Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 9, 2022, page 211-221

Crop Yield Prediction with Multiple Linear Regression in Malaysia

Muhammad Fikri Badaruzaman, Ong Chee Tiong
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

Abstract
Selected Agricultural Indicators, Malaysia, 2021 is a compilation of data on the agricultural sector's economic performance, employment, domestic production, and exports. The National Agro-Food Policy (DAN) 2011-2020-2021-2030 enacted by the government is a transformation of the country's productivity and food quality. However, there are certain flaws in achieving food self-sufficiency of at least 74.1 percent while only contributing 2.3 percent to the country's gross domestic product (GDP). The methodology of this study uses data collected from Department of Statistics Malaysia from year 1980 until year 2020. The multiple linear regression was used in this study. The modelling of multiple linear regression is purposed to estimate the average yield according to the planted area and rice production. Based on the obtained results, the two factors were found to affect the average yield. It needs to be emphasized to increase national rice supply quantity and quality in a comprehensive and adequate manner, as well as national income and clean food products. The study concludes that the planted area and rice production are important parameters to consider when developing a crop yield regression model. The MLR model for this research found to be $\hat{y} = 3044.960 + (-0.005)\, x_1 + 0.002\, x_2$. The study's findings imply that rice products in Malaysia should be raised in quantity and given priority over other commodities. The Ministry of Agriculture and Food Industry should pay close attention to the agricultural sector and is expected to play a role in it.

Keywords: Multiple Linear Regression; Rice; Average yield; Agriculture; Food quality

1. Introduction

Selected Agricultural Indicators, Malaysia, 2021 contains statistics on the agriculture sector's economic performance, employment, domestic production, and external trade. According to Chief Statistician Malaysia [1]. Malaysia's Gross Domestic Product (GDP) fell by 5.6 percent in 2020, owing to a drop in all sectors of the economy, compared to a growth of 4.4 percent in 2019. In 2020, the agricultural sector will contribute 7.4\% of Malaysia's GDP. In other hand, Agroeconomists need easy and precise estimation strategies to anticipate yields in the agricultural planning process [2][3]. Crop production prediction algorithms are based on input features such as area, irrigation methods, temperature, and so on.

Food deficiency was a global concern before COVID-19. One in nine people, or 820 million, are food insecure. Due to job losses and other income losses, more people around the world may go hungry. In 2019, when the Covid-19 epidemic began, Malaysia's unemployment rate was 3.31 percent. This group includes urban poor, rural dwellers, small farmers, and jobless people.

Consuming food requires a healthy environment, hygienic facilities, knowledge, and awareness. SW-Corp Malaysia reports that Malaysians produce 38,000 tonnes of rubbish per day, with 15,000 tonnes being food waste. Safe food care, preparation, and storage. Education may encourage safe diet and correct nutrition, as well as revenue, processing, storage, and handling procedures at all levels of the food supply chain, especially at the content stage home.

Pest and disease control are other food crop challenges. Farmers face blast disease (rice), leaf rust (corn), and fusarium (tomato). Most agricultural inputs, like fertilisers and chemicals, are imported, increasing production costs. Malaysia's agricultural output has fallen behind Thailand, Indonesia, and Vietnam. These challenges may affect Malaysia's food production and ability to meet demand.

2. Literature Review

2.1. *Artificial Neural Network (ANN)*

The ANN is the most widely used ML approach for agricultural yield prediction [4][5][6][7]. From the historical dataset, it learns patterns. It can describe the complicated nonlinear relationship between input and output [8]. The four input independent variables of the agricultural training data set are employed in much research, which are the same as input neurons in the input layer, such as area in hectare, amount of water, paddy production, and rice production [9]. Only production is dependent on another variable. Input and output layers are hidden. Each layer is interconnected. Signal travels from input to output through a hidden layer.

The input of the i^{th} node hidden layer ($HSum_h$) is obtained by (2.1)

$$HSum_h = \sum x_i^{in} w_i^{in} + b_i^{in} \qquad (2.1)$$

Where x_i^{in} in is the input value and $1 \leq i \leq n$, w_i^{in} in is the input weight and b_i^{in} in is the input layer bias. The output of the hidden layer node i is obtained by (2.2)

$$HSum_{hout} = f(Hsum_h) \qquad (2.2)$$

To approximate any nonlinear function with accuracy, only one hidden layer is required [10]. The trial-and-error method is used to select the number of nodes in the hidden layer [11]. The input layer does no calculations. The hidden layer accepts inputs, weights, and bias. Hidden layer output is sigmoid activation. Hidden node output is multiplied by layer weights and added to bias. The output layer's input is the obtained value. The output layer's result is subjected to the sigmoid activation function, which yields the predicted value [12]. Only one hidden layer and one hidden neuron are employed in this study (2.3)

$$0Sum_0 = \sum x_i^h w_i^h + b_i^h \qquad (2.3)$$

2.2. *Previous Study on Crops Yield Prediction*

2.2.1. Artificial Neural Networks for Rice Yield Prediction

Rice is the most important food grain in the diets of millions of people living in the tropics and subtropics. Rice production is influenced by a variety of varietals as well as environmental factors such as soil, weather and cultivation management. When plants are grown with plenty of nutrients and water, rice grain output is largely determined by local weather variables such as sunshine hours, solar radiation, and temperature. Artificial neural networks (ANN) are a powerful empirical modelling tool that is also quite simple to use. Because rice yield is non-linear and autoregressive in nature, it is thought that ANN models offer a more versatile empirical modelling approach than linear regression approaches. ANN models are useful in domains where there is little or no understanding of the problem to be solved but training data is available.

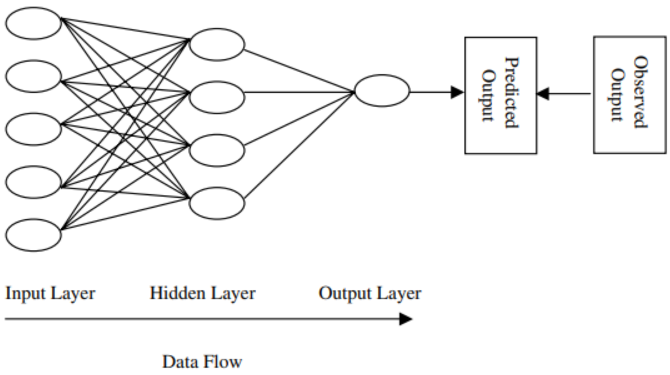

Figure 2.1 Model Feed Forward (MFF) networks

2.2.2. Tillage Systems and Soil Properties

Tillage is used to prepare the soil for planting. Physical, chemical, and biological changes must be monitored to measure tillage's impact on soil environment [15]. They're also used to improve performance and manage hazards. The effects of tillage systems on soil properties and crop production, as well as the predictive capabilities of artificial neural networks are investigated in this study. By turning the ploughed strip and displacing clods in the soil surface, deep tillage mechanical action improves the development of pores, which improves soil structure. Although, according to Estrade [16], the stability of soil structure in the surface is usually greater when using shallow tillage, direct sowing, and ridges than when ploughing. Tillage reduces surface organic matter and soil surface biological activity after ploughing. Limited tillage requires fewer mechanical passes while the soil is damp in autumn and spring and prevents plough pans.

2.2.3. ARMA, ARIMA, and ARMAX Models

2.2.3.1. Autoregressive Moving Average model (ARMA)

The ARMA model analyses and forecasts future data X_t values. AR involves regressing a variable on its lagged (past) values. The MA part models the error term as a linear combination of simultaneous and past errors. p signifies the AR part's order and q denotes the MA part's order in the ARMA (p, q) model.

2.2.3.2. Autoregressive Integrated Moving Average (ARIMA)

An autoregressive integrated moving average (ARIMA) is a generalization of the ARMA model. It is used in statistics and econometrics, particularly in time series analysis. The MA section of the equation denotes that the regression error is a linear combination of error factors.

2.2.3.3. Autoregressive Moving Average Model with Exogenous (ARMAX)

The dependent variable in the ARMA and SARIMA models is regressed exclusively on its own prior values, which works well for temperature prediction, but rainfall may be affected by other meteorological phenomena such as cloud cover, temperature, and vapor pressure. To turn our ARMA model into an ARMAX model, we add exogenous variables.

3. Methodology

3.1. *Description of the Data and Method*

The yearly of Malaysia Agriculture Statistics and Malaysia Economic Statistics are used in this study was obtained from Department of Statistics Malaysia (DOSM) which will be covering the period of year 1980 until the year 2020. The data consists of the average crop yield, planted and rice production. The data will be analyzed using Microsoft Excel and SPSS software. The average crop yield is the dependent variable for this research, while the others variable is independent variables. The average crop yield will be predicted by using Multiple Linear Regression model.

3.2. *Multiple Linear Regression*

Regression analysis is a type of predictive model that estimates the relationship between two or more variables. Remember that no assumptions are made about the causal relationship between two variables in a correlation analysis. The purpose of regression analysis is to examine the relationship between a dependent (target) variable and independent variables (predictors). The dependent variable is assumed to be the effect of the independent variables in this case. The value of predictors is used to estimate or predict the target variable's likely value [22].

Step to perform linear regression is assumed that the target and the predictors have a mathematical relationship. The relationship can be either a straight line (linear regression), a polynomial curve (polynomial regression), or a non-linear relationship (non-linear regression). Then, make a scatter plot of the target and predictor variables (simplest and most popular way) and determine the most likely values for the mathematical formula's coefficients [22]. The relationship between two or more variables is estimated using regression analysis. More specifically, regression analysis explains how the typical value of the dependent variable changes when any independent variable is changed while the other independent variables remain constant. One of the most widely used predictive modelling techniques is linear regression.
The formula for a multiple linear regression is (2.4):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \qquad (2.4)$$

Where:

y is the predicted value of the dependent variable

β_0 is the y-intercept (value of y when all other parameters are set to 0)

$\beta_1 x_1$ is the regression coefficient (β_1) of the first independent variables (x_1)

$\beta_n x_n$is the regression coefficient of the last independent variable

ϵ is model error (how much variation there is in our estimate of y)

3.3. Matrix Notation

Matrix notation is used to simplify the presentation of calculations that are performed in the linear regression. An r $\times$ c matrix is a rectangular array of elements with r rows and c columns. An r $\times$ c matrix is said to be of order r $\times$ c. The least squares approach is used to estimate the regression coefficients in multiple linear regression analysis. The regression coefficients represent the unrelated contributions of the independent variables to predicting the dependent variable. The computations to find the least squares estimators, ˆ turned out to be difficult. As a result, to make calculations easier, multiple linear regression can be represented in matrix notation. The model is in the form y = xβ + ϵ and when written in matrix notation we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The vector of least squares estimators is determined as the initial step in multiple linear regression analysis. β, this results in the linear combination y that minimises the error vector's length. Where y is in the form

$$\sum \boldsymbol{y_i} = \boldsymbol{n\widehat{\beta}_0} + \sum \boldsymbol{x_{i1}\widehat{\beta}_1} + \cdots + \sum \boldsymbol{x_{ik}\widehat{\beta}_k} \qquad (\mathbf{2.5})$$

$$\sum \boldsymbol{x_{ik}y_i} = \sum \boldsymbol{x_{i1}\widehat{\beta}_0} + \sum \boldsymbol{x_{i1}\widehat{\beta}_1} + \cdots + \sum \boldsymbol{x_{ik}^2\widehat{\beta}_k} \qquad (\mathbf{2.6})$$

The least squares estimators in the form

$$\boldsymbol{x^T x\widehat{\beta}} = \boldsymbol{x^T y} \qquad (\mathbf{2.7})$$

$$\boldsymbol{\widehat{\beta}} = (\boldsymbol{x^T x})^{-\mathbf{1}}\boldsymbol{x^T y} \qquad (\mathbf{2.8})$$

3.4. *Correlation Matrix*

A correlation matrix a table that shows the correlation coefficients for various variables. The correlation between all possible pairings of variables is represented by the matrix. It's a useful tool for quickly summarizing a large dataset and identifying and visualizing trends in the data. The variables are represented by rows and columns in a correlation matrix. The correlation coefficient is contained in each cell of a table. The correlation matrix is also commonly used in conjunction with other types of statistical analysis. It could be useful in the analysis of numerous linear regression models, for example. The models have a few independent variables. The correlation matrix determines the correlation coefficients between the independent variables in multivariate linear regression in a model.

3.5. *Assumption of Multiple Linear Regression*

To interpret a model and its limitations, it is important to understand the underlying assumptions of the method and how these affect the treatment of the data and modelling choices made. When we use linear regression to create a model, we make the following assumptions:

- Linearity between independent variables and dependent variables
- Absence of multicollinearity problem
- Constant variance

3.5.1. Linearity between independent variables and dependent variables

The main assumption of the MLR is to model the linear relationship between dependent variable and independent variables. To check the linearity, visually we used simple scatter plot between y and each x. If the scatterplot shows a non-linear relationship, the analyst must perform anon-linear regression or modify the data using SPSS.

3.5.2. Absence of Multicolinearity Problem

Multicollinearity, which arises when the independent variables (explanatory variables) are highly interrelated, should not be present in the data. When independent variables exhibit multicollinearity, determining the precise variable that contributes to the variance in the dependent variable becomes difficult. To check whether multicollinearity occurs, the correlation matrix of all independent variables is plotted as discussed in this research. We can see the pairwise correlation between all the variables after plotting the correlation matrix and colour scaling the background. The second method to check multicollinearity is to use the Variance Inflation Factor (VIF) for each independent variable. It is a measure of multicollinearity in the set of multiple regression variables. The higher the value of VIF the higher correlation between this variable and the rest. If the VIF value is higher than ten, it is usually considered to have a high correlation with other independent variables. However, the acceptance range is subject to requirements and constraints.

3.5.3. Constant Variance

The level of error in the residuals is assumed to be similar at each point of the linear model in multiple linear regression. Homoscedasticity is the term for this situation. The data can be plotted on a scatterplot or statistical software can be used to create a scatterplot.

3.5.4. Analysis of Variance

Analysis of Variance (ANOVA) provides the analysis of the variance in the model. The null hypothesis states that a population parameter (such as the mean, the standard deviation, and so on) is equal to a hypothesized value. Alternative hypothesis state that a population parameter is greater, or different than the hypothesized value in the null hypothesis.

4. Results and discussion

4.1. Scatter Plot

Each independent variables are investigated whether there is a linear relationship with the dependent variable. Scatter plot for each independent variable is presented in figure below. Based on the plot's, planted area is observed to have negative linear relationship as most of the point are not lie on a straight line. As the planted area decrease, the average crop yield will increase. The other plot indicates positive relationship exist between rice production with the dependent variable, the average crop yield. This implies, decreasing planted area will increase the average crop yield while if the rice production increase, then the average crop yield increase. The strength of the relationship is various for different independent variables and will be investigated using correlation matrix in the next subsection. All the data shown were analyzed by using Statistical Package for the Social Sciences (SPSS).

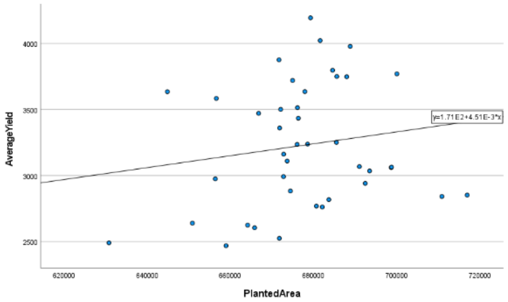

Figure 4.1 Simple Scatter Average Yield by Planted Area

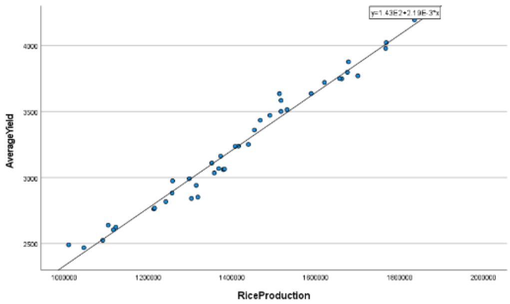

Figure 4.2 Simple Scatter Average Yield by Rice Production

4.2. Correlation Matrix

Based on the Table 4.1 above all x have significant relationship with y with low Pearson correlation coefficient ($\leq \pm 0.6$), hence we can conclude that we have overcome multicollinearity problem.

By default, SPSS always creates a full correlation matrix. Each correlation appears twice: above

Correlations

		AverageYield	PlantedArea	RiceProduction
AverageYield	Pearson Correlation	1	.160	.988**
	Sig. (2-tailed)		.316	.000
	N	41	41	41
PlantedArea	Pearson Correlation	.160	1	.309*
	Sig. (2-tailed)	.316		.049
	N	41	41	41
RiceProduction	Pearson Correlation	.988**	.309*	1
	Sig. (2-tailed)	.000	.049	
	N	41	41	41

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4.1 Correlation Matrix (2 Variables)

and below the main diagonal. The correlations on the main diagonal are the correlations between each variable and itself which is why they are all 1. Based on Table 4.1 above, the result shows that planted area and rice production have the weak correlation, ρ = 0.309. It's based on N = 41 years and its 2-tailed significance, p = 0.049. The more planted area used, the more rice will produce, but the effect is very small. At 5\% significance level the relationship exist between these 2 variables is not significant.

4.3. Variance Inflation Factor (VIF)

Model	Collinearity Tolerance	Statistics VIF
Planted Area	.904	1.106
Rice Production	.904	1.106

Table 4.2 Variance Inflation Factor (VIF)

The VIF values for each of the predictor variables:

- Planted Area: 1.106
- Rice Production: 1.106

Table 4.2 above shows the collinearity diagnostic value to check for multicollinearity. The tolerance values are all more than 0.1, suggesting that the multicollinearity assumption was not broken. This is confirmed further by VIF values less than 10, with the VIF value for planted area being 1.106, for rice production being 1.106. Overall, based on these findings, all regression assumptions were met, and so the regression analysis was carried out.

4.4. Constant Variance

Figure 4.3 shows the scatter plot of residual. A random pattern on the residual scatter plot implies that the residual has constant variance, is independent of the other factors, and is linearly connected to them. The residuals are randomly distributed around zero in no pattern, with a nearly constant variance at each level of the fitted values.

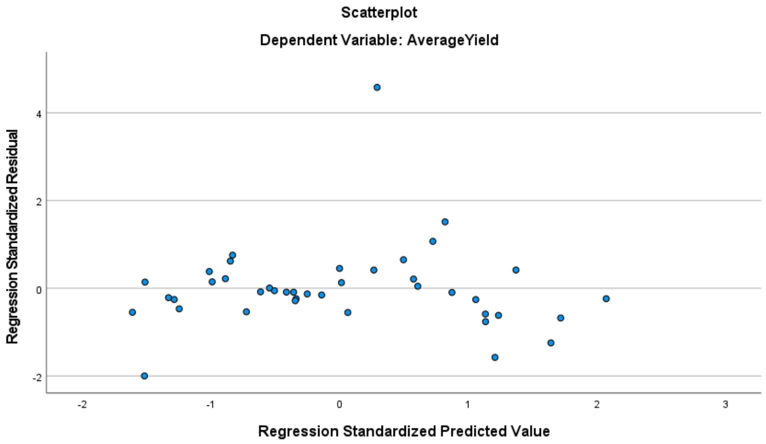

Figure 4.3 Residual Scatter Plot

4.5. Multiple Linear Regression

Table 4.3 provide R and R^2The R value represents the simple correlation and is 1.000 (the "R" Column), which indicates a high degree of correlation. The R^2 value (the "R Square" column) indicates how much of the total variation in the dependent variable, average yield of crop, can be explained by the independent variables which are planted area and rice production. In this case, 99.9% can be explained, which is very large.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000	0.999	0.999	14.952

Table 4.3 Model Summary

Table 4.4 indicates the statistical significance of the regression model that was run. Here, p-value 0.000, which is less than 0.05, and this indicates that, overall, the regression model statistically significant to predict the outcome variable. In other words, it is a good fit for the data to predict dependent variable.

Model	Sum of Square	df	Mean Square	F	Sig.
Regression	8728935.631	2	4364467.816	19521.198	.000
Residual	8495.881	38	223.576		
Total	8737431.512	40			

Table 4.4 Analysis of Variance (ANOVA)

The Coefficients Table (Table 4.5) provides us with the necessary information to predict average crop yield from planted area and rice production, as well as determine whether planted area and rice production contributes statistically significant to the model. The significance values for both independent variables are less than 0.05. Thus, both of it are significant variables to include in the regression model.

Model	Coefficient	Std. Error	Beta	t	P-value
Constant	3044.960	97.563		31.210	.000
Planted Area	-.005	.000	-.160	-30.147	.000
Rice Production	.002	.000	1.037	195.029	.000

Table 4.5 Coefficient Table

From table above, the parameter estimators are:

- $\hat{\beta}_0 = 3044.960$
- $\hat{\beta}_1 = -0.005$
- $\hat{\beta}_2 = 0.002$

Thus, the final multiple linear regression can be written as the following:

$$\boldsymbol{AverageYield = 3044.960 + (-0.005)PlantedArea + 0.002RiceProduction}$$

5. Conclusion

This research aimed to predict and explore the trends of crop yield production in Malaysia. In few studies have discussed potential factors affect average crop yield. Due to limitation of the data, only 2 independent variables that have been used for this research.

Except for rice, the subsistence level of all agro-food commodities is expected to rise. Paddy production will be intensified in the existing granary area, with enough irrigation and drainage infrastructure provided, notably in the granaries, which has the potential to be developed and land used efficiently. Consumption competition for land for construction is projected to put strain on food production in the area. From 784,069 hectares in 2019, to 841,000 hectares in 2020, the area dedicated to food production is predicted to shrink. Food production, particularly rice, would be raised from 11.2 million to 12.8 million metric tonnes to satisfy demand, which is estimated to reach 14.8 million metric tonnes by 2020.

Multicollinearity absence on this research. MLR can be used as one of the prediction models. From final model, the iteration of planted area and rice production for future years can be used to predict the average crop yield. The study can conclude that planted area and rice production are essential factors to be considered when designing regression model for crop yield production.

It is recommended to investigate what other researchers have done related to prediction in average crop yield. If any of the coefficient signs go against theory, we need to find out why and fix our model or explain the problem. One main point is that statistical data alone should not be used to dismiss the theoretical implications. Next, look at the residual plots because they can help us avoid faulty models and change our model for better results. The simplest model that generates random residuals is a potential alternative for creating a relatively precise and unbiased model. There is no single statistic that can be used to identify which model is the most effective.

Acknowledgement

The author is solely responsible for the content, which does not necessarily reflect the official views of the university. The researcher would like to thank everyone who has helped with the research.

References

[1] Selected Agricultural Indicators, Malaysia, 2021. URL https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=72&bul_id=TDV1YU4yc1Z0dUVyZ0xPV0ptRlhWQT09&menu_id=Z0VTZGU1UHBUT1VJMFlpaXRRR0xpdz09

[2] Ji, B., Sun, Y., Yang, S. and Wan, J. Artificial neural networks for rice yield prediction in mountainous regions. The Journal of Agricultural Science, 2007. 145(3): 249–261.

[3] Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S. et al. Brief history of agricultural systems modeling. Agricultural systems, 2017. 155: 240–254.

[4] Panda, S. S., Ames, D. P. and Panigrahi, S. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing, 2010. 2(3): 673–696.

[5] Gu, J., Yin, G., Huang, P., Guo, J. and Chen, L. An improved back propagation neural network prediction model for subsurface drip irrigation system. Computers & Electrical Engineering, 2017. 60: 58–65.

[6] Safa, M. and Samarasinghe, S. Modelling fuel consumption in wheat production using artificial neural networks. Energy, 2013. 49: 337–343.

[7] Khoshroo, A., Emrouznejad, A., Ghaffarizadeh, A., Kasraei, M. and Omid, M. Sensitivity analysis of energy inputs in crop production using artificial neural networks. Journal of cleaner production, 2018. 197: 992–998.

[8] Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J. and Kitchen, N. R. Statistical and neural methods for site–specific yield prediction. Transactions of the ASAE, 2003. 46(1): 5

[9] Zhang, B., Valentine, I. and Kemp, P. Modelling the productivity of naturalised pasture in the North Island, New Zealand: a decision tree approach. Ecological Modelling, 2005. 186(3): 299–311.

[10] Zhang, G., Patuwo, B. E. and Hu, M. Y. Forecasting with artificial neural networks: The state of the art. international journal of forecasting, 1998. 14(1): 35–62

[11] Shamseldin, A. Y. Application of a neural network technique to rainfall-runoff modelling. Journal of hydrology, 1997. 199(3-4): 272–294.

[12] Hill, D. J. and Minsker, B. S. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environmental Modelling & Software, 2010. 25(9): 1014–1022.

[13] Bishop, C. M. et al. Neural networks for pattern recognition. Oxford university press. 1995.

[14] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. Classification and regression trees. Routledge. 2017.

[15] Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R. and Chehaibi, S. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil and Tillage Research, 2019. 190: 202–208.

[16] Roger-Estrade, J., Richard, G., Boizard, H., Boiffin, J., Caneill, J. and Manichon, H. Modelling structural changes in tilled topsoil over time as a function of cropping systems. European Journal of Soil Science, 2000. 51(3): 455–474.

[17] Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. Time series analysis: forecasting and control. John Wiley & Sons. 2015.

[18] Khaki, S., Wang, L. and Archontoulis, S. V. A cnn-rnn framework for crop yield prediction. Frontiers in Plant Science, 2020. 10: 1750.

[19] Horie, T., Yajima, M. and Nakagawa, H. Yield forecasting. Agricultural systems, 1992. 40(1-3): 211–236.

[20] Khaki, S. and Wang, L. Crop yield prediction using deep neural networks. Frontiers in plant science, 2019. 10: 621.

[21] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R. et al. Random forests for global and regional crop yield predictions. PLoS One, 2016. 11(6): e0156571

[22] Panchotia, R. Predictive Modelling Using Linear Regression. 2020.