Gniewko NIEDBAŁA[1], Magdalena PIEKUTOWSKA[2], Mariusz ADAMSKI[1]
[1] Poznan University of Life Sciences, Institute of Biosystems Engineering, Faculty of Agronomy and Bioengineering
ul. Wojska Polskiego 50, 60-627 Poznań, Poland
[2] Koszalin University of Technology, Department of Agrobiotechnology, Faculty of Mechanical Engineering
ul. Racławicka 15-17, 75-620 Koszalin, Poland
e-mail: gniewko@up.poznan.pl

Received: 2018-09-19 ; Accepted: 2018-11-22

MULTIPLE REGRESSION ANALYSIS MODEL TO PREDICT AND SIMULATE WINTER RAPESEED YIELD

Summary

The aim of the work is to create a model for prediction and simulation of winter rapeseed yield. The model made it possible to perform a yield forecast on 30 June, directly before harvest in the current agrotechnical season. The prediction model was built using the multiple regression method (MLR). The model was based on meteorological data (air temperature and precipitation) and information about mineral fertilization. The data were collected from the years 2008-2017 from 291 production fields located in Poland, in the southern Opole region. The assessment of the quality of forecasts generated on the basis of the regression model was verified by determining prediction errors using RAE, RMS, MAE and MAPE error meters. An important feature of the created prediction model concerns the possibility of making the forecast in the current agrotechnical year on the basis of the current weather and fertilizer information.
Key words: *forecast, multiple regression, MLR, winter rapeseed, yield prediction*

MODEL ANALIZY REGRESJI WIELORAKIEJ DLA PROGNOZY I SYMULACJI PLONU RZEPAKU OZIMEGO

Streszczenie

Celem pracy było zbudowanie modelu do predykcji i symulacji plonu rzepaku ozimego. Model ten umożliwił wykonanie prognozy plonu na dzień 30 czerwca, bezpośrednio przed zbiorem w aktualnie trwającym sezonie agrotechnicznym. Do budowy modelu predykcyjnego użyto metody regresji wielorakiej (MLR). Model powstał w oparciu o dane meteorologiczne (temperatura powietrza i opady atmosferyczne) oraz informacje o nawożeniu mineralnym. Dane zostały zebrane z lat 2008-2017 z 291 pól produkcyjnych zlokalizowanych w Polsce, na obszarze południowej Opolszczyzny. Ocena jakości prognoz wytworzonych na bazie modelu regresyjnego została zweryfikowana poprzez określenie błędów prognozy za pomocą mierników błędów RAE, RMS, MAE oraz MAPE. Ważną cechą wytworzonego modelu predykcyjnego jest możliwość wykonania prognozy w bieżącym roku agrotechnicznym w oparciu o aktualne informacje pogodowe i nawozowe.
Słowa kluczowe*: prognoza, regresja wielokrotna, MLR, rzepak ozimy, prognoza plonu*

1. Introduction

Over the past 50 years, there has been a two-fold increase in the production of oilseeds and almost five-fold increase in the area of oil seed rape (*Brassica napus*) crops in the world. Currently, the discussed plant covers about 0.6% of the area of all crops in the world, which is 33 million ha, where the production of 67.7 million t provides 20 million t of edible oil and components needed to produce diesel oil [8, 22].

Rapeseed is grown mainly in Europe and Canada, China and India. In Poland, winter rapeseed was cultivated on 947,000 ha of sown area in 2015. The average yield per 1 ha was 28.5 dt, while in 2014 it reached the level of 34.4 dt·ha-1. Winter rape is the third most cultivated plant in Poland after winter wheat and winter triticale. The share of Polish winter oilseed rape production in comparison to the European Union in 2014 was 13.5% [5].

An increase in plant yields is related to the use of modern cultivation technologies and the use of crop yielding models to perform simulations and, consequently, to optimise the production process. For this reason, crop yield models can lead to the formation of forecasting tools, which can be an important element of precision agriculture. [17] and the main element of decision-making support systems [16].

Vegetation of plants is largely determined by meteorological conditions. Often in developed models, analyses of climate change impacts are made, which require the integration of meteorological and crop data. [13]. The process of forecasting yields during the growing season is the basis for estimating production volumes and expected yields at the end of the growing season [1]. Punctual and accurate forecasting of yields is essential for crop production, marketing, storage, transport and decision making, which supports risk management [3, 9].

The quantity and quality of the yield of plants depend on many factors which are often correlated with each other and directly or indirectly affect the yield. The most frequently considered are weather and climatic factors (air temperature, precipitation, sunshine), soil factors (pH, structure, organic matter content, soil nutrient abundance), soil cultivation technologies, plant varieties, fertilising technology and level, plant protection, harvesting technology and crop rotation [10, 15, 21].

That is why the research should be undertaken in order to produce a simple and accurate model of winter oilseed rape yield that has not been developed yet [2]. In the fol-

lowing paper the authors will attempt to develop such a model and evaluate it.

2. Materials and methods

The regression model was built on the basis of data collected in the years 2008-2017 from winter oilseed rape production fields located in Poland, in the southern part of the Opolskie Voivodeship in the Głubczyce, Branice, Kietrz, Baborów and Pawłowiczki communes (Fig. 1). All gathered data from 291 fields were used for the construction and verification of the regression model (Tab. 1). This information was the basis for the creation of the database which was divided into two sets I and II. Set I (246 fields) consisted of information from the years 2008-2016 which were used to build the model. Set II (45 fields) from 2017 consisted only of data for model validation and they did not take part in their construction. Meteorological data - average daily air temperature and daily precipitation sums referring to the area and the research period were obtained from a local meteorological station located in the researched area. The construction of a regressive prediction model was prepared on the basis of the forecasting deadline in a particular calendar year, i.e. 30 June.

The model takes into account factors collected from 1st January to 30th June of a given year, which affect the yield and are easily accessible for agricultural producers - Tab. 2.

The essence of this approach to the prediction of winter oilseed rape yield consists in the ability to make forecasts and simulations of the expected yield directly before harvest, in the current agrotechnical season.

Source: own work / Źródło: opracowanie własne

Fig. 1. Research area – southern part of the Opole Voivodeship
Rys. 1. Obszar badawczy – południowa część województwa opolskiego

Table 1. The number of productive fields of winter rapeseed divided into two sets, I and II
Tab. 1. Liczba pól produkcyjnych rzepaku ozimego podzielona na dwa zbiory I i II

	Set I									Set II
Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Number of fields	25	23	30	30	28	30	26	17	37	45

Source: own work / Źródło: opracowanie własne

Table 2. Data structure in MRL prediction model
Tab. 2. Struktura danych w predykcyjnym modelu MLR

Symbol	Unit of measure	Variable name	Scope of data
R1-4_CY	mm	The sum of precipitation from January 1 to April 15 of the current year	64.80-173
T1-4_CY	^{o}C	The average air temperature from January 1 to April 15 of the current year	-0.30-8.80
R4_CY	mm	The sum of atmospheric precipitation from April 1 to April 30 of the current year	14.20-119.20
T4_CY	^{o}C	The average air temperature from April 1 to April 30 of the current year	7.70-11.70
R5_CY	mm	The sum of precipitation from May 1 to May 31 of the current year	32-202.20
T5_CY	^{o}C	The average air temperature from May 1 to May 31 of the current year	12-14.70
R6_CY	mm	The sum of precipitation from June 1 to June 30 of the current year	37.60-147.50
T6_CY	^{o}C	Average air temperature from June 1 to June 30 of the current year	15.20-18.90
N_LY	kg · ha^{-1}	Total fertilization N - autumn in the previous year	0-42
N_CY	kg · ha^{-1}	Total fertilization N - autumn in the current year	127-280
P2O5_CY	kg · ha^{-1}	The sum of P2O5 fertilization in the current year	0-140
K2O_CY	kg · ha^{-1}	The sum of K2O fertilization in the current year	0-463
MGO_CY	kg · ha^{-1}	The sum of MgO fertilization in the current year	1-38
SO3_CY	kg · ha^{-1}	The sum of SO3 fertilization in the current year	23-140
B_CY	kg · ha^{-1}	The sum of B fertilization in the current year	0.16-2.38
CU_CY	g · ha^{-1}	The sum of Cu fertilization in the current year	0-115
MN_CY	g · ha^{-1}	The sum of Mn fertilization in the current year	0-358
MO_CY	g · ha^{-1}	The sum of Mo fertilization in the current year	0-204
ZN_CY	g · ha^{-1}	The sum of Zn fertilization in the current year	0-205

Source: own work / Źródło: opracowanie własne

2.1. Method of construction of the MLR model

Multiple regression is a statistical method whose main goal is to quantify the connections between many independent variables and a dependent variable. Even if there is no reasonable dependence between variables, one can try to link them by the use of a mathematical equation. This equation may not have a physical sense, but under some assumptions it allows to forecast values determined on the basis of knowledge of other variables [19].

Multiple regression is preceded by examination of the determination coefficient R^2 for the examined features. It is used to evaluate the degree of explanation of the total variability of a dependent variable by an independent variable. It is equal to the square of the multiple correlation coefficient between the analyzed traits. The continuation of the regression analysis is the determination of the probability factor for absolute statistics "t", verified at the level of significance α = 0.05 (statistically significant difference). In the final phase of this stage the regression equation is constructed in the form:

$$Y = a + b1X1 + b2X2 + ... + bpXp, \quad (1)$$

where:
Y – dependent variable (examined feature),
a – constant,
Xp – value of the independent variable,
bp – regression rate.

Equation (1) presents a regression model for the predicted trait - winter rapeseed yield.

2.2. Methodology of evaluation of the created model

Evaluation of the predictive ability of the produced model is being done with the use of indicators of forecast error (*ex post*), comparing data from set II to the results of prediction created on the basis of set I. These errors are characterised by the fact that they are calculated on the basis of past data, i.e. on the basis of information on predictions that have already expired and on the corresponding realisation of the forecast variable. A forecast error is the difference between the realisation of a forecast variable over time and a forecast realised for the same period [18].

The validation of the produced models was carried out on the basis of data from the year 2017, which included 45 winter rape fields. These data did not participate in the construction of the model. The methodological methods widely described in the literature were used to evaluate the quality of forecasts [4, 9, 11, 12, 14, 18].

– RAE – relative approximation error;

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}} \quad (2)$$

– RMS – root mean square error;

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (3)$$

– MAE – mean absolute error;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

– MAPE – mean absolute percentage error;

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\ \%, \quad (5)$$

where:
n - number of observations,
y_i - the real values obtained during the research,
$\hat{y}_i$ - the values determined by the model.

In order to illustrate better the relations between the real yield and the forecast yield, a graph is made, showing the mutual relations and a linear equation is determined.

3. Results and discussion

The produced regression model is based on 19 independent features contained in Tab. 2. The dependent feature refers to the yield of winter oilseed rape [t·ha^{-1}]. Tab. 3 presents the results for the produced regression model.

Determination of the statistical significance level:
- not significance,
* significance for α = 0,05

On the basis of the above results, the multiple regression equation takes the form:

Yield = - 4.87613701 – 0.01320****R1-4_CY*** + 0.08832****T1-4_CY*** + 0.01940****R4_CY*** – 0.07458****T4_CY*** – 0.00218****R5_CY*** + 0.74854****T5_CY*** + 0.02118****R6_CY*** – 0.15313****T6_CY*** – 0.00245****N_LY*** + 0.00019****N_CY*** + 0.00169****P2O5_CY*** + 0.00097****K2O_CY*** + 0.00414**MGO_CY* – 0.00493**SO3_CY* + 0.65611**B_CY* + 0.00833**CU_CY* – 0.00054****MN_CY*** + 0.00222****MO_CY*** – 0.00565****ZN_CY***

In order to determine the quality of the forecast, the calculations used for the *ex post* methods have been carried out using formulae (2 - 5), with the results shown in Tab. 4.

Table 4. Measures prediction *ex post* of analyzed MLR model
Tab. 4. Mierniki predykcyjne ex post w analizowanym modelu MLR

RAE [-]	RMS [-]	MAE [t·ha^{-1}]	MAPE [%]
0.4232	1.6704	1.5950	44.21

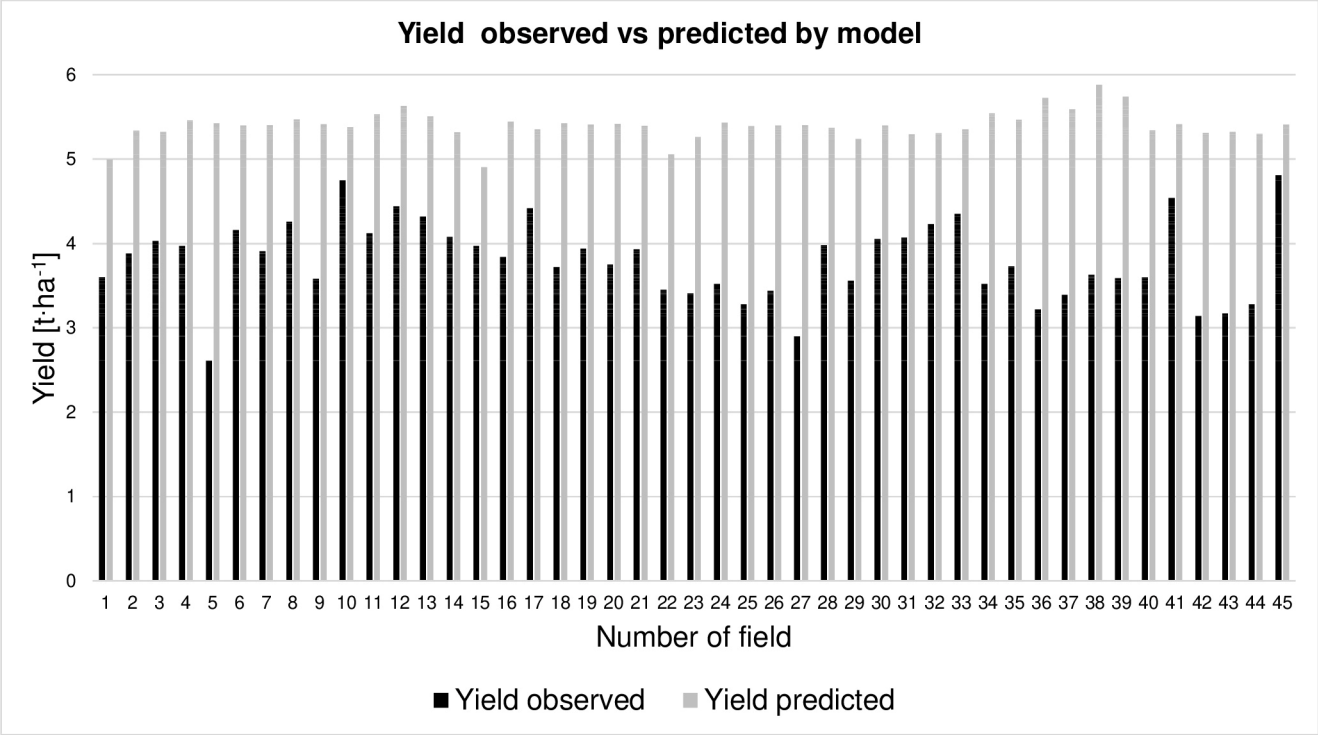
Source: own work / Źródło: opracowanie własne

In the next step, a graph of relations between the real yield and the MLR model forecast was created (Fig. 3) and a linear equation was determined based on the results obtained (Fig. 4).

Table 3. Regression coefficients, standard errors and probability levels for the MLR model
Tab. 3. Współczynniki regresji, błędy standardowe oraz poziomy prawdopodobieństwa dla modelu MLR

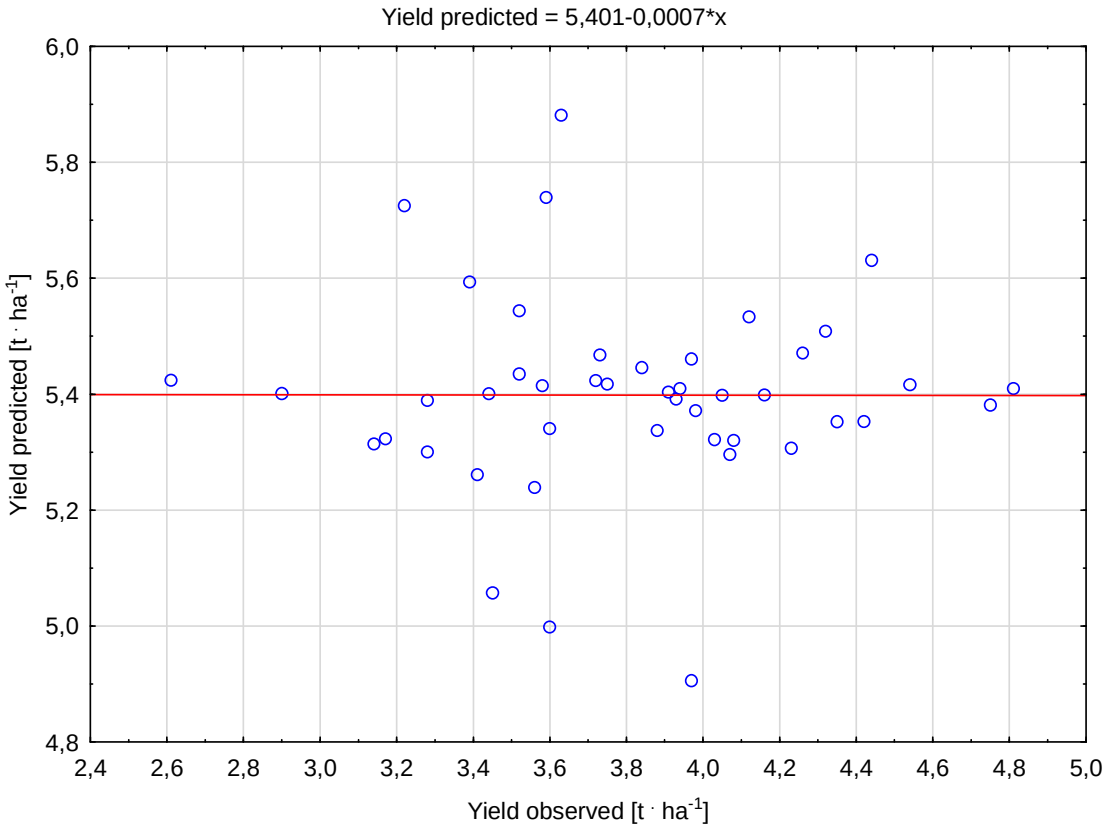
Variable	**Yield: R= 0.81744181, R^2= 0.66821111, Constant= -4.87613701**			
	b	standard error b	p	significance
R1-4_CY	**-0.01320**	**0.003032**	**0.000020**	*
T1-4_CY	**0.08832**	**0.023122**	**0.000173**	*
R4_CY	0.01940	0.019531	0.321614	-
T4_CY	-0.07458	0.091784	0.417293	-
R5_CY	-0.00218	0.004968	0.661118	-
T5_CY	**0.74854**	**0.191453**	**0.000122**	*
R6_CY	**0.02118**	**0.00671**	**0.001813**	*
T6_CY	**-0.15313**	**0.074569**	**0.041171**	*
N_LY	-0.00245	0.006062	0.686849	-
N_CY	0.00019	0.001433	0.894877	-
P2O5_CY	0.00169	0.002142	0.431717	-
K2O_CY	0.00097	0.000625	0.120390	-
MGO_CY	0.00414	0.009121	0.650208	-
SO3_CY	-0.00493	0.003315	0.138037	-
B_CY	**0.65611**	**0.324602**	**0.044429**	*
CU_CY	0.00833	0.006237	0.182973	-
MN_CY	-0.00054	0.00167	0.745490	-
MO_CY	0.00222	0.005732	0.699083	-
ZN_CY	-0.00565	0.003849	0.143504	-

Source: own work / Źródło: opracowanie własne



Source: own work / Źródło: opracowanie własne

Fig. 3. Observed and predicted yield of winter rapeseed in MLR model
Rys. 3. Rzeczywisty i prognozowany przez model MLR plon rzepaku ozimego



Source: own work / Źródło: opracowanie własne

Fig. 4. Relation between observed and predicted yield with linear equation
Rys. 4. Relacja pomiędzy plonem rzeczywistym i prognozowanym wraz z równaniem liniowym

The determination coefficient for the produced MLR model took the following value R^2= 0.66821111, and the constant in the regression equation was -4.87613701. These results show that the model is on a medium adjustment to the empirical data on which the MLR model was created. Coefficient "b" obtained the highest value for two independent features. For the T5_CY feature it was 0.748542514, while for the B_CY feature it was 0.656113104. This means that both T5_CY and B_CY had the greatest impact on the shaping of the volume of the forecasted winter oilseed rape yield.

The produced MLR model is based on empirical data, which are usually easily accessible to every grower and these are weather data and fertilisation information. The advantage of the produced model is the possibility to use it in the current agrotechnical year, before harvesting, for example on the 1st of July or at a later date. Often the forecasting models are based solely on experimental data [3, 7, 20]. Such an approach makes it difficult to use models and create forecasts by a wide range of interested persons or institutions. It was assumed that the correct functioning of the model would be verified by comparing the obtained forecasts with the actual rapeseed yields in the last year of the study.

In view of the above, four *ex post* error measures were used in this paper: relative approximation error (RAE), root mean square error (RMS), mean absolute error (MAE), mean absolute percentage error (MAPE). They were applied to determine the quality of the model and to determine the errors in the forecast of winter oilseed rape yield.

Table 4 shows the error values for the model produced. To the most commonly used indicators characterizing the values of prediction errors belongs MAPE which is easy to interpret [6, 9, 15]. The MAPE error value for the MLR model was 44.21%. Considering a critical MAPE error rate of up to 10%, in cases that are significantly affected by random conditions [18], the results are unsatisfactory.

For this reason, further work should be undertaken in order to reduce the forecast error by selecting another set of independent features or changing the method of building the forecasting model.

4. References

[1] Bussay A., van der Velde M., Fumagalli D., Seguini L.: Improving operational maize yield forecasting in Hungary. Agric. Syst., 2015, 141: 94–106.

[2] Diepenbrock W.: Yield analysis of winter oilseed rape (Brassica napus L.): A review. F. Crop. Res., 2000, 67: 35–49.

[3] Domínguez J.A., Kumhálová J., Novák P.: Winter oilseed rape and winter wheat growth prediction using remote sensing methods. Plant, Soil Environ., 2015, 61: 410–416.

[4] Emamgholizadeh S., Parsaeian M., Baradaran M.: Seed yield prediction of sesame using artificial neural network. Eur. J. Agron., 2015, 68: 89–96.

[5] FAO: Food and Agriculture Organization of the United Nations (FAO). FAOSTAT Online Statistical Service. http://faostat.fao.org. 2017.

[6] Farjam A., Omid M., Akram A., Fazel Niari Z.: A neural network based modeling and sensitivity analysis of energy inputs for predicting seed and grain corn yields. J. Agric. Sci. Technol., 2014, 16: 767–778.

[7] Guérif M., Duke C.: Calibration of the SUCROS emergence and early growth module for sugar beet using optical remote sensing data assimilation. Eur. J. Agron., 1998, 9: 127–136.

[8] Jiang C., Shi J., Li R., Long Y., Wang H., Li D., Zhao J., Meng J.: Quantitative trait loci that control the oil content variation of rapeseed (Brassica napus L.). Theor. Appl. Genet., 2014, 127: 957–968.
[9] Kantanantha N., Serban N., Griffin P.: Yield and price forecasting for stochastic crop decision planning. J. Agric. Biol. Environ. Stat., 2010, 15: 362–380.
[10] Khairunniza-Bejo S., Mustaffha S., Ishak W., Ismail W.: Application of Artificial Neural Network in Predicting Crop Yield: A Review. J. Food Sci. Eng., 2014, 4: 1–9.
[11] Li F., Qiao J., Han H., Yang C.: A self-organizing cascade neural network with random weights for nonlinear system modeling. Appl. Soft Comput., 2016, 42: 184–193.
[12] Logan T.M., McLeod S., Guikema S.: Predictive models in horticulture: A case study with Royal Gala apples. Sci. Hortic. (Amsterdam)., 2016, 209: 201–213.
[13] Nelson G.C., Valin H., Sands R.D., Havlík P., Ahammad H., Deryng D., Elliott J., Fujimori S., Hasegawa T., Heyhoe E., Kyle P., Von Lampe M., Lotze-Campen H., Mason d'Croz D., van Meijl H., i wsp.: Climate change effects on agriculture: economic responses to biophysical shocks. Proc. Natl. Acad. Sci. U. S. A., 2014, 111: 3274–9.
[14] Niazian M., Sadat-Noori S.A., Abdipour M.: Artificial neural network and multiple regression analysis models to predict essential oil content of ajowan (Carum copticum L.). J. Appl. Res. Med. Aromat. Plants, 2018, 9: 124–131.
[15] Niedbała G., Przybył J., Sęk T.: Prognosis of the content of sugar in the roots of sugar-beet with utilization of the regression and neural techniques. Agric. Engienering, 2007, 2: 225–234.
[16] Park S.J., Hwang C.S., Vlek P.L.G.: Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agric. Syst., 2005, 85: 59–81.
[17] Shearer J.R., Burks T.F., Fulton J.P., Higgins S.F.: Yield Prediction Using A Neural Network Classifier Trained Using Soil Landscape Features and Soil Fertility Data . Annu. Int. Meet. Midwest Express Cent., 2000, 5–9.
[18] Stańko S.: Prognozowanie w agrobiznesie. Teoria i przykłady zastosowania. Wydawnictwo SGGW, Warszawa 2013.
[19] Trzepieciński T.: Zastosowanie regresji wielokrotnej i sieci neuronowej do modelowania zjawiska tarcia. Zesz. Nauk. WSInf, 2010, 9: 31–43.
[20] Vandendriessche H.J.: A model of growth and sugar accumulation of sugar beet for potential production conditions: SUBEMOpo I. Theory and model structure. Agric. Syst., 2000, 64: 21–35.
[21] Velička R., Marcinkevičienė A., Pupalienė R., Butkevičienė L.M., Kosteckas R., Čekanauskas S., Kriaučiūnienė Z.: Winter oilseed rape and weed competition in organic farming using non-chemical weed control. Zemdirbyste-Agriculture, 2016, 103: 11–20.
[22] Wenda-Piesik A., Hoppe S.: Evaluation of hybrid and population cultivars on standard and high-input technology in winter oilseed rape. Acta Agric. Scand. Sect. B — Soil Plant Sci., 2018, 1–12.