Environment and Ecology 40 (3B) : 1444—1449, July—September 2022
ISSN 0970-0420

Weather Based Potato Yield Modelling using Statistical and Machine Learning Technique

Akhilesh Kumar Gupta, Kader Ali Sarkar, Digvijay Singh Dhakre, Debasis Bhattacharya

Received 24 April 2022, Accepted 26 May 2022, Published on 11 August 2022

ABSTRACT

The crop yield is greatly affected by environmental factors and this study focuses on modelling the effect of weather parameters on potato yield. The crop yield forecasting is essential for better planning and policy. The weather variables, viz., maximum temperature, minimum temperature, rainfall, and relative humidity were used and their weekly effects on yield were incorporated by the means of weather indices. These weather indices and time variable *t* were used as independent variable and potato yield as dependent variable to develop a multiple regression model and different neural network models. The criterion for model selection was lowest RMSE, MAE and MAPE. The study revealed that the neural network with 4 hidden nodes was best fitted model. The best model can be used to obtain a reliable forecast of potato yield at 6-8 weeks before harvest for various policy decisions.

Keywords Potato, Weather indices, Regression, Neural networks.

INTRODUCTION

Potato (*Solanum tuberosum* L.) is one of the major important crops around the globe as well as in India. The Indo-Gangetic plains regions in the states of UP and WB contributes the major proportion of potato production in India. West Bengal is second highest producer of potato accounting for around 23% of overall production in India (Monthly report potato, May 2020). The crop yield modelling and forecasting is an essential process for better planning and policy decisions relating to its storage, marketing, pricing, export-import, distribution. The availability of these forecasts before harvests is very important for these decisions.

The potato crop is grown under irrigated conditions during the *rabi* season due to climatic requirement. The crop growth and yield are affected by many biological and environmental factors. Potato requires specific climatic conditions in relation to temperature for proper establishment and growth of the crop. The vegetative growth of the plant is favored at a relatively high temperature while tuber development is favored at low temperature (<20°C). In addition, the rainfall

Akhilesh Kumar Gupta*[1], Kader Ali Sarkar[2], Digvijay Singh Dhakre[2], Debasis Bhattacharya[3]

[1]Assistant Professor, Department of Agricultural Statistics, College of Agriculture, OUAT, Bhubaneswar 751003, Odisha, India
[2]Assistant Professor, Department of Agricultural Statistics, Institute of Agriculture, Visva-Bharati, Sriniketan (WB), India
[3]Professor & Head, Department of Agricultural Statistics, Institute of Agriculture, Visva-Bharati, Sriniketan (WB), India

Email: akhileshgupta.ouat@gmail.com
*Corresponding author

and relative humidity throughout the cropping season affect the disease and pest incidence in the potato crop. The present study focuses on the development of weather based pre-harvest yield forecast models for potato by accounting the effect of weather variables over the cropping season.

The research area of crop yield modelling is explored by many researchers over the time. The pre-harvest model to forecast rice and wheat was developed using weather indices based multiple linear regression (Jain *et al*. 1980, Gill *et al*. 2015). Weather based crop forecasting model using discriminant function analysis for different crops (Aditya and Das 2012, Agarwal *et al*. 2012, Pandey *et al*. 2015). The use of machine learning has picked up in recent times for crop yield modelling using weather and satellite data (Laxmi and Kumar 2011, Drocsch 2018, Khaki and Wang 2019, Gupta *et al*. 2021).

MATERIALS AND METHODS

Data

Yearly yield data of the potato has been collected from the released issues of yield estimates of Bureau of Applied Economics and Statistics (BAES), Department of Statistics and Program Implementation, Government of West Bengal for 43 years from 1977-1978 to 2019-20. This paper focuses on Hooghly district, which is the highest contributor in the overall production of potato in the WB.

Data on different weather variables viz., minimum temperature, maximum temperature, rainfall, and relative humidity have been collected from NASA Power data access viewer (power.larc.nasa.gov/data-access-viewer) for the study period. The weather data from pre-sowing period i.e., 40th standard meteorological week through to harvesting period i.e., 6th standard meteorological week have been taken.

Methodology

In this study we have focused on modelling the crop yield based on weather variables through weather indices using multiple linear regression and artificial neural network models using Multilayer Perceptron (MLP) architecture with resilient backpropagation algorithm.

Multiple linear regression model

The weather indices and time variable *t* were taken as independent variable and yield as dependent variable to develop a multiple linear regression model. Stepwise regression analysis have been used for selecting significant variables. The regression model is given as:

$$y_t = a + \sum_{i=1}^{p} b_i Z_{t(i)} + ct + \varepsilon$$

The indices $Z_{t(i)}$ are defined as

$$Z_{t(i)} = \sum_{w=1}^{m} r_{\{y_t, X_{tw(i)}\}} X_{tw(i)} / \sum_{w=1}^{m} r_{\{y_t, X_{tw(i)}\}}$$

where m denote the number of weeks in any particular crop season and p denote the number of weather variables taken in the study, X_{iw} (i=1, 2, …, p; w=1, 2, …,m) denote the value of ith weather variable in wth week and yt denote the crop yield for the year t (t=1,2,…,43).

Multi-layered perceptron artificial neural network model

A MLP feed forward neural network is a data driven, nonlinear fully connected network, which connect each neuron in one layer to each neuron in the other layer. It consists of a series of fully connected layers that connect each neuron in one layer to each neuron in the other layer. A feed forward network has one-way flow and no cycles. Fig.1 gives an example of a fully connected MLP with one hidden layer.

For a causal relationship problem, the information given to input layer in an ANN are the predictor variables. The functional relationship estimated by the ANN can be written as $y=f(x_1, x_2,...,x_p)$, where x_1, x_2,..., x_p are predictor variables and y is the response. The output of jth node in the neural network is given by

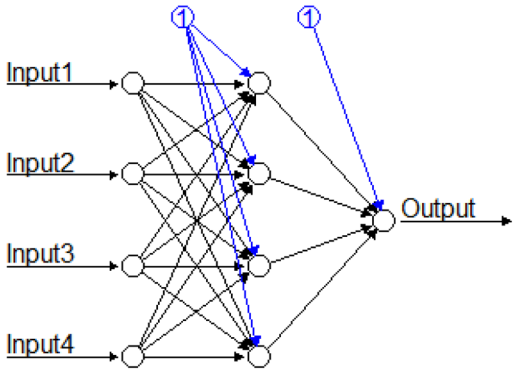

Fig. 1. A typical fully connected MLP with 1 hidden layer.

$$Output_j = g(\theta_j + \sum_{i=1}^{p} w_{ij}\, x_i)$$

where g is a transfer or activation function, θ_j is the bias of the node j, $w_{1j}, \ldots, w_{pj}$ are weights of node j and x_i (i=1,2,…,p) are the input variables. The activation function determines the relationship between input and outputs of a node and also introduces non-linearity in the model which is the core of ANNs. The sigmoid (logistic) function is the most popular activation function in neural networks (Shumeli *et al*. 2018). The sigmoid (logistic) function is given by $f(x)=1/1+e^{-x}$. The prominent task in designing multilayer feed forward neural network architecture for a prediction problem is to determine numbers of hidden layers and nodes in each hidden layer. The most common way of determining numbers of hidden layers and nodes is by trial-and-error as there is not theoretical basis till date (Zhang *et al*. 1998). As learning algorithm, the resilient backpropagation (RPROP) algorithm (Riedmiller and Braun 1993) have been used which is a faster and improved version of commonly used backpropagation training algorithm as it doesn't require specifying any learning rate.

For model fitting, the whole data was divided into training and validation set. The dataset of first 38 years was used for training and last 5 years for validation. Before training the ANNs, the dataset was scaled in [0, 1] scale and as rescaled back for comparing the predicted values. The best network was selected by considering the lowest error measures like root mean squared error (RMSE) and mean absolute percentage error (MAPE) and prediction error percentage.

RESULTS AND DISCUSSION

The weekly weather data on four weather variables were converted into weather indices to fit a multiple linear regression model. The four weather indices and time variable t were taken as independent variables and yield as output variable. The stepwise variable selection method was used and result of multiple linear regression model is presented in the Table 1. The final model after variable selection shows that there is significant effect of relative humidity (Z_{rh}), minimum temperature (Z_{mint}) and precipitation (Z_{prep}). Fig. 2 represents the fitting performance of regression model and it shows that the fitted and predicted values are deviating much from actual values.

For neural network models, only significant variables form the regression model and time variable t were taken as input variables and yield as output variables. Therefore, for the ANN models, there were total four input variables and one output variable. The number of hidden nodes were varied from one to five to find out which model best fits the dataset.

Table 1. The model structures and their error measures for each fitted model.

Model	Model structure	Training			Testing			Remark
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	
Regression	$396.375+0.394Z_{rh}-26.3792Z_{mint}+0.616Z_{prep}$	44.05	34.48	18.02	82.12	72.8	35.46	
ANN-1	4:1:1	44.54	35.57	18.16	80.21	70.83	33.29	
ANN-2	4:2:1	32.29	27.1	12.81	62.35	51.19	29.1	
ANN-3	4:3:1	23.84	18.56	8.2	53.51	44.61	17.07	
ANN-4	4:4:1	22.64	16.16	6.59	32.82	24.51	9.6	Selected model
ANN-5	4:5:1	21.21	16.72	7.4	52.06	40.6	19.17	

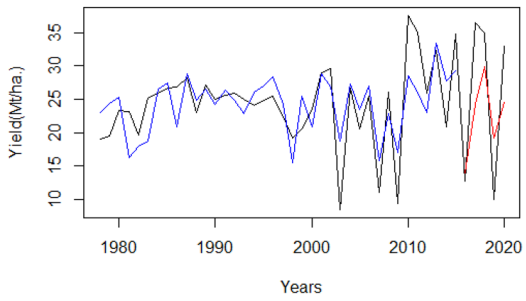

Fig. 2. The actual, fitted and predicted values from multiple linear regression model.

The results of all the ANN models are presented in Table 1. The results shown that as the number of hidden nodes increases the error measures decreases till four hidden nodes and then again increases for five hidden nodes. The lowest RMSE, MAE and MAPE for training and testing set were found for the ANN-4 model which have 4 input nodes, 4 hidden nodes and 1 out nodes. The network structure along with its weights on each neuron are presented in Fig. 3. Fig. 4 represents the fitting performance of ANN-4 model. Fig. 4 shows that the fitted and predicted values are very close to actual values which was not in case of regression model.

Each model was validated for the period of 5 years from 2016 to 2020 by compairing actual yield during this period with predicted yield from each model. The actual yield , predicted yield and corresponsing percentage errors are presented in the Table 2. These results again confirms that the multiple linear regression has higher errors while ANN-4 model has lower error percentages on avearge as compare to other models.

CONCLUSION

The weather variables viz., relative humidity (Z_{rh}), minimum temperature (Z_{mint}) and precipitation (Z_{prep}) significantly affected the potato yield in Hooghly district over the time. The comparative performance of regression model and ANN models shows that ANN models consistently performed better than regression model for crop yield prediction based on lower RMSE and MAPE. The best models can be used to obtain a reliable and timely forecast of potato yield at 6–8 weeks before harvest using the meteorological data.

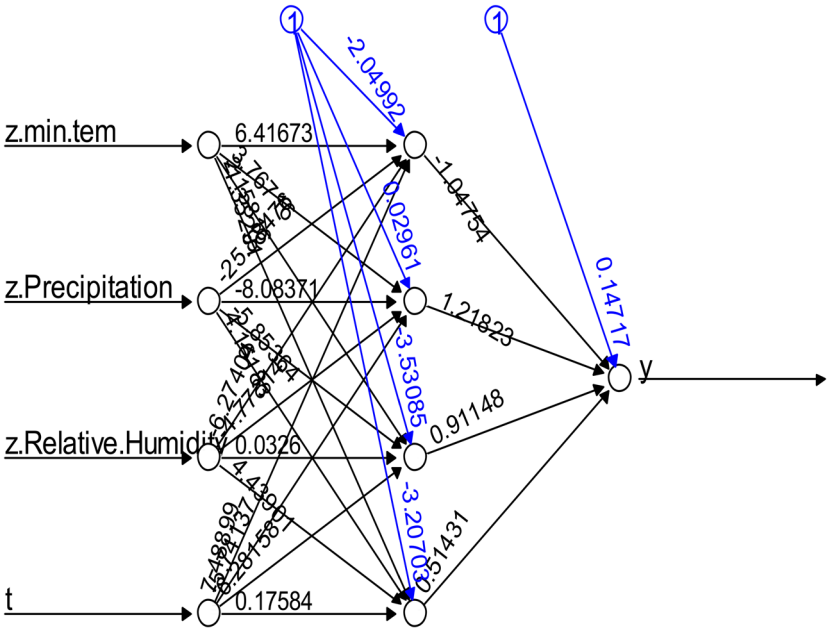

Fig. 3. The network structure of best fitted neural network model.

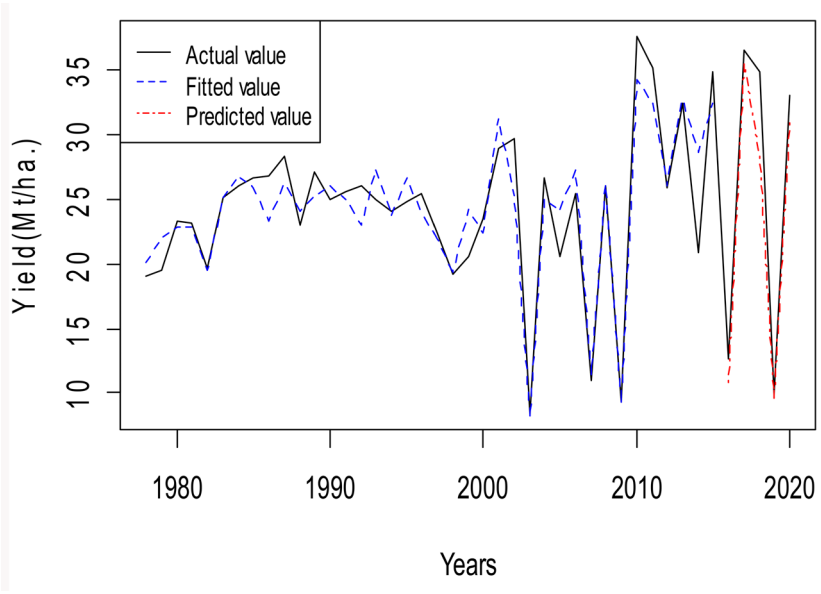

Fig. 4. The Actual, fitted and predicted values from ANN model.

Table 2. The actual yield, predicted yield and their corresponding prediction errors for each fitted model.

Year	Actual yield	Regression		ANN-1		ANN-2	
		Predicted yield	% error	Predicted yield	% error	Predicted yield	% error
2016	126.6	139.96	-9.54	133.52	-5.18	127.97	-1.07
2017	365.9	241.52	51.50	244.03	49.94	282.55	29.50
2018	349.4	299.42	16.69	288.91	20.94	315.32	10.81
2019	99.1	191.25	-48.18	184.69	-46.34	198.69	-50.12
2020	330.2	246.08	34.19	250.95	31.58	292.65	12.83

Table 2. Continued.

Year	Actual yield	ANN-3		ANN-4		ANN-5	
		Predicted yield	% error	Predicted yield	% error	Predicted yield	% error
2016	126.6	140.36	-9.80	108.38	16.81	121.78	3.96
2017	365.9	320.16	14.29	354.76	3.14	300.29	21.85
2018	349.4	251.87	38.72	282.82	23.54	267.46	30.64
2019	99.1	119.00	-16.72	94.22	5.18	149.16	-33.56
2020	330.2	284.07	16.24	308.45	7.05	329.64	0.17

REFERENCES

Agrawal R, Chandrahas, Aditya K (2012) Use of discriminant function analysis for forecasting crop yield. *Mausam* 63(3): 455-458.

Drocsch AC (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ Res Lett* 13(114003).

Gill KK, Sandhu SS, Kaur P, Babuta R, Bhatt K (2015) Wheat yield prediction using weather based statistical model in

Central Punjab. *J Agric Phys* 15(2): 157-162.
Gupta AK, Sarkar KA, Bhattacharya D, Dhakre DS (2021) Potato yield modeling based on meteorological factors using discriminant analysis and artificial neural network. *Int J Veg Sci* DOI: 10.1080/19315260.2021.2021342
Jain RC, Agrawal R, Jha MP (1980) Effect of climatic variables on rice yield and its forecast. *Mausam* 37(4): 591-596.
Khaki S, Wang L (2019) Crop Yield Prediction Using Deep Neural Networks. *Front Pl Sci* doi:103389/fpls201900621.
Laxmi RR, Kumar A (2011) Weather based forecasting model for crops yield using neural network approach. *Stat Appl* 9: (1 - 2): 55-69.
Pandey KK, Rai VN, Sisodia BVS, Singh SK (2015) Effect of weather variables on rice crop in eastern Uttar Pradesh India. *Pl Arch* 15(1): 575-579.
R version 4.1.1 (2019) R: A language and environment for statistical computing. Available at http://wwwR-projectorg/
Riedmiller M, Braun H (1993) A direct adaptive method for faster back propagation learning: The RPROP algorithm. Proceedings of the IEEE International Conference on Neural Networks (ICNN), San Francisco, pp 586-591.
Shmueli G, Bruce PC, Yahav I, Patel NR, Lichtendahl KC (2018) DATA MINING FOR BUSINESS ANALYTICS-Concepts, Techniques, and Applications in R. John Wiley and Sons, Inc.
Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. *Int J Forecast* 14: 35-62.