOMs) (Ballabio, 2012), and used for wheat yield prediction in a recent study (Pantazi et al., 2016). XGBoost is an efficient implementation of the gradient boosting method (Friedman, 2001). XGBoost was applied successfully in a wide variety of classification problems XGBoost (Chen, 2016).

Evaluation measures for binary classification

Sensitivity, specificity, accuracy and ROC AUC were calculated to evaluate the classification models (AUC is calculated as the area under the ROC curve). Sensitivity means the proportion of actual positives (low yield samples) that are correctly identified, and sensitivity means the proportion of actual negatives (middle-high yield samples) that are correctly identified. Accuracy is the proportion of the total number of samples that are correctly identified. ROC curve (Receiver Operating Characteristic Curve) is defined by the point pairs of true positive rates (sensitivity) and false positive rates (1-specificity) at different threshold settings.

Results and discussion

Three neural networks models (CP-ANN, XY-F, SKN) and a gradient boosting method (XGBoost) were trained on a feature table of 47 measurements of 63 treatment units in five years, altogether 315 samples. Recall that we used 50 randomly chosen treatment units for training and 13 for testing. The statistics of the training and testing datasets are shown in Table 1. The distribution of features that influence maize yield are shown in Fig. 2 and the yield distribution by year in Fig. 3.

Table 1. Summary statistics of the selected soil and crop parameters and the maize yield for the training and test datasets. Min, minimum; Max, maximum; Mean, mean

(average); SD, standard deviation; CV coefficient of variation (the ratio of the standard deviation to the mean).

	Training dataset (n = 250)					Test dataset (n = 65)				
	Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV
Maize yield (t/ha)	3.25	15	9.39	2.62	0.28	2.46	15.05	9.23	2.84	0.31
pH (H2O)	7.62	7.82	7.75	0.04	0.01	7.7	7.93	7.78	0.06	0.01
pH (KCl)	7.09	7.83	7.39	0.14	0.02	7.14	7.71	7.4	0.14	0.02
P2O5 (mg/kg)	123	415	234.24	55.67	0.24	151	380	250.95	57.84	0.23
K2O (mg/kg)	22.6	518	153.05	102.06	0.67	60.1	400	148.6	91.4	0.62
Zn (mg/kg)	1.21	4.7	2.91	0.66	0.23	1.74	4.1	2.95	0.57	0.19
Clay content %	8.4	21	13.28	3.07	0.23	7.9	18.4	13.14	3.45	0.26
Draught force (kN)	1.52	4.72	3.1	0.72	0.23	1.9	6.12	3.4	1.31	0.38
Relative elevation (m)	122.18	123.6	122.74	0.34	0	121.96	123.11	122.62	0.36	0
NDVI	0.2	0.69	0.45	0.14	0.31	0.22	0.68	0.46	0.14	0.3
VERIS N3	5.5	39.12	13.37	6.58	0.49	5.51	26.6	12.82	6.18	0.48
VERIS N4	7.16	37.83	17.06	6.81	0.4	6.94	33.26	16.58	7.17	0.43
Cone index (Mpa)	30.27	58.74	44.88	6.85	0.15	22.47	58.38	43.88	10.55	0.24

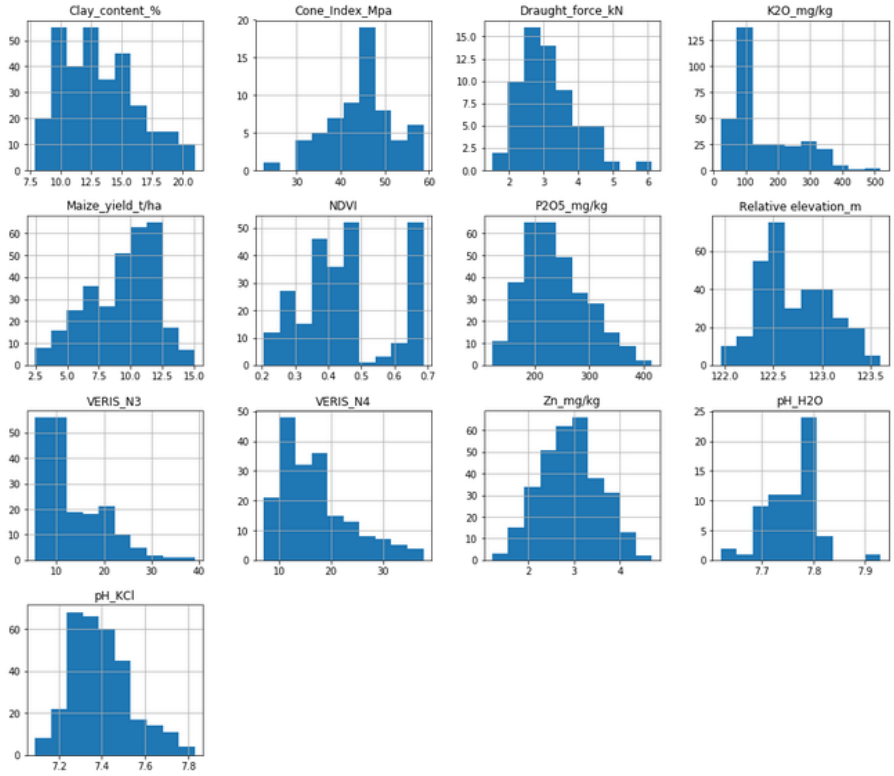

Figure 2. Value distributions of influencing factors of maize yields.

The distribution of maize yield is shown in *Figure 3* in Boxplot.

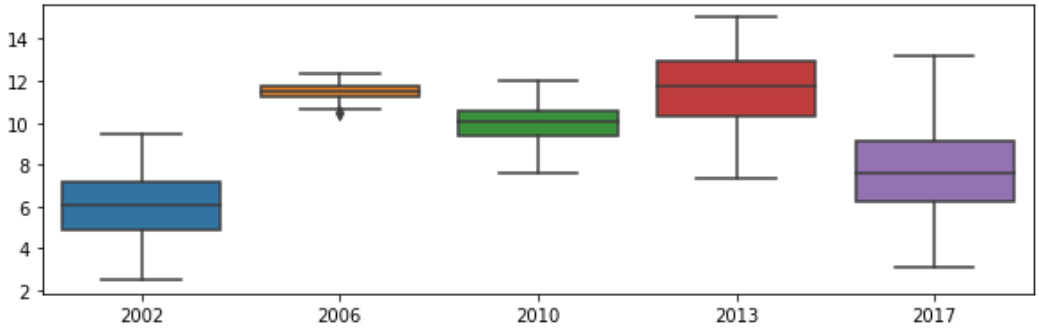

Figure 3. Yield distribution in each year considered in the study.

In Table 2, we show the results of 5-fold cross-validation to find the optimal parameters of CP-ANN, XY-F, SKN, XGBoost and SVM. Sensitivity, specificity, accuracy and ROC AUC were calculated to evaluate the classification models. Mean and standard deviation of each evaluation measure were generated by repeating the 5-fold cross-validation ten times. XGBoost reached the highest performance in 3 of the 4 evaluation measures on the calibration set using cross-validations: 92.1% of accuracy, 87.9% of sensitivity, 96.7 of ROC AUC. Results of independent validation are shown in Table 3 for the same models. XGBoost reached the highest performance in 3 of the 4 evaluation measures on the validation set: 95.38% of accuracy, 91.3% of sensitivity, 97.62 of specificity.

Table 2. Results of 5-fold cross-validations (CV) on the training dataset for Counter-propagation artificial neural network (CP-ANN), XY-fused networks (XY-Fs), supervised Kohonen network (SKN), extreme gradient boosting (XGBoost) and support-vector machine (SVM). Mean and standard deviation of accuracy, sensitivity, specificity and ROC AUC values were calculated by 10 experiment of 5-fold cross-validation. CP-ANN, XY-F and SKN were trained by 30 by 30 neurons and 50 epochs. XGBoost was trained with 10 trees and maximum depth 2. Linear kernel function with standard and L^2 normalization was used to train SVM.

		Accuracy		Sensitivity		Specificity		ROC AUC	
	No. exps	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
CP-ANN	10	0.9076	0.0097	0.8159	0.0272	0.9524	0.0097	NA	NA
XY-F	10	0.916	0.0078	0.8549	0.0272	0.9458	0.0082	NA	NA
SKN	10	0.9144	0.0091	0.8634	0.0138	0.9393	0.0115	NA	NA
XGBoost	10	0.921	0.007	0.879	0.016	0.941	0.007	0.967	0.008
SVM	10	0.9001	0.0065	0.8595	0.0205	0.9205	0.0079	0.9405	0.0060

Table 3. Results of the predictions on the independent test dataset for Counter-propagation artificial neural network (CP-ANN), XY-fused networks (XY-Fs), supervised Kohonen network (SKN), extreme gradient boosting (XGBoost) and support-vector machine (SVM), trained on the whole training set with the same parameter settings as described in the legend of Table 2. Accuracy, sensitivity, specificity and ROC AUC values were calculated.

	Accuracy	Sensitivity	Specificity	ROC AUC
CP-ANN	0.8923	0.8261	0.9286	NA
XY-F	0.8615	0.7826	0.9048	NA
SKN	0.8923	0.8261	0.9286	NA
XGBoost	0.9538	0.9130	0.9762	0.9829
SVM	0.9538	0.9565	0.9524	0.9917

The positive and negative influencing factors for maize yield were divided. In this study, NDVI was identified as the overall most important positive factor in all years. The following influencing parameters were K_2O and soil electrical conductivity measurements (Veris N3).
However, soil draught force affected on yield as second largest impact (negative correlation). The additional influencing variables were *ECa* (Veris N4), relative elevation, P_2O_5, Zn and pH (H2O), in general, negative for yield in all years.

Conclusions

Five machine learning analysis methods, namely, CP-ANN, XY-F, SKN, XGBoost and SVM were presented. Machine learning to adapt site-specific data on soil and crop with two classification of maize yield were applied. In this paper, we presented the yield prediction model of site-specific maize yield in 5 years. We use CP-ANN, XY-F, SKN, XGBoost and SVM for defining the machine learning models between the maize yield and the influencing factors of yield. We used meteorological parameters, soil dataset and crop parameter. The results showed that the best prediction method (accuracy) was the XGBoost. This prediction accuracy reached on the test set: 95.38% of accuracy, 91.3% of sensitivity and 97.62% of specificity. The results showed that XGBoost was very effective in medium-high maize yield prediction.

Acknowledgements

This work was supported by the EFOP-3.6.3-VEKOP-16-2017-00008 project. The project is co-financed by the European Union at the European Social Fund. The work was supported by the 2018-1.2.1-NKP-00008 "Exploring the Mathematical Foundations of Artificial Intelligence" project.

References

Ballabio et al. 2012. A Matlab toolbox for self organizing maps and supervised neural networks learning strategies. Chemometrics and intelligent laboratory systems 118: 24-32.

Chen, T. and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minig, 785-794.

Chlingaryan, A., Sukkarieh, S., Whelan, B. 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Computers and Electronics in Agriculture. Vol. 151, 61-69.

Dahikar, S. S., Rode, S. V. 2014. Agricultural crop yield prediction using artificial neural network approach. Int. J. of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering. Vol. 2.

Elgaali, L., Garcia, L. 2004. Neural Network Modeling of Climate Change Impact on Irrigation Water Supplies in Arkansas River Basin, Hyrology Days, Colorado State University).

Folberth, C., Baklanos, A., Balkovic, J., Skalsky, R., Khabarov, N., Obersteiner, M. 2019. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. Agricultural and Forest Meteorology. Vol. 264, 1-15.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics. 1189-1232.

Iizumi, T., Shin, Y., Kim, W., Kim, M., Choi, J. 2018. Global crop yield forecasting using seasonal climate information from a multi-model ensemble. Climate Services. Vol. 11, 13-23.

Irmak, A., Jones, J. W., Batchelor, W. D., Irmak, S., Boote, K. J., Paz, J. O. 2006. Artificial neural network model as a data analysis tool in precision farming. Trans. Of ASABE. Vol. 49(6), 2027-2037.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., Shearer, S. 2018. Integration of hihg resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Computers and Electronics in Agriculture. Vol. 153, 213-225.

Laacouri, A., Nigon, T., Mulla, D., Yang, C. 2018. Case study comparing machine learning and vegetation indices for assessing corn nitrogen status in an agricultural field in Minnesota. Paper from the Proceedings of the 14th International Conference on Precision Agriculture, June 24-June 27, Montreal, Quebec, Canada

Longchamps, L., Tremblay, N., Panneton, B. 2018. Observational studies in agriculture:paradigm shift required. Paper from the Proceedings of the 14th International Conference on Precision Agriculture, June 24-June 27, Montreal, Quebec, Canada

Maeda, Y., Goyodani, T., Nishiuchi, S., Kita, E. 2018. Yield prediction of paddy rice with machine learning. Int'l Conf. Par. and Dist. Proc. Tech. and Appl. PDPTA'18. 361-365.

Miao, Y., Mulla, D. J., Robert, P. C. 2006. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agriculture 7 117-135.

Mike-Hegedűs F. 2006. Applying fuzzy logic and neural networks to database evaluation in precision agriculture. PhD thesis in Hungarian. University of West-Hungary, Mosonmagyaróvár.

Neményi, M., Mesterházi, P. Á. and Milics, G. 2006. An Application of Tillage force Mapping as a Cropping Management Tool. Biosystems Engineering. Vol. 94, 3, pp. 351-357.

Nyéki, A., Milics, G., Kovács, A. J., Neményi, M. 2013. Improving yield advisory models for precision agriculture with special regards to soil compaction in maize production. In: Stafford, J.V. (ed.), Precision Agriculture ’13. Academic Publishers. Wageningen, The Netherlands. pp. 443–451.

Nyéki, A., Milics, G., Kovács, A. J., Neményi, M. 2017: Effect of soil compaction on cereal yield. Review. Cereal Research Communication. 45.1:1-22.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L. and Mouaen, A. M. 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture 121 57-65.

Szőllősi, I. 2003. A 3T SYSTEM készülékkel mért penetrációs ellenállás és nedvességtartalom összefüggése vályog fizikai féleségű talajon. (English abstract: Correlations between the penetration resistance registered with a 3T SYSTEM instrument and the moisture content of a soil with loam texture.) Agrokémiai és Talajtan Vol. 52, 3-4, pp. 263-274.