Check for
updates

Application of spatio-temporal data in site-specific maize yield prediction with machine learning methods

A. Nyéki[1] · C. Kerepesi[2] · B. Daróczy[3] · A. Benczúr[4] · G. Milics[1] · J. Nagy[5] ·
E. Harsányi[5] · A. J. Kovács[1] · M. Neményi[1]

Abstract

In order to meet the requirements of sustainability and to determine yield drivers and limiting factors, it is now more likely that traditional yield modelling will be carried out using artificial intelligence (AI). The aim of this study was to predict maize yields using AI that uses spatio-temporal training data. The paper has advanced a new method of maize yield prediction, which is based on spatio-temporal data mining. To find the best solution, various models were used: counter-propagation artificial neural networks (CP-ANNs), XY-fused Querynetworks (XY-Fs), supervised Kohonen networks (SKNs), neural networks with Rectangular Linear Activations (ReLU), extreme gradient boosting (XGBoost), support-vector machine (SVM), and different subsets of the independent variables in five vegetation periods. Input variables for modelling included: soil parameters (pH, P_2O_5, K_2O, Zn, clay content, ECa, draught force, Cone index), micro-relief averages, and meteorological parameters for the 63 treatment units in a 15.3 ha research field. The best performing method (XGBoost) reached 92.1% and 95.3% accuracy on the training and the test sets. Additionally, a novel method was introduced to treat individual units in a lattice system. The lattice-based smoothing performed an additional increase in Area under the curve (AUC) to 97.5% over the individual predictions of the XGBoost model. The models were developed using 48 different subsets of variables to determine which variables consistently contributed to prediction accuracy. By comparing the resulting models, it was shown that the best regression model was Extreme Gradient Boosting Trees, with 92.1% accuracy (on the training set). In addition, the method calculates the influence of the spatial distribution of site-specific soil fertility on maize grain yields. This paper provides a new method of spatio-temporal data analyses, taking the most important influencing factors on maize yields into account.

Keywords Maize yield prediction · Machine learning methods · Gradient boosting (XGBoost) · Random fields · Yield influencing variables · Decision support in crop production

✉ A. Nyéki
nyeki.aniko@sze.hu

Extended author information available on the last page of the article

🖄 Springer

Introduction

The current undesirable effects of agriculture (greenhouse gas emission, groundwater pollution or nitrification) on the biosphere cannot be reduced using traditional experiments (Longchamps et al., 2018). The gap between environmentally adverse impacts and the slowly expanding knowledge base is growing. Yield prediction is nowadays an essential research field, as it has an equally important point of reference for farm management during planning, agrotechnological intervention and preharvest processes. Farm and in-situ observations together with existing databases provide the opportunity to not only predict yields using "simpler" statistical methods or decision support systems that are already used as an extension, but they also enable the potential for machine learning (ML) to be used. The latter has the advantage of being able to handle many parameters indefinitely in time and space, i.e., big data databases created using precision management tools and data collection capabilities can be used mostly in the area of meteorological, technological or soil-related information, including characteristics of different plant species.

One viable solution for this problem is site-specific data collection, i.e., obtaining as much information about events as possible during the vegetation period, as it is possible to determine the processes and their causes, and to find solutions to prevent negative effects on the environment after precise intervention. This approach will result in a controlled, monitored environment in space and time. Furthermore, it also paves the way for a new decision support system that will be much more complex than currently used systems. In earlier studies (Nyéki et al., 2013, 2017), it was concluded that increasing input data for plant physiological models could increase the accuracy of prediction only to a limited extent. Consequently, there has to be a change in scientific methods and the potential of big data needs to be better exploited than before, which will result in the need to use artificial intelligence (AI), particularly ML methods.

Therefore, one of the requirements of precision farming and crop production is to use on-line and on-the-go systems that are capable of logging a large amount of data and soil parameters in a relatively short period of time. This representative data collection results in mean cost-effectiveness, but it could also refine current knowledge by analysing the data, and repeating measurements. Vis–NIR spectroscopy can be used to determine soil moisture content, pH and soil organic matter content (Mouazen et al., 2007). Furthermore the total nitrogen content, the total carbon content, magnesium, calcium, cation exchange capacity and the available phosphorus content determination tests have been performed successfully in Europe (Mouazen et al., 2009). Looking into the sources of data used to make deep learning models, mainly large data sets of images are used; in some cases, they contain thousands of images (Dyrmann et al., 2016; Mohanty et al., 2016). In general, the more complicated the problem to be solved, the more data is required (Kamilaris & Prenafeta-Boldú, 2018).

The most successful yield prediction ML techniques have been artificial neural networks (Fortin et al., 2011; Liu et al., 2001; Safa et al., 2004), support vector machines (Ruß, 2009), M5-Prime Regression Trees (Frausto-Solis et al., 2009; Ruß & Kruse, 2010; Wang & Witten, 1997) and k-nearest neighbour methods (Zhang et al., 2010). Cao et al. (2013) calculated the influence of the spatial distribution of adjacent-area soil fertility on maize yields by using neural networks.

AI is used by an increasing number of researchers to model a wide range of tasks in agriculture. According to the available technical literature sources, various methods are used for precision agriculture decision support. There are a large number of

studies focusing on the relationship between soil properties and yield (Miao et al., 2006; Mike-Hegedűs, 2006). CP-ANN, SKN and XY-F models were utilized with high resolution data on soil and crop with isofrequency classes of wheat yield productivity (Pantazi et al., 2016). The best results were obtained from the SKN network for prediction of the low category of wheat yield with a correct classification reached 91.3% for both cross and independent validation. Climate effects have also been analysed by AI (Elgaali & Garcia, 2004; Iizumi et al., 2018).

The gradient boosting machine (GBM) is a very effective state-of-the-art ML method ML (Chen & Guestrin, 2016; Friedman, 2001). XGBoost has been used to determine rice yields using weather data, cultivation data and location information (Maeda et al., 2018). The best accuracy was estimated to be 74.4%.

Laacouri et al. (2018) assessed maize nitrogen deficiency levels comparing ML and vegetation indices. Hyperspectral images were collected using a hexacopter UAV platform. Eight ML algorithms were compared for their accuracy, among other gradient boosting models. The authors concluded that hyperspectral imagery combined with ML improved the assessment of maize N stress status at the V5 growth stage (Abendroth et al., 2011), and achieved more than 90% classification accuracy when the entire spectrum was mined. GBM achieved an overall classification accuracy of 89%. SVM, logistic regression (LR), Multi-Layer Perceptron (MLP) and GBM showed promising results.

Folberth et al. (2019) explored two ML approaches: XGBoost and random forests. ML was trained on a global scale maize simulation of a GGCM (Global Gridded Crop Model). This method provided a very high accuracy ($R^2 > 0.96$) for predictions of maize yields, evapotranspiration and crop available water. ML algorithms were trained on a global scale maize yield simulation. Extreme gradient boosting showed moderately better results especially for inter-annual yield variability (Folberth et al., 2019).

Fan et al. (2018) proposed two ML methods, SVM and XGBoost. Overall, the XGBoost model had comparable prediction accuracy with the SVM model to estimate daily solar radiation.

Khanal et al. (2018) integrated high resolution remotely sensed data and ML techniques for spatial prediction of soil properties and maize yields. High spatial resolution (< 1 m) was used for the capturing of soil properties (SOM, CEC, K, Mg, and pH) and maize yields. The accuracies of seven statistical methods (LM—linear regression, RF—random forest, SVM with linear and radial kernel function, GBM, NN—neural network, and CUB—cubist) were compared for their ability to predict soil properties and maize yields. For pH and maize yield prediction, the GBM and the RF model performed better than other models. These analyses showed that the XGBoost ML seems to be a favourable method. However, XGBoost has not been applied to spatio-temporal maize yield prediction.

The aim of the study was to increase the accuracy of maize yield prediction. The above literature methods confirm that the XGBoost gives more accurate yield prediction than ANNs. However, further studies are needed to prove the advantages of the XGBoost method and to overcome the limitations of individual unit predictions. Among these, spatio-temporal studies may be highlighted, as conclusions can be drawn from several years of experiments in agronomy. The goal of this study was to improve a model that highlights decision making in crop production. This aim was achieved by applying ML methods comprising of soil data, meteorological data and satellite-based data.

Materials and methods

Experimental site

Field experiments were carried out in a 23.52 ha experimental field (Fig. 1) belonging to Széchenyi István University in the vicinity of Mosonmagyaróvár, Hungary [47°54′20.00″N; 17°15′10.00″E]. The experimental field—with 63 management units (Mike-Hegedűs, 2006)—is an alluvial plain of the Leitha River on which precision agriculture has been applied since 2001. The soil textures according to the USDA (ND) 2020 indicated the soil types of loam, silty loam and sandy loam. The terrain has a slight slope of 5% with elevation varying between 133 and 138 m.

Data sources

Data from various sources were integrated into this research.

Crop and soil parameters affecting maize yield

Soil samples were collected from each management zone in order to be subjected to standard methods of laboratory analysis. Clay, silt, sand content (%), pH, phosphorous (P_2O_5), potassium (K_2O) and zinc (Zn) were measured.

The relative elevation (microrelief parameter, m) was determined by a Sokkia-Radian-IS GSR-2600 (Sokkia BV, Essebaan, The Netherlands) differential global navigation satellite system.

Crop and yield parameters were diagnosed in order to estimate crop characteristics. NDVI was calculated from satellite data by Landsat on 26 July 2006, 5 July 2010, 29 July

Fig. 1 Site-specific maize yield (2017) of the experimental field over the 63 treatment units
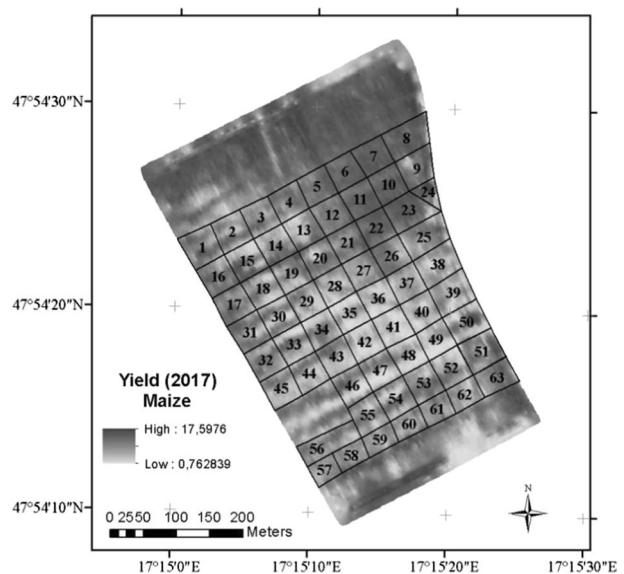
2013 and 9 August 2017. The satellite images were collected under high quality (maximum cloud coverage was under 5%) atmospheric conditions during the maize flowering period.

The apparent soil electrical conductivity (EC_a) was measured by a Veris Soil EC-3100 (Salina, KS, USA) instrument at two depths, shallow (less than 0.3 m below the surface) and deep (between 0.3 and 0.9 m below the surface). In total, data was collected from 63 management zones (~0.25 ha) of the research field, covering approximately 15 ha.

Meteorological data

Of the 35 meteorology features observed, the following five values were measured in each month from April to October during the vegetation period of maize: precipitation sum, average temperature, average relative precipitation, evapotranspiration, aridity index (evapotranspiration/precipitation). Monthly averages were estimated using daily databases including the sum of precipitation, average temperature, average relative precipitation, evapotranspiration and aridity index in each month. The observed field was 1.78 km from the weather station of the university.

On-the-go soil draft measurements were carried out by a self-developed system (Neményi et al., 2006). A conventional tillage system after cereal harvest consists of stubble tillage (~15 cm) using cultivator and tooth harrows, mouldboard ploughing (~20 cm) and disking in the end of summer, and tooth harrow to prepare seedbed for winter crops. Ploughing is applied for spring cereals in late fall and the same tillage system are prepared in spring for maize sowing. A Trimble AgGPS 542 RTK base system (Trimble Ltd., Sunnyvale, CA, USA) was used in order to find the correct position of the location of measurements. Yield data was measured with a Deutz Fahr grain harvester until 2010, which was equipped with an Agrocom ACT 260 (Class GmbH., Harsewinkel, Germany) yield sensor. From 2011, an AgLeader InCommand™ 1200 (AgLeader Technology Inc., Ames, IA, USA) yield mapping system was used, which was mounted on a CLAAS Medion 340 grain harvester. Impact based mass flow measurement was used on both machines. A site-specific penetrometer measurement was carried out by means of a 3 T system (Szőllősi, 2003). The soil measurements were done at 20–25 points across the management units in the topsoil (upper 30 cm). All of the site-specific measurements were carried out before the maize vegetation seasons.

The obtained yield, electrical conductivity and soil draft measurements were averaged in each management zone.

Feature table for supervised machine learning/dataset definition

Prior to supervised ML, a feature table was created from the measured parameters (Table 1). The dataset combined data from sources mentioned above.

The feature table consists of 315 (63×5) rows for each treatment unit and year, and also 48 columns for each parameter (11 soil parameters, 35 meteorology parameters and one crop parameter) and maize yield. In summary, four types of parameters were used:

(i) Soil parameters measured over five years (2002, 2006, 2010, 2013 and 2017): pH, P_2O_5, K_2O, Zn; measured over 3 years (2010, 2013 and 2017): EC_a in two layers, Veris N3 and N4; measured over 1 year (2010): Cone Index (MPa), pH (water) and draught force (kN, 2002);

(ii) Crop parameter (NDVI) measured over 4 years: 2006, 2010, 2013 and 2017;

Table 1 Descriptive statistics of the selected soil and crop parameters and the maize yields for the training (80% of samples) and test datasets (20% of samples)

	Years	Training dataset (n = 250)					Test dataset (n = 65)				
		Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV
Maize yield (t/ha)	5	3.25	15	9.39	2.62	0.28	2.46	15.05	9.23	2.84	0.31
pH (water)	1	7.62	7.82	7.75	0.04	0.01	7.7	7.93	7.78	0.06	0.01
pH (KCl)	5	7.09	7.83	7.39	0.14	0.02	7.14	7.71	7.4	0.14	0.02
P2O5 (mg/kg)	5	123	415	234.24	55.67	0.24	151	380	250.95	57.84	0.23
K2O (mg/kg)	5	22.6	518	153.05	102.06	0.67	60.1	400	148.6	91.4	0.62
Zn (mg/kg)	5	1.21	4.7	2.91	0.66	0.23	1.74	4.1	2.95	0.57	0.19
Clay content %	5	8.4	21	13.28	3.07	0.23	7.9	18.4	13.14	3.45	0.26
Draught force (kN)	1	1.52	4.72	3.1	0.72	0.23	1.9	6.12	3.4	1.31	0.38
Relative elevation (m)	5	122.18	123.6	122.74	0.34	0	121.96	123.11	122.62	0.36	0
NDVI	4	0.2	0.69	0.45	0.14	0.31	0.22	0.68	0.46	0.14	0.3
VERIS N3	3	5.5	39.12	13.37	6.58	0.49	5.51	26.6	12.82	6.18	0.48
VERIS N4	3	7.16	37.83	17.06	6.81	0.4	6.94	33.26	16.58	7.17	0.43
Cone index (MPa)	1	30.27	58.74	44.88	6.85	0.15	22.47	58.38	43.88	10.55	0.24

Min minimum, *Max* maximum, *Mean* mean (average), *SD* standard deviation, *CV* coefficient of variation (the ratio of the standard deviation to the mean)

(iii) Non-yearly-changing soil parameters [clay content (%), relative elevation (m) meas-
　　　ured over one year (2010) but used in every year];

(iv) Out of the 35 meteorology features observed, the following six values were meas-
　　　ured in each month from April to October during the growing season of maize: sum
　　　of precipitation, average temperature, average relative precipitation, vaporization,
　　　evapotranspiration, aridity index.

Machine learning methods

Various ML methods were utilized in this research, including neural networks, tree ensem-
bles and a kernel method to predict maize yields. Previous research and results were driven
by the analysis and unknown correlations in the feature space as traditional neural net-
works and tree ensembles are more suitable for different problems. Self-organizing maps
(SOM, Kohonen, 1982) were trained as an unsupervised latent representation method in
a comparison to Pantazi et al. (2016). The results of low dimensional feature represen-
tation were used in a supervised manner as in Kohonen (2001) networks, where classifi-
cation was trained with the maps. Additionally, two modified versions of self-organizing
maps, counter-propagation artificial neural networks (CP-ANN, Gasteiger & Zupan, 1993)
together with XY-fused networks (XY-F, Melssen et al., 2006), were used in a similar
fashion but with a different structure. For more details, see Ballabio and Vasighi (2012).
Additionally, fully connected deep networks with ANN were trained using rectangular
linear units (ReLU, Montufar et al., 2014) as activations and the parameters were opti-
mized with adaptive momentum (Kingma & Ba, 2015). In comparison to SOMs and ReLU
networks, tree ensemble methods are less powerful as a function approximation while the
smaller capacity helps in the case of small datasets or non-spatio-temporal structural vari-
ables. Tree ensembles build a set of "weak" classifiers from small, almost random decision
trees. There are several methods to determine the set of decision trees and their impor-
tance e.g., random forest (Breiman, 2001), adaptive boosting (Freund & Schapire, 1997)
or GBM (Friedman, 2001). A recent implementation of gradient boosted trees was used in
the experiments because these outperform other methods. Additionally, the support vector
machine (Cortes & Vapnik, 1995) models were trained with various kernel functions. As
the current dataset describes a single site with 63 treatment units, the units are connected
and therefore form a lattice. Besag (1974) suggested practical methods over botanical lat-
tice systems, e.g., *Plantago lanceolata,* and showed that typical systems act as a Markov
random field (MRF). The idea was examined based on the individual previously-performed
predictions per unit by smoothing and predictions with the neighbouring predictions of the
best methods (spatial smoothing) in a simple setup.

Prediction of maize yields

Cross-validation and train-test split

The structure of the dataset for model training and testing was defined. Before training,
the feature table was split into two parts: a training set of 50 and a testing set of 13 ran-
domly selected treatment units. 5 years of measurements were used and 250 training and
65 testing samples were obtained. The best parameters of the ML models were selected by
a fivefold cross-validation method over the training set. The final models were trained on

the entire training set and tested on the independent testing set. For experiments over the site structure, the units in the study site were split into three smaller sub-sites according to their position.

Binary classification

Maize yield values 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 obtained 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 ML 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 et al., 2012), and have been used for wheat yield prediction in a recent study (Pantazi et al., 2016). XGBoost is an efficient implementation of the GBM (Friedman, 2001). XGBoost has been applied successfully in a wide variety of classification problems (Chen & Guestrin, 2016). The raw predictions by the best XGBoost model were used for spatial smoothing and were trained on a sub-site. The best neighbourhood model was determined on a separate site.

Evaluation measures for binary classification

Sensitivity, specificity, accuracy and ROC AUC (area under the receiver operating characteristic curve) were calculated to evaluate the classification models. Sensitivity means the proportion of actual positives (low yield samples) and specificity means the proportion of actual negatives (middle-high yield samples). Accuracy is the proportion of the total number of samples that are correctly identified. The ROC curve is defined by the point pairs of true positive rates (sensitivity) and false positive rates (1-specificity) at different threshold settings. The AUC can be interpreted as the probability of classifying a positive sample with higher confidence than a negative sample.

Regression

Additionally, regression methods were experimented with for continuous maize yield prediction. The main difference is the target variable, thus the same ML methods were applied as in binary classification without the underperforming SOM methods since, besides XGBoost, the best performing models were ReLU networks and SVM.

Evaluation measures for regression

Model performance was evaluated following Tan et al. (2016) using models with mean absolute error (MAE, the mean absolute difference between the predicted value and the ground truth), relative error (RE, the mean ratio of absolute error and the ground truth) and fraction of unexplained variance (FVU, the ratio of error deviation and target deviation).
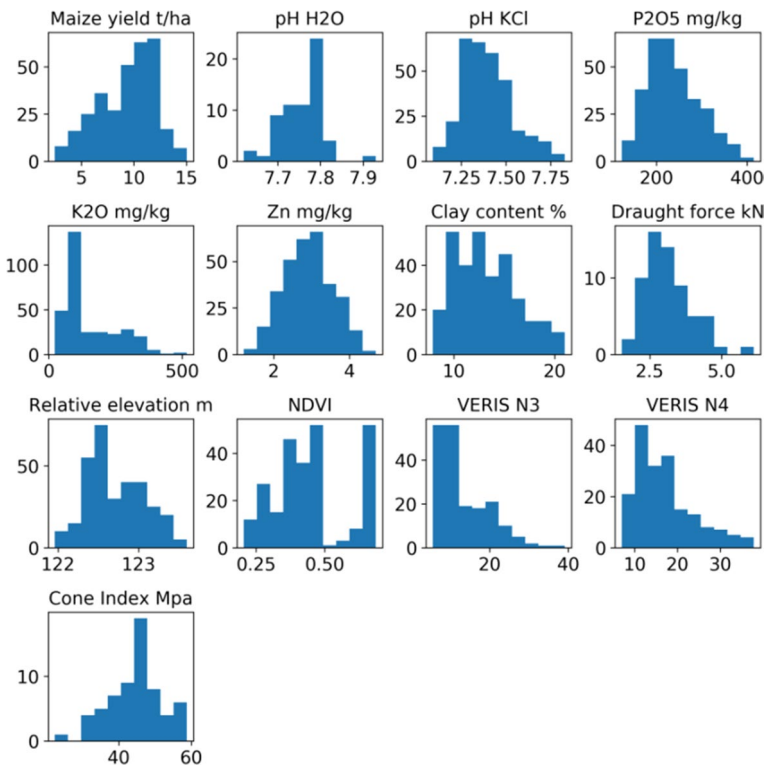
Results

## Accuracy of yield prediction with supervised models

Four neural networks models (CP-ANN, XY-F, SKN, SVM) and a GBM (XGBoost) were trained on a feature table of 47 measurements of 63 treatment units in 5 years, resulting in 315 samples altogether. 50 randomly chosen treatment units (80%) were used for training and 13 for testing (20%). The statistics of the training and testing datasets are shown in Table 1. Table 1 summarises the performance of linear regression models, showing the minimum, maximum, mean and standard deviation for each parameter for the training and the test datasets.

The distributions of features that influence maize yields are shown in Fig. 2 and the yield distribution by year in Fig. 3 in a boxplot.

Table 2 shows the results of the fivefold 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 fivefold 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, and 96.7%

Fig. 2 Value distribution of maize yields and its influencing factors. Horizontal-axis: value of the given variable. Vertical axis: frequency
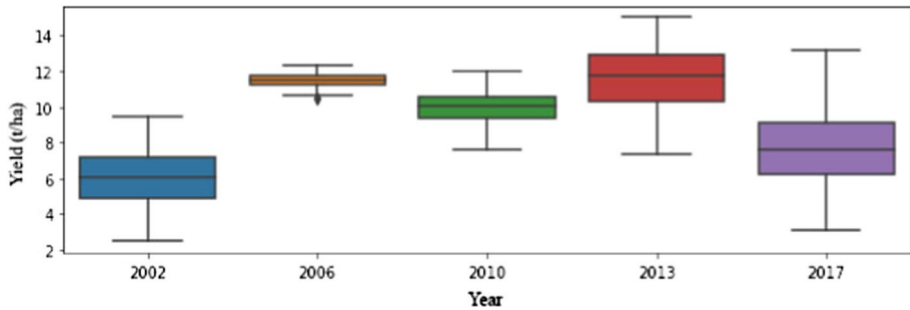
Fig. 3 Yield distribution in each year of the study

Table 2 Results of fivefold 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 machine (XGBoost) and support-vector machine (SVM)

	No. Exps	Accuracy		Sensitivity		Specificity		ROC AUC	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
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
ReLU	10	0.9001	0.0162	0.9495	0.0137	0.6452	0.0439	0.8785	0.0156
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

Mean and standard deviation of accuracy, sensitivity, specificity and ROC AUC values were calculated by ten experiments of fivefold cross-validation. CP-ANN, XY-F and SKN were trained by 30 by 30 neurons and 50 epochs. XGBoost was trained with ten trees and a maximum depth of 2. Linear kernel function with standard and L^2 normalization were used to train SVM

Table 3 Results of the predictions on the independent test dataset for the counter-propagation artificial neural network (CP-ANN), XY-fused networks (XY-Fs), supervised Kohonen network (SKN), extreme gradient machine (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	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
ReLU	0.8951	0.9762	0.7794	0.9398
XGBoost	0.9538	0.9130	0.9762	0.9829
SVM	0.9538	0.9565	0.9524	0.9917

Accuracy, sensitivity, specificity and ROC AUC values were calculated

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 in the validation set: 95.38% of accuracy, 91.3% of sensitivity, and 97.62 of specificity.

Gradient boosted tree algorithms are capable of selecting the most important uncorrelated features by building small decision trees of a few of the most important features and gradually refining the small models by adding new trees (Kerepesi et al., 2018). The positive and negative influencing factors for maize yields were divided. In this study, NDVI was identified as the overall most important positive factor in every year. The following influencing parameters were K_2O and soil electrical conductivity measurements (Veris N3).

However, soil draught force had the second largest impact (negative correlation) on yields. The additional influencing variables, i.e., ECa (Veris N4), relative elevation, P_2O_5, Zn and pH (H_2O) were, in general, negative for yields in every year.

Generally speaking, the XGBoost model yielded the minimum prediction errors and generated equitable results compared to the other four models.

Extreme gradient boosted trees

In this study, the classified low and moderately high maize yields were examined, based on the field results from 5 years: 315 data points, involving 250 training data.

Twelve gradient boosted models were trained on the feature set, defined in Table 1. Each dataset was randomly divided into training and test sets. The training set consisted of 80% of the samples obtained. The performance of gradient boosted models can be seen in Table 4. Various maximal depth and number of ensembles were used in this project. The least complex model, a single decision tree, reached 87% accuracy with a maximal depth of ten. The label "mean ACC" denotes the prediction accuracy of a yield. The order is set to mean ACC. Gradient boosted models achieved results between 92.2 and 87.2%, although the prediction model with 100 trees and a maximal depth of 2 performed best in this experiment.

Table 4 contains the results of 12 models, which are based on the mean ACC results. They are based on two main parameters: maximal depth of decision trees and the number of trees. The 12 runs have different parameters and different models were used. The simplest model (Run 0) has an ACC of 0.872 (87.2% accuracy). In this model, results were calculated based on the amount of precipitation. The models (the runs) differ in the number of trees and leaves. Run 8 performed with an accuracy of 89.5% (mean ACC). It is composed of three depths, which is necessary in this case.

The decision to train eleven combinations was made mainly based on field experiments, database limitations and the test dataset, so the models (trees), parameters and analysis results are not necessarily optimal in other cases. For more details see Tan et al. (2016).

Important influencing factors on maize yields

Using the trees of the XGBoost model, it is possible to rank the features based on their position in the trees. Even a simple count of occurrence may yield a generously good ranking of the features but modern XGBoost implementations rank the features upon their positions in the trees too and, most importantly, based on how their split gained better separation, (for more details see Chen & Guestrin, 2016). 13 influencing factors were selected (Fig. 4) as the additional features had low score in the ranking list. Over

Table 4 Statistical values of twelve models

Number of trees	Max_d	n_est	n_exps	Mean AUC	Std AUC	Mean ACC	Std ACC	Mean sens	Std sens	Mean spec	Std spec
7	2	100	10	0.966	0.004	0.922	0.009	0.942	0.009	0.882	0.016
11	3	100	10	0.968	0.004	0.910	0.008	0.932	0.009	0.866	0.019
3	1	100	10	0.937	0.012	0.907	0.006	0.922	0.004	0.877	0.017
10	3	10	10	0.957	0.007	0.901	0.014	0.914	0.009	0.874	0.034
8	3	1	10	0.946	0.007	0.895	0.010	0.909	0.014	0.866	0.036
9	3	2	10	0.946	0.010	0.895	0.010	0.910	0.018	0.865	0.035
6	2	10	10	0.947	0.006	0.895	0.011	0.909	0.016	0.866	0.018
4	2	1	10	0.924	0.012	0.894	0.009	0.924	0.013	0.830	0.022
5	2	2	10	0.917	0.011	0.883	0.011	0.896	0.013	0.856	0.025
0	1	1	10	0.859	0.006	0.872	0.000	0.851	0.000	0.915	0.000
1	1	2	10	0.856	0.006	0.872	0.000	0.851	0.000	0.915	0.000
2	1	10	10	0.889	0.012	0.872	0.000	0.851	0.000	0.915	0.000
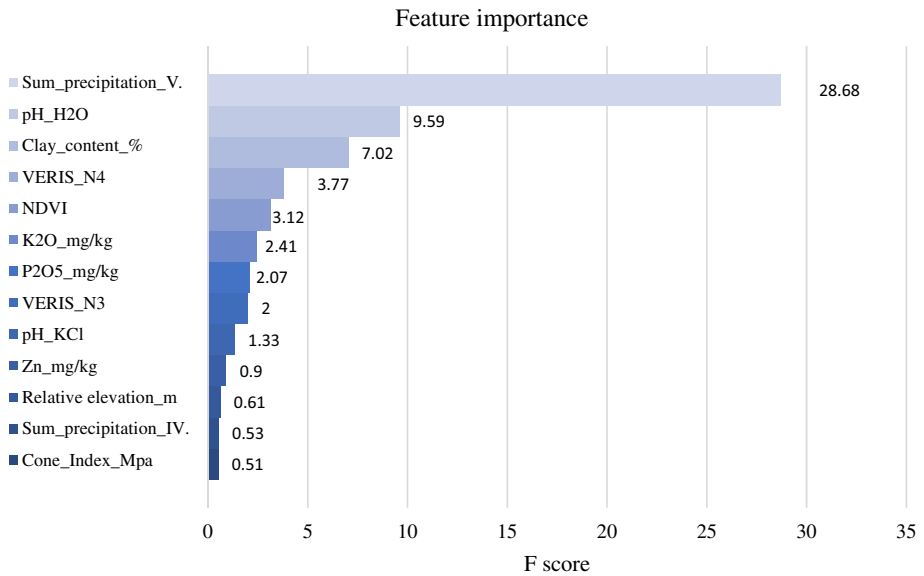
**Fig. 4** Ranking of importance features on maize yields

5 years, precipitation in May was identified as the most important influencing factor for yields. Other important factors among the top five included pH, clay content of soil, electrical conductivity in upper layer and normalized difference vegetation index. Thus, yields were strongly affected by both soil physical properties and pH. These parameters have similar results to electrical conductivity in this field, as well. The potassium content in soil was the sixth most important factor. Soil Zn content, relative elevation and amount of precipitation in June (Table 5) were also important for grain yields. Miao et al. (2006) also found that the relative elevation and Zn content have significant effect on maize yields.

Since the amount of precipitation in May was identified as the most important influencing factor, it is necessary to compare the different meteorological parameters during the different vegetation seasons. The "Sum_precipitation_V" indicates the water demand in the juvenile phase in maize vegetation.

Figure 5 shows the site-specific maize yield distribution in management zones over the 5 years examined. The maps show similar patterns in different years, where low yield values appeared in the middle part of the field. In this part, the sand content of the soil is higher than in other treatment units.

Table 5 Meteorological parameters

	2002	2006	2010	2013	2017
Sum. precipitation during vegetation period (IV–X)	353.8	405.4	597.9	398.4	332.7
Evapotranspiration (mm)	911	881	690	845	951
Aridity index	2.6	2.2	1.2	2.1	2.9
Precipitation in May (mm)	25.9	90	150.3	125.9	26.7
Precipitation in June (mm)	40.2	59.1	100	42.2	40
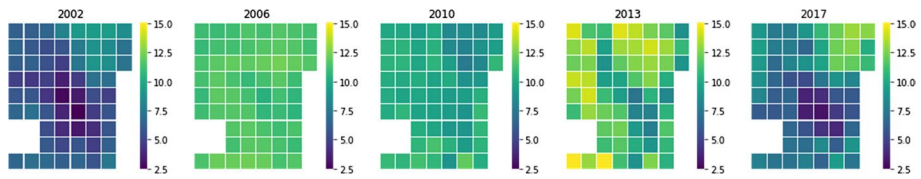
Fig. 5 Maize yield distribution in the field using the same scale (absolute yield distribution, t/ha)

In 2002, the growing season (April–October) was very dry, having 358 mm rainfall (Table 5). In 2017, rainfall was lower than normal. These results show the Run 0 booster model results. If the precipitation in May is relatively low (under 58 mm) and the management zone has lower clay content (<14.65%), the model predicts a lower yield.

Experiments with important features

From the trees of the XGBoost model, analysis can identify a subset of the original feature space, the features of which were selected by any of the trees ("important" features) in the model without changing the output. The "important" features were selected and experimented on using ReLU networks and SVM because the new feature space is more likely to be less noisy than a vector space due to the tree building method in XGBoost. The new ReLU networks and SVM models were trained in the new feature space and evaluated the models in the independent test set. In Table 6 the results suggest that there were some features which were not selected by XGBoost but it was necessary for SVM; however ReLU results stayed generally the same, as shown in Table 3.

Spatial smoothing

The study site was split into three separate sub-sites to measure the correlation between adjacent sites and form a lattice system similar to Besag (1974). Treatment units 1–31 were used for training and treatment units 32–45 for validation, and units 46–63 for test sets (see Fig. 1). The prior models, XGBoost (100 trees with a maximal depth of 2) and SVM (radial basis function, RBF) kernel with a cost of 32 over standard, power and l2 normalized features (Daróczy et al., 2015) with the best parameters and normalization, were trained over the training set, while the best linear combination of adjacent unit confidence scores was determined for a single unit based on the validation set and evaluated the findings on the test units. It is worth mentioning that as the training set is a significantly smaller set of treatment units than in previous experiments, the baseline results for XGBoost and SVM were lower in the test set in Table 6. Since the only aspect was how adjacent predictions can be used to smooth predictions of individual units, the results show measurable gain over the performance of individual models. Interestingly, the best linear combination was a simple addition with a uniform 0.1 weight for the confidence score of neighbouring units (see Table 7).

Table 6 Experiments on "important" features (imp. Feat.) selected by XGBoost		Accuracy	Sensitivity	Specificity	ROC AUC
	ReLU with imp. feat	0.9024	0.9737	0.8007	0.9340
	SVM with imp. feat	0.9451	0.9793	0.8529	0.9451

Table 7 Effect of spatial smoothing

	Accuracy	Sensitivity	Specificity	ROC AUC
XGBoost	0.8533	0.9828	0.6216	0.9235
SVM	0.9263	0.9828	0.8378	0.9101
Spatial smooth. + XGB	0.9263	0.9655	0.8649	0.9748
Spatial smooth. + SVM	0.8947	0.9828	0.7568	0.9739

The results show increased results on both XGBoost and SVM

Continuous maize yield prediction

In the previous section, various ML methods were tested in a binary, above threshold, maize yield setting. This section investigated further on how linear regression and the best performing methods [XGBoost and SVM or as usually referenced support vector regression (SVR)] and the spatial smoothing performed in the case of actual yield prediction in comparison to classification. Since, the problem can be treated as a traditional regression problem where the model output is continuous and the models were evaluated with MAE, relative error (RE, the mean ratio of absolute error and the ground truth) and FVU. It is worth mentioning that outliers and subset errors may shift these measures.

Results in Table 8 show that XGBoost outperformed other methods in mean MAE, RE and FVU while SVM showed slightly lower standard deviation. The results were reported for the best hyper-parameters per method. RBF kernel on standard normalized features with a cost parameter of four was used for SVM. The depth and width of ReLU networks were utilized according to Adam (Kingma & Ba, 2015) for optimization. The best network contained five layers with 50 hidden units per layer, while on the other hand the XGBoost model had 50 trees with a maximal depth of 7. Additionally, spatial smoothing was tested in the case of continuous maize yield prediction. It was concluded, based on Table 9, that spatial smoothing does not affect the prediction as much as in the case of binary classification; however it showed some significant gain in FVU performance especially when applied over SVR.

Table 8 Fivefold cross-validation regression performance of linear regression, ReLU networks, XGBoost and SVR

	No. Exps	Mean absolute error		Relative error		Fraction of variance of unexplained	
		Mean	SD	Mean	SD	Mean	SD
Linear regression, w/o normalization	10	5.8752	0.6825	0.6404	0.0706	28.683	4.4713
ReLU	10	2.4574	0.3094	0.2935	0.0309	8.2085	1.9711
XGBoost	10	0.8999	0.0198	0.1362	0.0045	0.4376	0.0106
SVR	10	0.9320	0.0131	0.1373	0.0022	0.4555	0.0076

Table 9 Regression performance of XGBoost and SVM with and without spatial smoothing evaluated in management zones number 32–63

	Mean absolute error	Relative error	Fraction of variance of unexplained
XGBoost	1.0277	0.1528	0.3741
Spatial smooth. + XGBopt. for MAE	1.0277	0.1528	0.3741
Opt. for RE	1.0410	0.1523	0.3716
Opt. for FVU	1.1772	0.1584	0.3626
SVR	1.2168	0.2044	0.4094
Spatial smooth. + SVR opt. for MAE	1.2037	0.1977	0.3876
Opt. for RE	1.282	0.1900	0.3397
Opt. for FVU	1.425	0.1945	0.3227

The models were trained on the management zones number 0–31

Conclusions

The aim of this study Querywas to compare different models and to develop a model that predicts maize yield according to a spatio-temporal database. The dataset captured weather features, soil parameters and a crop vegetation index. Five ML analysis methods, namely, CP-ANN, XY-F, SKN, XGBoost and SVM were presented. ML was applied to adapt to site-specific data on soil and crop with two classifications of maize yields. In this paper, site-specific maize yields in different years were presented with the yield prediction model. CP-ANN, XY-F, SKN, XGBoost and SVM were used to define the ML models between the maize yield and the influencing factors of yield. The results obtained showed that XGBoost was the best prediction method (in terms of accuracy). The prediction accuracy obtained on the test set was 95.38% for accuracy, 91.3% for sensitivity and 97.62% for specificity. The results obtained showed that XGBoost was very effective in medium–high maize yield prediction. In addition, the XGBoost model was improved at different levels (amount of trees and leaves). The validation results of this study demonstrated that the proposed ML method of the XGBoost model could accurately estimate the spatio-temporal distribution of maize yields by combining soil parameters and environmental variables. Contrary to the available research papers and simulations about XGBoost, this study provided a new predicting method for site-specific maize yield changes in the field. Additionally, a simple method was suggested to increase the quality of the individual unit predictions by smoothing the confidence scores based on the position of the units.

In order to adopt the effect of weather information, soil and others parameters, these variables were ranked for prediction accuracy.

References

Abendroth L.J., Elmore R.W, Boyer M.J., & Marlay S.K. (2011). Corn growth and development. PMR 1009. Iowa State University of Science and Technology, Cooperative Extension Service

Ballabio, D., & Vasighi, M. (2012). A Matlab toolbox for self organizing maps and supervised neural networks learning strategies. *Chemometrics and Intelligent Laboratory Systems, 118*, 24–32. https://doi.org/10.1016/j.chemolab.2012.07.005

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, 36*(2), 192–225. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:10109 33404324

Cao, L., San, X., Zhao, Y., & Chen, G. (2013). The application of the spatio-temporal data mining algorithm in maize yield prediction. *Mathematical and Computer Modelling, 58*, 507–513.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery. pp 785–794

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Daróczy, B., Vaderna, P., & Benczúr, A. (2015). Machine learning based session drop prediction in LTE networks and its SON aspects. *IEEE.* https://doi.org/10.1109/VTCSpring.2015.7145925 81st Vehicular Technology Conference.

Dyrmann, M., Karstoff, H., & Midtiby, H. S. (2016). Plant species classification using deep convolutional neural network. *Biosystems Engineering, 151*, 72–80. https://doi.org/10.1016/j.biosystems eng.2016.08.024

Elgaali, L., & Garcia, L. (2004). Neural Network Modeling of Climate Change Impact on Irrigation Water Supplies in Arkansas River Basin, Hydrology Days, Colorado State University, Fort Collins. Retrieved April 14, 2021, from http://hydrologydays.colostate.edu/Papers_2004/Elgali_paper.pdf

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., & Xiang, Y. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management, 164*, 102–111. https://doi.org/10.1016/j.enconman.2018.02.087

Folberth, C., Baklanov, A., Balkovic, J., Skalsky, R., Khabarov, N., & Obersteiner, M. (2019). Spatiotemporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and Forest Meteorology, 264*, 1–15. https://doi.org/10.1016/j.agrformet.2018.09.021

Fortin, J. G., Anctil, F., Parent, L. -É., & Bolinder, M. A. (2011). Site-specific early season potato yield forecast by neural network in Eastern Canada. *Precision Agriculture, 12*, 905–923. https://doi.org/10.1007/s11119-011-9233-6

Frausto-Solis, J., Gonzalez-Sanchez, A., & Larre, M. (2009). A new method for optimal cropping pattern. In A. H. Aguirre, R. M. Borja, & C. A. R. García (Eds.), *MICAI 2009: Advances in artificial intelligence.* Springer.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. In P. Vitányi (Ed.), *Computational learning theory.* Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics. *29*(5), 1189–1232. Retrieved October 17, 2020, from https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451

Gasteiger, J., & Zupan, J. (1993). Neural networks in chemistry. *Angewandte Chemie, 32*, 503–527. https://doi.org/10.1002/anie.199305031

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, 11*, 13–23. https://doi.org/10.1016/j.cliser.2018.06.003

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Kerepesi, Cs., Daróczy, B., Sturm, Á., Vellai, T., & Benczúr, A. (2018). Prediction and characterization of human ageing-related proteins by using machine learning. *Scientific Reports, 8*(1), 1–13. https://doi.org/10.1038/s41598-018-22240-w

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and Electronics in Agriculture, 153*, 213–225. https://doi.org/10.1016/j.compag.2018.07.016

Kingma, D. P. & Ba, J. L (2015): ADAM: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*. Retrieved April 7, 2021, from https://arxiv.org/pdf/1412.6980.pdf

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (2001). *Self-organizing maps*. Springer.

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. In *Proceedings 15th International Conference on Precision Agriculture*. Retrieved April 14, 2021, from https://www.ispag.org/proceedings/?action=abstract&id=4974&title=A+Case+Study+Comparing+Machine+Learning+and+Vegetation+Indices+for+Assessing+Corn+Nitrogen+Status+in+an+Agricultural+Field+in+Minnesota%0D%0A&search=years

Liu, J., Goering, C. E., & Tian, L. (2001). A neural networks for setting target corn yields. *Transactions of the American Society of Agricultural Engineers, 44*(3), 705–713. https://doi.org/10.13031/2013.6097

Longchamps, L., Tremblay, N., & Panneton, B. (2018). Observational studies in agriculture: paradigm shift required. In *Proceedings 15th International Conference on Precision Agriculture*. Retrieved April 14, 2021, from https://www.ispag.org/proceedings/?action=abstract&id=5436&title=Observational+Studies+in+Agriculture%3A+Paradigm+Shift+Required&search=years

Maeda, Y., Goyodani, T., Nishiuchi, S., & Kita, E. (2018). Yield prediction of paddy rice with machine learning. The 24th International Conference on Parallel and Distributed Processing Techniques and Applications (pp. 361–365). CSREA PRESS.

Melssen, W., Wehrens, R., & Buydens, L. (2006). Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems, 83*(2), 99–113. https://doi.org/10.1016/j.chemolab.2006.02.003

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*(2), 117–135. https://doi.org/10.1007/s11119-006-9004-y

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

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*, 1–10. https://doi.org/10.3389/fpls.2016.01419

Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems, 27*, 1–17.

Mouazen, A. M., Maleki, M. R., Cockx, L., Van Meirvenne, M., Van Holm, L. H. J., Merckx, R., De Baerdemaeker, J., & Ramon, H. (2009). Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorous measured using an on-line visible and near infrared sensor. *Soil and Tillage Research, 103*(1), 144–152. https://doi.org/10.1016/j.still.2008.10.006

Mouazen, A. M., Maleki, M. R., De Baerdamaeker, J., & Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS-NIR sensor. *Soil and Tillage Research, 93*, 13–27. https://doi.org/10.1016/j.still.2006.03.009

Neményi, M., Mesterházi, P. Á., & Milics, G. (2006). An application of tillage force mapping as a cropping management tool. *Biosystems Engineering, 94*(3), 351–357. https://doi.org/10.1016/j.biosystemseng.2006.04.005

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 Proceedings of the 9th European Conference on Precision Agriculture. Wageningen Academic Publishers

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. https://doi.org/10.1556/0806.44.2016.056

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture, 121*, 57–65. https://doi.org/10.1016/j.compag.2015.11.018

Ruß, G. (2009). Data mining for agricultural yield data: A comparison of regression models. In P. Perner (Ed.), *Advances in data mining. Applications and theoretical aspects: 9th Industrial Conference* (pp. 24–37). Springer.

Ruß, G., & Kruse, R. (2010). Feature selection for wheat yield prediction. In M. Bramer, R. Ellis, & M. Petridis (Eds.), *Research and development in intelligent systems XXVI: Incorporating applications and innovations in intelligent systems XVII* (pp. 465–478). Springer.

Safa, B., Khalili, A., Teshnehlab, M., & Liaghat, A. (2004). Artificial neural networks application to predict wheat yield using climatic data. In *Proceedings of the 20th International Conference on Interactive Information on Processing Systems for Meteorology, Oceanography and Hydrology*. American Meteorological Society. (pp. 1–39). Retrieved April 19, 2021, from https://ams.confex.com/ams/pdfpapers/66796.pdf

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

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

USDA. ND. (2020). *Natural Resources Conservation Service*. Soil classification. Retrieved October 17, 2020, from https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/class/

Wang, Y., & Witten, I. (1997). Inducing model trees for continuous classes. In M. van Someresn, & G. Widmer (Eds.), *Proceedings of the 9th European Conference on Machine Learning.Prague*. (pp. 128–137). Retrieved April 5, 2021, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.9768

Zhang, L., Zhang, J., Kyei-Boahen, S., & Zhang, M. (2010). Simulation and prediction of soybean growth and development under field conditions. *American-Eurasian Journal Agricultural and Environmental Science., 7*(4), 374–385.

Authors and Affiliations

A. Nyéki[1] · **C. Kerepesi[2]** · **B. Daróczy[3]** · **A. Benczúr[4]** · **G. Milics[1]** · **J. Nagy[5]** · **E. Harsányi[5]** · **A. J. Kovács[1]** · **M. Neményi[1]**

[1]   Department of Biosystems and Food Engineering, Faculty of Agricultural and Food Sciences, Széchenyi István University, Vár square 2, Mosonmagyaróvár 9200, Hungary

[2]   Gladyshev Lab, Harvard Medical School, Avenue Louis Parker 77, Boston, MA 02115, USA

[3]   Université Catholique de Louvain, ICTEAM, B-1348 Louvain-la-Neuve, Belgium

[4]   Hungarian Academy of Sciences, Institute for Computer Science and Control, Kende Str. 13-17, Budapest 1111, Hungary

[5]   Faculty of Agricultural and Food Sciences and Environmental Management, Institute of Land Utilization, Technology and Regional Development, University of Debrecen, Böszörményi Str. 138, Debrecen 4032, Hungary