A given model of yield forecasting using an artificial neural network connects the wheat crop with the amount of productive moisture in the soil, soil fertility, weather, and factors in the presence of pests, diseases, and weeds. The difficulty of creating a yield forecast system is in the correct choice of predictors that have the greatest impact on yield.

To build the model, moisture in the 100 cm layer of the soil, the content of nitrogen, phosphorus, humus, and soil acidity in the soil were used as input parameters. The amount of precipitation over 4 months, the average air temperature for the same period, as well as the presence of diseases, pests, and weeds were also taken into consideration. Data on 13 districts of the North Kazakhstan region in the period from 2008 to 2017 were used. The output parameter was the yield of spring wheat over the same time period.

The relative importance of input variables in relation to the output variable was used to determine the weight values of input variables.

An artificial neural network of error backpropagation was used as a method. The advantage of this method is that the quality of the forecast increases with a large amount of training data, as well as the ability to model nonlinear relationships between different data sources.

After training the artificial neural network and obtaining predictive data, good results were achieved for predicting wheat yields ($p=0.52$, mean absolute error in percentage (MAPE) = 12.02 %, root mean square error (RMSE) = 3.368).

Thus, it is assumed that the developed model for forecasting wheat yields based on data can be easily adapted for other crops and places and will allow the adoption of the right strategies to ensure food security

Keywords: yield forecasting, artificial neural network, wheat yield, independent variables

UDC: 004.032.26

DOI: 10.15587/1729-4061.2022.259653

APPLICATION OF ARTIFICIAL NEURAL NETWORK FOR WHEAT YIELD FORECASTING

Gailya Aubakirova
Corresponding author
Doctorant*
E-mail: gailyaaibatova@gmail.com
Victor Ivel
Doctor of Sciences in Engineering, Professor*
Yuliya Gerassimova
Candidate of Engineering Sciences*
Sayat Moldakhmetov
PhD*
Pavel Petrov
PhD*
*Department of Energetic and Radioelectronics
Manash Kozybayev North Kazakhstan University
Pushkin str., 86, Petropavlovsk,
Republic of Kazakhstan, 150000

Received date 31.05.2022
Accepted date 20.06.2022
Published date 29.06.2022

How to Cite: *Aubakirova, G., Ivel, V., Gerassimova, Y., Moldakhmetov, S., Petrov, P. (2022). Application of artificial neural network for wheat yield forecasting. Eastern-European Journal of Enterprise Technologies, 3 (4 (117)), 31–39. doi: https://doi.org/10.15587/1729-4061.2022.259653*

1. Introduction

Yield forecasting is an important but complex problem necessary for sustainable intensification and efficient use of natural resources [1, 2]. Advance and accurate forecasting of yields has been and remains an urgent problem for any state because the effectiveness of a long agri-food chain depends on the accuracy of the forecast. Farmers, agronomists, and politicians participate in this chain and rely on yield forecasts given by experts in their activities [3–10].

The yield of different crops depends on environmental conditions, management actions, and many other specific parameters [11]. Various approaches are used in predicting yields, the main ones being expert estimates (for example, interviews and field studies), statistical models, and models based on processes. Interviews with farmers tend to provide highly subjective yield expectations towards the end of the season [12, 13]. Field studies with crop pruning provide an objective assessment of yields prior to harvesting. Statistical models use various methods (regression, Bayesian approaches, machine learning methods) to construct regression dependences between various statistical data obtained using remote and meteorological observations [7, 14]. One of the most common methods of forecasting yields is a statistical model based on agrometeorological data. This model is relatively easy to develop and use. However, one of the main disadvantages of this method is that numerous environmental factors are nonlinear, i.e. can have large deviations from the average values. Such factors, for example, include air temperature, and the amount of precipitation, they have the greatest impact on the formation of wheat yields. That is why it is necessary to move away from traditional methods in favor of more accurate forecasting methods.

The most suitable alternative is models based on artificial neural networks.

Crop simulation models produce not only end-of-season yields but also yield distribution based on crop genotypes, soil condition, typical management techniques, and seasonal weather. These data are obtained on the basis of historical climate or weather forecasts or by assimilation of information obtained by remote sensing [15–17].

The advantage of models using neural networks is the high accuracy of the forecast and the possibility of increasing yields. Algorithms for building and training a neural network are based on functions that determine the dependence of features and predictors on output data; in this case, it is yield. Artificial neural networks have some distinct advantages over traditional models. Thus, they can simulate nonlinear

relationships between multiple data sources [18]; their performance usually improves with a large amount of training data [19]. Therefore, research on the development of a model for predicting the yield of grain crops, using an artificial neural network, is relevant.

2. Literature review and problem statement

Paper [20] predicted the yield of winter wheat in Guangzhou Province, China. The Leaf Area Index (LAI) and the Vegetation Temperature Condition Index (VTCI) were used as predictors of yield. These two indices are closely related to plant growth and water scarcity, and were used to indicate crop growth conditions and estimate yields in the Guanzhong Plain, China. The LAI and VTCI indices were used as variables of the neural backpropagation network (BP) and the neural network of the improved particle swarm optimization algorithm IPSO-BP. In the paper, the authors compared the results of these two methods. As a result of that study, the neural network of the improved particle swarm optimization algorithm showed better results.

Study [21] predicted the yield of winter wheat, rapeseed, corn, and sunflower in Hungary. One of the predictors was the Normalized difference vegetation index (NDVI). The use of a normalized vegetation index in forecast models is associated with some difficulties. One of the disadvantages of using this predictor is that the index does not have feedback (open structure). And this makes it susceptible to numerous errors and uncertainties associated with changing weather conditions and the background of vegetation cover.

Study [22] demonstrated a correlation between the NDVI vegetation index obtained at the vegetative, reproductive, and maturity stages and the final yield of maize, both in rainfed and irrigated treatments. The main advantage of that study is that the inclusion of remote sensing information in the statistical model increases the accuracy of forecasts. Prospects for the use of satellite information in yield forecasting models are limited by the quality of remote sensing data (i.e., the presence of clouds in the images). Since most cereals are non-irrigated, except for rice [23], therefore, the growing season and the rainy season coincide. Getting a series of cloudless images can be tricky; and it may not be possible to obtain good quality images on the forecasting date, when derived vegetation indices are best correlated with final yields.

Work [24] used combinations of different climate variables, including minimum and maximum temperature, relative humidity in the morning and evening, and rainfall, as predictors to predict the yield of maize, wheat, and rice. The model used data over a long time period (1980–2009) and for different parts of India. The model was tested on 2010 and 2011 yield data, and the deviation between predicted and actual yields was less than 15 %, indicating satisfactory results.

Paper [25] presents a model based on statistical regression. In the paper, the dependence of corn yield on weather factors is established. Daily data on air temperature (maximum and minimum), relative soil moisture (morning and evening), and rainfall were used as dependent variables. Statistical models based on agrometeorological data rely on the use of weather and/or agronomic variables as independent variables to predict yields. Data for the period 1985–2012 were used to develop the forecast model, and data for the remaining three years (2013–2015) were used to validate the models. The advantage of that model is the simplicity and availability of the data employed. The use of additional input parameters, for example, soil fertility indicators, such as nitrogen, phosphorus, etc., would improve this predictive model.

The authors of [26] predicted the yield of corn and soybeans 2 months before harvest in the main producing countries of the world using conventional regression models of the least squares with temperature, precipitation. To make a forecast of yield, regression equations are used depending on the yield indicators obtained for previous years in the study region or similar regions. In general, statistical regression-based models are usually simple and easy to understand and require less parameter adjustment, so they are widely used. As the quantity and quality of the observed data increase, regression-based statistical models usually produce satisfactory results. This is especially observed in conditions characterized by significant interannual variations in yield due to several factors. However, models based on statistical regression are also not without problems. Because the relationship between dependent and independent variables is not linear, such models do not work well under conditions with frequent extreme climatic conditions. Moreover, the same meteorological factors occurring at different stages of growth, namely factors specific to the growth stage, can affect yields in different ways. For example, heat or drought events occurring during the flowering stage can cause greater crop losses than during the growing stages.

According to [27], artificial neural networks are one of the best tools for obtaining information from inaccurate and nonlinear data. An additional advantage of artificial neural networks is the ability to use qualitative variables without the need to pre-encode them, as is the case with conventional statistical methods [28].

Many studies have demonstrated the advantage of ANNs over multiple linear regression in yield prediction. Paper [29] analyzed the possibility of using ANN and multiple linear regression to predict the yield of barley grown in Ardabil, Iran. The study used a multilayer perceptron with three input neurons, 15 neurons in the hidden layer, and one output neuron. Based on the average absolute error, the authors found that the ANN is more accurate than multiple linear regression.

All this suggests that it is advisable to conduct a study on forecasting yields using the method of an artificial neural network, taking as a basis the predictors that have the greatest impact on the result of forecasting.

3. The aim and objectives of the study

The purpose of this study is to build a model for forecasting wheat yields using an artificial neural network, using meteorological, agrochemical, and phytosanitary data. This will make it possible to predict and improve the accuracy of the forecast of wheat yields under conditions of risky farming.

To accomplish the aim, the following tasks have been set:

– to identify predictors that have the greatest impact on wheat yield;

– to select the settings of the neural network for building predictive models;

– to calculate the accuracy of the forecast model using the mean absolute error in percentage MAPE and the RMSE root mean square error.

4. The study materials and methods

The North Kazakhstan region is one of the leading areas of agricultural production in the Republic of Kazakhstan. The region occupies 3.6 % of the territory of the Republic of Kazakhstan while it gives 16 % of agricultural products or 25 % of the grain harvest. The region is in the north of the Republic of Kazakhstan, within the southern outskirts of the West Siberian Plain. The area of the region is 98 thousand km^2, of which farmland occupies 58.8 thousand km^2, which is 60 % of the territory of the region [30, 31].

The region includes 13 districts: Kyzylzhar, M. Zhumabayeva and Akkayinsky, G. Musrepova, Aiyrtau, Yesil, Mamlyut, Tayynshinsky, Timiryazevsky, Zhambylsky, Akzharsky, Shalakyn, and Ualikhanovsky. To develop the ANN, data from the above 13 districts of the North Kazakhstan region were used.

This study suggests that wheat yields vary within and between areas depending on environmental factors, soil fertility and moisture, and the presence of plant diseases, pests, and weeds. Functionally, this is expressed as:

$$Y=f(AF, WF, SF), \quad (1)$$

Y is the wheat harvest (kg/ha); AF is the agrochemical factors; WF is the meteorological factors; SF is the phytosanitary factors.

During the study, a large number of different factors were considered, from which 16 names of input data for each field were selected (Table 1). The input data were divided into 3 groups of factors: agrochemical, meteorological, and phytosanitary. Agrochemical factors affecting the yield of wheat include the content of nitrogen and phosphorus in the soil, as well as the percentage of humus and soil moisture. Meteorological factors include rainfall and air temperature. Phytosanitary factors include the presence of weeds, pests, and plant diseases. Inputs and data preparation are described below.

The first type of factor is agrochemical. Wheat's requirements for the required soil moisture depend on the stage of development. Ears of wheat reach the peak of water consumption during the filling of the ear; it is at this time that the plants are more sensitive to drought. During this period, large yield losses may occur on soils with low water retention capacity [32]. The amount of precipitation and the supply of productive moisture in the soil is the main reason for fluctuations in the annual yield in this study.

The data of the agrochemical factor used the results of measurements made by hydrometeorological stations and posts in order to obtain data on the supply of productive moisture in the 100 cm layer of the soil as input indicators of the ANN.

Four factors of soil fertility were also included in the ANN: soil pH, phosphorus, nitrogen, as well as the percentage of humus in the soil (Table 1). Soil fertility coefficients were obtained from the analysis of soil samples taken from fields. The data were obtained for 10 years (from 2008 to 2017). Average values were calculated for further use in the experiment.

The second type of factor is meteorological. Daily climatic observational data (2008–2017) for 29 sites, including precipitation, as well as the minimum and maximum air temperature, were obtained from hydrometeorological stations and posts of the Branch of RGP «Kazhydromet» in the North Kazakhstan region.

Table 1

Layers of information used to build ANN

Input layer				Output layer
Type	Factor	Dimensionality	Designation	
Agrochemical factors (*AF*)	Soil moisture in 100 cm soil layer	%	X_1	Yield (y)
	Nitrogen content	mg/kg	X_2	
	Phosphorus content	mg/kg	X_3	
	Soil pH (Ph)	mg/kg	X_4	
	humus	%	X_5	
Meteorological factors (*WF*)	Rainfall for May	mm	X_6	
	Rainfall for June	mm	X_7	
	Amount of precipitation in July	mm	X_8	
	Rainfall for August	mm	X_9	
	average air temperature for May	°C	X_{10}	
	average air temperature for June	°C	X_{11}	
	average air temperature for July	°C	X_{12}	
	average temperature for August	°C	X_{13}	
Phytosanitary factors (*SF*)	weeds	0–4 a.u.	X_{14}	
	pests	0–4 a.u.	X_{15}	
	disease	0–4 a.u.	X_{16}	

Precipitation has been divided into four periods and is referred to as May, June, July, and August.

The third type of factor is phytosanitary. In the north of Kazakhstan, such common pests as bread striped flea, cross non-herd locust, gray grain armyworm, etc. cause huge damage to plants and crops of spring wheat grain.

Under the conditions of the modern agricultural economy, the contamination of fields does not decrease but, on the contrary, increases. The reason for this is the use of zero technologies, the presence of waste land, etc. There is a change in the species composition of weeds, which is a consequence of the use of herbicides of one group. Now, almost all crops of grain crops are littered, and more than half to a moderate and strong extent. The danger of malicious weeds is especially high: milkweed, thistle (species), field bindweed, wormwood (species). As well as fescue, bristles, creeping wheatgrass, white maria, shchiritsa, creeping mustard, which belongs to the quarantine weeds for many countries importing grain from Kazakhstan. Weed prevalence has been observed in all fields and has a rating ranging from zero (density=0) to very high (density=4).

Great damage to durum wheat crops in the North of Kazakhstan is caused by root rot, leading to thinning of seedlings, weakening of plants, and a decrease in their productivity, as well as brown, yellow, stem rust, leaf spotting (septoriosis, helminthosporiosis, etc.) [33]. Data on the number of these pests, diseases, and weeds were used as input parameters for NN because they significantly affect the damage to wheat yields.

Weed prevalence has been observed in all fields and has a rating ranging from zero (density=0) to very high (density=4).

To build and train a neural network, the Neural Network Toolbox of the MATLAB software environment (The Mathworks, USA) was used.

The output parameter of the ANN was the yield of spring wheat (Table 2).

Table 2

Yield of spring wheat in the North Kazakhstan region

Year	North Kazakhstan region districts												
	Aiyr-tau	Akzharsky	Zhuma-baeva	Esil-sky	Zham-byl	Kyzylz-har	Mam-lyutsky	Shala-kyn	Akkayin-sky	Taiyn-shinsky	Timirya-zevsky	Ualikha-novskiy	Musre-pova
2008	13.0	5.9	14.9	13.0	13.7	14.2	13.5	11.4	13.9	12.5	12.7	6.5	13.4
2009	14.1	13.0	17.1	15.7	13.6	16.8	15.7	13.1	17.6	13.9	14.4	13.1	12.7
2010	9.2	12.0	11.8	10.0	7.6	12.6	12.5	8.9	10.6	9.3	7.6	9.6	7.2
2011	23.0	15.8	21.9	23.4	19.4	22.8	21.0	19.1	21.7	19.1	25.5	14.6	22.5
2012	15.5	7.7	13.7	16.1	12.7	16.5	15.9	8.7	14.0	10.1	9.5	7.4	8.5
2013	12.6	9.1	15.2	11.5	11.0	13.7	11.2	9.1	12.8	12.5	9.4	12.3	14.5
2014	13.8	12.9	6.7	13.1	13.6	16.0	13.6	11.9	14.1	13.4	12.3	12.5	13.7
2015	14.8	15.0	18.3	17.7	17.3	18.7	15.9	13.2	18.2	12.9	14.4	15.1	14.5
2016	14.1	13.2	17.4	16.4	17.4	16.6	14.3	14.1	18.2	11.8	13.8	12.7	14.8
2017	16.8	12.6	19.9	18	18.8	17.4	17.4	15.9	19.8	14.6	17.3	11.6	16.2
2018	13.1	14.2	16.7	17	19.8	16.9	14.6	14.3	16.9	13.9	15.6	15.3	14.6
2019	12.7	12.4	13.5	17.3	15.4	18.2	17.7	14.1	16.3	14.2	11.8	14.5	13.2
2020	14.0	13.2	14.0	18.1	14.3	18.3	17.9	15.5	16.1	12.4	14.5	10.2	14.0

Wheat yield data for 2008–2020 (Table 2) are taken from the official website of the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan [34]. The choice of such a period is due to the availability of all necessary statistical data.

5. Results of studying an artificial neural network for forecasting wheat yields

5. 1. Selection of predictors for building a predictive model of yield

To build and train the neural network, a trial and error method was used to select the optimal parameter value (weights of connections) that would give the most accurate results.

The input is 16 neurons representing the input data. As input data, 3 groups of factors were used: agrochemical, meteorological, and phytosanitary, including 16 parameters that affect yield. Agrochemical parameters include soil moisture in the meter layer, the content of phosphorus, nitrogen, and pH of the soil, as well as the percentage of humus in the soil. Meteorological parameters include the average air temperature for 4 months (May, June, July, August), the average amount of precipitation for the same 4 months. Phytosanitary parameters include the weediness of crops, the presence of diseases and pests.

5. 2. Selection of artificial neural network settings

There are several types of artificial neural network (ANN) models, distinguished by the way nodes are connected, the methods of calculating weights, the number of nodes in hidden layers, and the type of transfer function between layers. The architecture determines how weights are interconnected in the network and what training rules can be used [35].

The choice of learning rules is important because it affects which input function, transfer function, and parameters will be used for the ANN model. The backpropagation algorithm is one of the methods of training multilayer neural networks of direct propagation. Training by the backpropagation algorithm involves two passes through all layers of the network: forward and reverse. In a direct pass, the input vector is fed to the input layer of the neural network, and then propagates through the network from layer to layer. As a result, a set of output signals is generated, which is the actual reaction of the network to this input image. During a direct passage, all synaptic weights of the network are fixed. During the return pass, all synaptic weights are adjusted according to the error correction rule, namely: the actual output of the network is subtracted from the desired one, resulting in an error signal. This signal subsequently propagates through the network in the opposite direction of synaptic connections [36–40].

A reverse back propagation neural network has been proposed [41], and is the most widely used algorithm for trainer-assisted learning in multi-level, direct-link networks. Its main idea is to revise the weights and thresholds of the network by backpropagation to minimize the error between the actual output value and the expected output value. Neural networks with at least one hidden layer are necessary and sufficient to approximate arbitrary nonlinear functions. In practice, neural networks with one or two hidden layers, that is, three-layer or four-layer perceptrons (including input and output layers), are usually used. The topology of the neural network of error backpropagation with one hidden layer is shown in Fig. 1. In the training of a neural network of error backpropagation, there are two processes: the direct propagation of the input signal and the reverse propagation of the error. In direct propagation, the input signal acts on the output node through a hidden layer to generate an output signal. The state of the neuron in each layer affects only the state of the neuron in the next layer. If the actual result does not match what one expects, the error is canceled. The inverse propagation of the error is to pass the output error back to the input layer through the hidden layer and minimize the error signal by changing the weights of each layer of neurons.

Initializing the network. There are m input neurons, n hidden neurons, and one output neuron.

The first step in training is to initialize the weight parameters w and usually small random values are offered. Fig. 1 shows the weight of the connection between the 16th node in the input layer and the n node in the hidden layer, h_n is the

result of the n-th node in the hidden layer. W_n is a weighted value between the n-th node in the hidden layer and the output layer. Y is the output of the neuron in the output layer. The calculation was performed according to the formulas:

$$h_k = \int\left(\sum_{j=1}^{m} w_{jn} a_j - \theta_n\right), \quad (2)$$

$$y = \int\left(\sum_{n=1}^{m} w_n h_n - \theta\right), \quad (3)$$

where θ_n is the displacement of the nth node in the hidden layer, and θ is the displacement of the neuron in the output layer. Displacements are assigned by random values from 0 to 1 before the direct propagation of working signal.

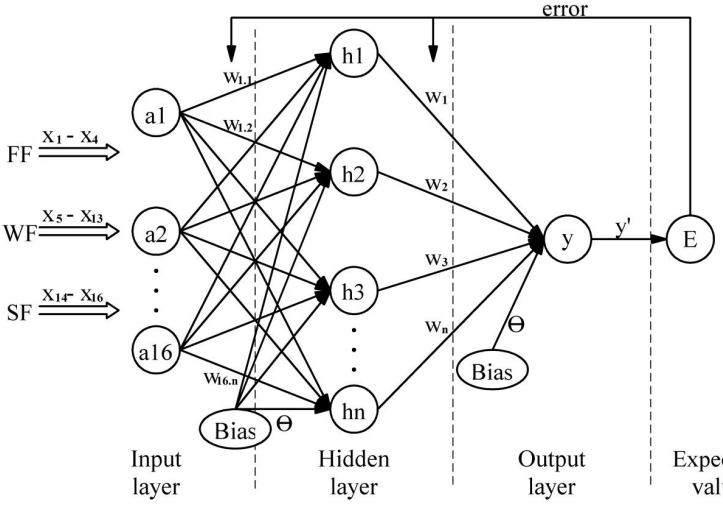

Fig. 1. Structure of an artificial neural network of backpropagation

The most commonly used latent function of neuronal activation is the sigmoid, determined by the formula:

$$\int(x) = \frac{1}{1+e^{-x}}. \quad (4)$$

Our study used a standard fully connected backpropagation neural network with direct connection (BPNN); its structural diagram is shown in Fig. 2. Sigmoid function was used as input. Details of the implementation of a neural network with direct connection and error backpropagation are described in [42]. The architecture of the neural network with direct connection and error backpropagation used in this study was as follows:

– the number of layers=3 (input, hidden, and output);

– the number of neurons in the hidden layer=from 1 to 3;

– the type of activation functions= sigmoid for the hidden layer, linear for the output layer;

– the number of nodes in the input layer=16 (Table 1);

– the number of nodes in the output layer=1;

– the type of network error=root mean square error.

Fig. 2 shows a neural network model with direct communication and error backpropagation. Hidden Layer is a layer to the input of which signals are given, after which they are multiplied by weights (each signal – by its own weight) (in Fig. 2, indicated by the letter w). To this amount is added the displacement of the neuron (in Fig. 2, indicated by the letter b) and then entered on the summation block. The summing block algebraically adds the weighted inputs, creating an output. The resulting signal is transformed by the activation function of the neuron, which forms the output signal.

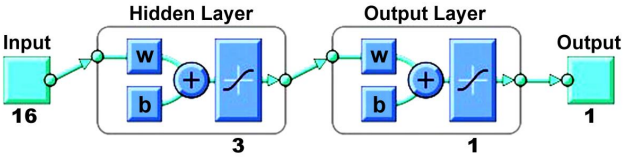

Fig. 2. Model of an artificial neural network in the MATLAB environment

5. 3. Calculating the accuracy of a predictive model

An important element of predictive modeling is an accurate assessment of the correctness of the functionality of the model. For this purpose, retrospective forecast quality indices are used. One of the most used forecast error indicators is the Average Absolute Error Percentage (MAPE), which is calculated from (5) [43–49]:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i'}{y_i}\right| \times 100\,\%. \quad (5)$$

MAPE measures error as a percentage and indicates the average percentage deviation between the predicted value and the actual implementation. Papers [50, 51] reported that if the MAPE value is below 10 %, the degree of model compliance is ideal; if it falls within the range of 10–20 %, the quality of the model is good. In the range of 20 % to 30 %, an error rate is acceptable, while an error greater than 30 % is considered a bad result and, therefore, the model should be rejected.

According to [52], when the real value is close to or equal to zero, MAPE provides infinite or undefined values, which is considered its significant disadvantage. Therefore, a combination of MAPE and root mean square error (RMSE) is used to test in detail the evaluation of the effectiveness of the forecasting model [53]. RMSE indicates that the observed data point absolutely matches the predicted values. RMSE is defined according to (6) as the second-order root of the mean square of all errors [54, 55].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{n}}. \quad (6)$$

The lower the MAPE and RMSE values, the higher the accuracy of the resulting forecast model [56]. Other measures to identify an error include mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), relative square root error (RRSE), and others [57–59].

The analysis of the forecasting results obtained was carried out by calculating the average absolute error in percentage MAPE and the RMSE root mean square error. The results of the calculation of errors and the correlation coefficient between the predicted neural network yield and actual data are given in Table 3. The artificial neural network was trained on the data of the period from 2007 to 2016 and tested on the data of 2017–2020. The table also shows the results of calculating the average absolute error in percentage

MAPE and the RMSE mean square error based on the results of training and testing of the artificial neural network.

For clarity, linear regression equations for test results are also calculated. Fig. 3 shows the results of the projected yield in comparison with the actual one. The data used for training are indicated by «▲», and for testing – «●».

Table 3

Results of the analysis of the performance of the developed NN

Stage	Year	Iteration	Error		Correlation coefficient	Linear regression
			MAPE	RMSE		
Training	2007–2016	60	8.76	2.241	0.689	$y=0.86x+2.4$
Testing	2017–2020	10	12.02	3.368	0.534	$y=0.59x+6.8$

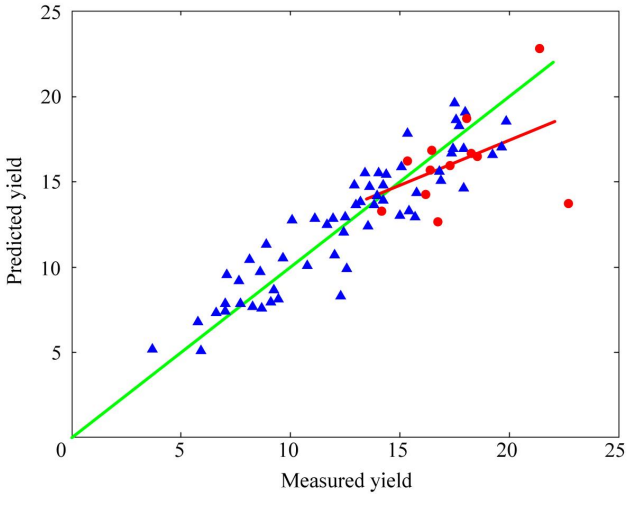

Fig. 3. Results of predicting the yield of an artificial neural network in comparison with the actual data

A study of the degree of influence of factors on the yield forecast was also conducted. To do this, the ANN was tested without taking into consideration one factor or another and the change in the MAPE and RMSE error was considered. When testing an artificial neural network, the data of selected factors were fed to the input and the change in the MAPE and RMSE error was considered, in order to determine which of them have the greatest impact on wheat yield. Table 4 gives the results of the MAPE and RMSE error for this study.

Table 4

Influence of factors on the accuracy of yield forecasting

Missing factor	MAPE	RMSE	Effect
Soil moisture	8.87	2.269	medium
Soil acidity (pH)	8.79	2.248	low
Humus content	8.81	2.253	medium
Precipitation	9.13	2.335	large
Weeds	8.81	2.253	low
Temperature	9.09	2.325	large
NDVI	8.76	2.241	no effect

The results obtained show that the developed neural network has successfully mastered the basic dependences of yield on input data. With test forecasting on the created neural network, a yield forecast with a MAPE error of no more than 12 % is obtained, which is a confirmation of the adequacy of the forecast issued.

6. Discussion of results of studying the artificial neural network for forecasting wheat yields

Table 4 demonstrates that not all factors have the same effect on wheat yields. The most significant of them were combined into 3 groups of factors: agrochemical, meteorological, and phytosanitary.

To train the developed neural network (Fig. 2), yield data from 2007 to 2016 were used. Thus, a sample for 10 iterations and 14 epochs was used for training. The training was carried out using the method of error backpropagation. At the same time, the created neural network learns quickly enough. Depending on the power of the computer, the training time is from 10 to 25 seconds.

Testing the already trained neural network implied the ability to predict yields in the years 2017 to 2021. The calculated average absolute error in percentage MAPE and RMSE root mean square error, given in Table 3, allow us to talk about good predictive results of the model.

In works [24, 25], only climatic factors were used as predictors, such as minimum and maximum temperature, relative soil moisture in the morning and evening, and the amount of precipitation. Taking into consideration temperature and precipitation in models predicting crop yields is justified because these factors strongly influence the growth and development of crops [60]. The distribution of temperature during the growing season has the greatest impact on plant productivity. However, if the plant is properly supplied with water, this impact is reduced [61]. Phytosanitary indicators are also very important because the presence of diseases and pests reduces both the quality of grain and yield indicators in general. Table 4 demonstrates that the greatest influence on wheat yield is exerted by the amount of precipitation and air temperature. Soil moisture and the percentage of humus in the soil have a significant impact on yields. The least influence is exerted by the pH of the soil and the presence of soil debris. Satellite images of the NDVI vegetation index did not have a significant impact on yields in this model. This is due to the poor quality of the images, as the ripening period of wheat is often accompanied by poor weather conditions in the study area. That is why it was decided to exclude this indicator in the forecast model. In our study, air temperature, rainfall, nitrogen and phosphorus content in the soil, soil pH, soil moisture up to 100 cm of the layer, the presence of pests, diseases, and weeds were used as predictors.

In contrast, the use of NDVI remote sensing data in an artificial neural network did not lead to significant improvements in the model. In this study, the NDVI vegetation index does not rank high as an important variable influencing the prediction of wheat yields. This may be due to the NDVI index data used in the study. NDVI values with a spatial resolution of 500 m did not have a sufficiently high resolution to reflect the vegetation conditions of specific test sites. For yield predictions on a larger scale, applying this information to systems is likely to contribute more to the accuracy of the model. In addition, it is possible that the low quality of the images was negatively affected due to increased cloudiness and, therefore, this predictor was not used in the construction of this artificial neural network.

A given forecasting system showed good results in forecasting wheat yields. However, this method requires a large amount of data, from a variety of sources including soil

indicators, climate and phytosanitary data, and may be limited due to human error or failure of measuring devices. One of the disadvantages of this method is that to improve the accuracy of the yield forecast, it is necessary to use a large amount of training data, ideally for 10 years or more of time.

In the future, our research team plans to develop a wireless system for remote monitoring of air and soil temperature, soil moisture at different depths and acidity directly in the sown field. In addition, in the future, the neural network can be retrained by increasing the number of training samples to develop a more accurate assessment model.

7. Conclusions

1. To build a model for forecasting yields, predictors were selected that have the greatest impact on wheat yields. These data are divided into 3 groups of factors: agrochemical, meteorological, and phytosanitary and include 16 indicators.

2. In this study, an artificial neural network with the number of neurons in the covered layer from 1 to 3, with a direct connection and error backpropagation, was used to build a predictive model of yield. Wheat yield indicators from 2007 to 2016 were used as training data, and data from 2017–2020 were used for testing. The training of the neural network was carried out by the method of error backpropagation. The choice of this method is due to good resistance to the influence of external factors. An additional advantage of the method of error backpropagation is high efficiency with sufficient simplicity of implementation, although the learning process can take quite a long time.

3. As a result of our study, the average absolute error in percentage (MAPE) was calculated, which was 12.02 %. The root mean square error (RMSE), which was 3.368, was also calculated. Since the result of the MAPE calculation is in the range of 10–20 %, this indicates good results of the forecast.

Acknowledgments

This study was carried out with the financial support of the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (grant No AP13268732).

References

1. Phalan, B., Green, R., Balmford, A. (2014). Closing yield gaps: perils and possibilities for biodiversity conservation. Philosophical Transactions of the Royal Society B: Biological Sciences, 369 (1639), 20120285. doi: https://doi.org/10.1098/rstb.2012.0285
2. Tilman, D., Balzer, C., Hill, J., Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences, 108 (50), 20260–20264. doi: https://doi.org/10.1073/pnas.1116437108
3. Basso, B., Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. Advances in Agronomy, 201–255. doi: https://doi.org/10.1016/bs.agron.2018.11.002
4. Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van der Velde, M., Makowski, D. (2018). Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. Nature Communications, 9 (1). doi: https://doi.org/10.1038/s41467-018-04087-x
5. Funk, C., Shukla, S., Thiaw, W. M., Rowland, J., Hoell, A., McNally, A. et. al. (2019). Recognizing the Famine Early Warning Systems Network: Over 30 Years of Drought Early Warning Science Advances and Partnerships Promoting Global Food Security. Bulletin of the American Meteorological Society, 100 (6), 1011–1027. doi: https://doi.org/10.1175/bams-d-17-0233.1
6. Headey, D. (2011). Rethinking the global food crisis: The role of trade shocks. Food Policy, 36 (2), 136–146. doi: https://doi.org/10.1016/j.foodpol.2010.10.003
7. Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment, 141, 116–128. doi: https://doi.org/10.1016/j.rse.2013.10.027
8. MacDonald, R. B., Hall, F. G. (1980). Global Crop Forecasting. Science, 208 (4445), 670–679. doi: https://doi.org/10.1126/science.208.4445.670
9. Puma, M. J., Bose, S., Chon, S. Y., Cook, B. I. (2015). Assessing the evolving fragility of the global food system. Environmental Research Letters, 10 (2), 024007. doi: https://doi.org/10.1088/1748-9326/10/2/024007
10. Stone, R. C., Meinke, H. (2005). Operational seasonal forecasting of crop performance. Philosophical Transactions of the Royal Society B: Biological Sciences, 360 (1463), 2109–2124. doi: https://doi.org/10.1098/rstb.2005.1753
11. Fischer, R. A. (2015). Definitions and determination of crop yield, yield gaps, and of rates of change. Field Crops Research, 182, 9–18. doi: https://doi.org/10.1016/j.fcr.2014.12.006
12. Nandram, B., Berg, E., Barboza, W. (2013). A hierarchical Bayesian model for forecasting state-level corn yield. Environmental and Ecological Statistics, 21 (3), 507–530. doi: https://doi.org/10.1007/s10651-013-0266-z
13. Pease, J. W., Wade, E. W., Skees, J. S., Shrestha, C. M. (1993). Comparisons between Subjective and Statistical Forecasts of Crop Yields. Review of Agricultural Economics, 15 (2), 339. doi: https://doi.org/10.2307/1349453
14. Lobell, D. B., Schlenker, W., Costa-Roberts, J. (2011). Climate Trends and Global Crop Production Since 1980. Science, 333 (6042), 616–620. doi: https://doi.org/10.1126/science.1204531
15. Arkin, G. F., Richardson, C. W., Maas, S. J. (1980). Forecasting Grain Sorghum Yields Using Simulated Weather Data and Updating Techniques. Transactions of the ASAE, 23 (3), 0676–0680. doi: https://doi.org/10.13031/2013.34645
16. Kadaja, J., Saue, T., Vii, P. (2009). Probabilistic Yield Forecast Based on Aproduction Process Model. Computer and Computing Technologies in Agriculture II, Volume 1, 487–494. doi: https://doi.org/10.1007/978-1-4419-0209-2_50
17. Reynolds, C. A., Yitayew, M., Slack, D. C., Hutchinson, C. F., Huete, A., Petersen, M. S. (2000). Estimating crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. International Journal of Remote Sensing, 21 (18), 3487–3508. doi: https://doi.org/10.1080/014311600750037516

18. 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, 151, 61–69. doi: https://doi.org/10.1016/j.compag.2018.05.012
19. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press. Available at: https://www.deeplearningbook.org/
20. Tian, H., Wang, P., Tansey, K., Zhang, S., Zhang, J., Li, H. (2020). An IPSO-BP neural network for estimating wheat yield using two remotely sensed variables in the Guanzhong Plain, PR China. Computers and Electronics in Agriculture, 169, 105180. doi: https://doi.org/10.1016/j.compag.2019.105180
21. Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P. et. al. (2018). Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. Agricultural and Forest Meteorology, 260-261, 300–320. doi: https://doi.org/10.1016/j.agrformet.2018.06.009
22. Singh, P. K., Singh, K. K., Singh, P., Balasubramanian, R., Baxla, A. K., Kumar, B. et. al. (2017). Forecasting of wheat yield in various agro-climatic regions of Bihar by using CERES-Wheat model. Journal of Agrometeorology, 19 (4), 346–349. doi: https://doi.org/10.54386/jam.v19i4.604
23. Portmann, F. T., Siebert, S., Döll, P. (2010). MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. Global Biogeochemical Cycles, 24 (1). doi: https://doi.org/10.1029/2008gb003435
24. Giri, A. K., Bhan, M., Agrawal, K. K. (2017). Districtwise wheat and rice yield predictions using meteorological variables in eastern Madhya Pradesh. Journal of Agrometeorology, 19 (4), 366–368. doi: https://doi.org/10.54386/jam.v19i4.610
25. Singh, M., Sharma, S. (2017). Forecasting the maize yield in Himachal Pradesh using climatic variables. Journal of Agrometeorology, 19 (2), 167–169. doi: https://doi.org/10.54386/jam.v19i2.715
26. Schauberger, B., Gornott, C., Wechsung, F. (2017). Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting. Global Change Biology, 23 (11), 4750–4764. doi: https://doi.org/10.1111/gcb.13738
27. Caselli, M., Trizio, L., de Gennaro, G., Ielpo, P. (2008). A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model. Water, Air, and Soil Pollution, 201 (1-4), 365–377. doi: https://doi.org/10.1007/s11270-008-9950-2
28. Niedbała, G., Kurasiak-Popowska, D., Stuper-Szablewska, K., Nawracała, J. (2020). Application of Artificial Neural Networks to Analyze the Concentration of Ferulic Acid, Deoxynivalenol, and Nivalenol in Winter Wheat Grain. Agriculture, 10 (4), 127. doi: https://doi.org/10.3390/agriculture10040127
29. Zaefizadeh, M., Jalili, A., Khayatnezhad, M., Gholamin, R., Mokhtari, T. (2011). Comparison of multiple linear regressions (MLR) and artificial neural network (ANN) in predicting the yield using its components in the hulless barley. Advances in Environmental Biology, 5, 109–113. Available at: https://www.academia.edu/77348556/Comparison_of_Multiple_Linear_Regressions_MLR_and_Artificial_Neural_Network_ANN_in_Predicting_the_Yield_Using_its_Components_in_the_Hulless_Barley
30. Agentstvo Respubliki Kazakhstan po statistike. Portret sela (2011). Astana, 92. Available at: https://stat.gov.kz/api/getFile/?docId=WC16200032726
31. Gribskiy, A. A. (2005). Pochvy i zemel'nye resursy Severo-Kazakhstanskoy oblasti. Petropavlovsk, 34.
32. Ritchie, S. W., Hanway, J. J., Thompson, H. E. (1985). How a soybean plant develops. Special Report No. 53. Ames, Iowa. Available at: http://publications.iowa.gov/14855/1/1985%20How%20a%20Soybean%20Plant%20Develops.pdf
33. Arinov, K. K., Musynov, K. M., Shestakova, N. A., Serekpaev, A. A. (2013). Rastenievodstvo. Astana, 507.
34. Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. Available at: https://www.gov.kz/memleket/entities/stat?lang=ru
35. Deep Learning Toolbox. MathWorks. Available at: https://www.mathworks.com/products/deep-learning.html
36. Drummond, S. T., Sudduth, K. A., Birrell, S. J. (1995). Analysis and correlation methods for spatial data. ASAE Paper No. 951335. St. Joseph.
37. Irmak, A., Jones, J. W., Batchelor, W. D., Irmak, S., Paz, J. O., Boote, K. J. (2006). Analysis of spatial yield variability using a combined crop model-empirical approach. Transactions of the ASABE, 49 (3), 811–818. doi: https://doi.org/10.13031/2013.20464
38. Wilkerson, J. B., Sui, R., Hart, W. E., Wilhelm, L. R., Howard, D. D. (1999). Artificial neural networks for determining nitrogen status in corn. ASAE Paper No. 99-3042. St. Joseph, Mich.: ASAE.
39. Braga, R. P. (2000). Predicting the spatial pattern of grain yield under water limiting conditions. Gainesville.
40. Liu, J., Goering, C. E., Tian, L. (2001). A neural network for setting target corn yields. Transactions of the ASAE, 44 (3). doi: https://doi.org/10.13031/2013.6097
41. Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323 (6088), 533–536. doi: https://doi.org/10.1038/323533a0
42. Dayhoff, J. E. (1990). Neural Network Architectures: An Introduction. Van Nostrand Reinhold Company, 259.
43. Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H. (2014). Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs. Information Processing in Agriculture, 1 (1), 14–22. doi: https://doi.org/10.1016/j.inpa.2014.04.001
44. Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H. (2014). Prediction of potato yield based on energy inputs using multi-layer adaptive neuro-fuzzy inference system. Measurement, 47, 521–530. doi: https://doi.org/10.1016/j.measurement.2013.09.020

45. Amid, S., Mesri Gundoshmian, T. (2016). Prediction of output energies for broiler production using linear regression, ANN (MLP, RBF), and ANFIS models. Environmental Progress & Sustainable Energy, 36 (2), 577–585. doi: https://doi.org/10.1002/ep.12448
46. Vivas, E., Allende-Cid, H., Salas, R. (2020). A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. Entropy, 22 (12), 1412. doi: https://doi.org/10.3390/e22121412
47. Wang, X., Huang, J., Feng, Q., Yin, D. (2020). Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. Remote Sensing, 12 (11), 1744. doi: https://doi.org/10.3390/rs12111744
48. Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B., Hammer, G. L. (2020). Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. Remote Sensing, 12 (6), 1024. doi: https://doi.org/10.3390/rs12061024
49. Felipe Maldaner, L., de Paula Corrêdo, L., Fernanda Canata, T., Paulo Molin, J. (2021). Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. Computers and Electronics in Agriculture, 181, 105945. doi: https://doi.org/10.1016/j.compag.2020.105945
50. Sharma, L. K., Singh, T. N. (2017). Regression-based models for the prediction of unconfined compressive strength of artificially structured soil. Engineering with Computers, 34 (1), 175–186. doi: https://doi.org/10.1007/s00366-017-0528-8
51. Peng, J., Kim, M., Kim, Y., Jo, M., Kim, B., Sung, K., Lv, S. (2017). Constructing Italian ryegrass yield prediction model based on climatic data by locations in South Korea. Grassland Science, 63 (3), 184–195. doi: https://doi.org/10.1111/grs.12163
52. Kim, S., Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32 (3), 669–679. doi: https://doi.org/10.1016/j.ijforecast.2015.12.003
53. Bhojani, S. H., Bhatt, N. (2020). Wheat crop yield prediction using new activation functions in neural network. Neural Computing and Applications, 32 (17), 13941–13951. doi: https://doi.org/10.1007/s00521-020-04797-8
54. Singh, R., Umrao, R. K., Ahmad, M., Ansari, M. K., Sharma, L. K., Singh, T. N. (2017). Prediction of geomechanical parameters using soft computing and multiple regression approach. Measurement, 99, 108–119. doi: https://doi.org/10.1016/j.measurement.2016.12.023
55. Chen, J.-F., Do, Q., Nguyen, T., Doan, T. (2018). Forecasting Monthly Electricity Demands by Wavelet Neuro-Fuzzy System Optimized by Heuristic Algorithms. Information, 9 (3), 51. doi: https://doi.org/10.3390/info9030051
56. Gandhi, N., Petkar, O., Armstrong, L. J. (2016). Rice crop yield prediction using artificial neural networks. 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR). doi: https://doi.org/10.1109/tiar.2016.7801222
57. Gandhi, N., Armstrong, L. J., Petkar, O., Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). doi: https://doi.org/10.1109/jcsse.2016.7748856
58. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 284, 107886. doi: https://doi.org/10.1016/j.agrformet.2019.107886
59. Mishra, S., Paygude, P., Chaudhary, S., Idate, S. (2018). Use of data mining in crop yield prediction. 2018 2nd International Conference on Inventive Systems and Control (ICISC). doi: https://doi.org/10.1109/icisc.2018.8398908
60. Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U. et. al. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agriculture, 20 (5), 1015–1029. doi: https://doi.org/10.1007/s11119-018-09628-4
61. Tao, F., Xiao, D., Zhang, S., Zhang, Z., Rötter, R. P. (2017). Wheat yield benefited from increases in minimum temperature in the Huang-Huai-Hai Plain of China in the past three decades. Agricultural and Forest Meteorology, 239, 1–14. doi: https://doi.org/10.1016/j.agrformet.2017.02.033