Reviews of Literature
Impact Factor : 1.4716 (UIF)
Volume 2 , Issue 7 / Feb 2015 ISSN:-2347-2723

# CROP YIELD FORECASTING USING ARTIFICIAL INTELLIGENCE METHODS

**Md. Atheeq Sultan Ghori**
**Assistant Professor , Department of CSE ,**
**Telangana University , NIZAMABAD.**

**ABSTRACT**

The crop production forecasting has become an important issue, now, as it is a key factor for our economy and sustainable development on account of increased demand of the food grains with growing population. It encourages agriculturists and government to build up a superior post-reap administration at nearby/territorial/national level, e.g., transportation, stockpiling, dissemination. Furthermore, it encourages agriculturists to design one year from now's product and government to design import/send out techniques. This work depends on the yield estimating of the pearl millet (bajra) in the Jaipur locale of Rajasthan, India. The proposed technique utilizes a back spread counterfeit neural system to figure current yield of the product concerning the ecological variables utilizing time arrangement information. The acquired outcomes are empowering and much better in comparison to a recent fuzzy time series based methods for forecasting.

**KEYWORDS:** Crop yield forecasting, Pearl millet, Time series, Correlation analysis, Neural network.

## INTRODUCTION

Crop yield forecasting plays an important role in farming planning and management, domestic food supply, international food trade, ecosystem sustainability, and so on. For instance, China has the largest population in the world but with limited agricultural land so accurate crop forecasting helps the government provide sufficient food supply to the people. Australia has a small population with vast agricultural land so its concern on crop production is how to optimize revenue from international crop export to countries like China. There are many factors that have an influence on crop yield, such as plantation area, efficiency of irrigation systems, variations in rainfall and temperature, quality of crop seeds, topographic attributes, soil quality and fertilisation, and disease occurrences. Crop growing follows seasonal cycles but many of the factors above are largely irrelevant to the temporal factor. For example, plantation area, rainfall, fertilising, and disease occurrence vary yearly; efficiency of irrigation systems,

quality of crop seeds, and soil quality may be improved or degraded from year to year; and topographic attributes may largely remain the same for a long period of time.

Effort has been made in using either statistics to identify relationships or neural networks to establish mappings between crop yield and some of these factors. Our recent study using historic data of wheat yield and associated plantation area, rainfall, and temperature in Queensland, Australia, has shown that incorporating statistics and artificial neural networks can produce a high level of satisfactory forecasting of wheat yield. The neural network employed in this study was a spatial model that treats the wheat plantation areas and yields as mutual mappings, rather than yearly time series. Doubts have been raised about the lack of comparison between the outcomes from this spatial neural network model and commonly used temporal neural network models in crop forecasting. To address this issue, using the wheat yield in Queensland as a reference, this paper presents our research outcomes from using both the spatial and temporal neural network models in crop forecasting. Comparison and discussion are made in terms of their usefulness in crop yield forecasting.

Forecasting is the use of historic data to determine the direction of future trends. It is an age old phenomenon and finds its application in almost every walk of life, e.g., weather forecasting, economic forecasting, energy forecasting, transport forecasting, sales forecasting, technology forecasting, crop yield forecasting. As risk and uncertainty are central to forecasting, it is not possible to forecast with 100% accuracy. Therefore, the forecasting methods aim to reduce the forecasting error and obtain the best possible forecast. The crop production forecasting is determining future value of a crop yield for any given region/country for a particular year or season. It has gained significance due to the rapidly growing population, industrialization and globalization. A successful forecast of a crop yield bears significant profits. An accurate crop yield forecast helps agriculturists away administration and getting ready for the following year's product and enables governments for better post-to collect administration regarding capacity, transportation and circulation at nearby/territorial/national level, and plan import/trade systems appropriately. It is particularly useful for us as our economy primarily depends on agriculture and agriculture based products. Moreover, almost two-third of the employed class in our country lives on the business of agriculture. In spite of this, there is a high degree of uncertainty as the agriculture is heavily dependent on rains.

The crop yield is affected by the physical factors, the economic factors, and the technological factors. The prime physical factors are temperature, humidity, rainfall, soil, and topography. For example, the moisture requirements vary from plant to plant and region to region; regions having low maximum temperature are not suitable for plant growth whereas agriculture is successful in the tropical regions, where temperature is high throughout the year. The prime economic factors are market, transport facilities, capital, labour, and government policies. For example, the supply of labour determines the character and type of agriculture; the government policies may restrict / promote a crop and influence agricultural land use. The

Government may restrict or force the cultivation of a crop. The prime technological factors are fertilizers, pesticides, machinery and high yielding variety seeds. The scientific and technological development has a great impact on the crop production. The use of primitive methods of farming results in poor farm yield whereas the use of modern farm technologies increases the farm yield substantially. For example, per hectare yield of rice in India is just 2000 kg in contrast with around 5600 kg in Japan because of logical and mechanical contrasts in cultivating.

In this investigation, we show our back spread manufactured neural system based way to deal with gauge yield of the pearl millet (bajra) in the Jaipur district of Rajasthan, India. However, in absence of vast and varied data required for forecasting, we use data related to physical factors only, which we collected through official websites and by other official means from the government of Rajasthan and the government of India. Here, we use the monthly data for every season as our primary concern is to obtain a high degree of accuracy in the forecasting. In other words, this study aims to provide a simple computational method which uses available data to the best possible extent to obtain a high degree of accuracy in the forecasting. However, this is not a maiden attempt to a novel problem (agricultural production forecasting). Kumar et al. present a fuzzy time series model and two variations of it to forecast wheat production. The results clearly indicate that the models are not appropriate as the forecasting error is high; it is probably because they use only previous years' yield to forecast. Kumar and Kumar too present a fuzzy time series model for wheat forecasting. Here, too the authors use only previous years' yield to forecast and obtain poor results.

Rest of the paper is organized as follows. Section 2 includes the data analysis part. The proposed forecasting method is presented in Section 3. Section 4 presents simulation results and is followed by the conclusion in Section 5.

## 2. FACTORS AFFECTING THE CROP PRODUCTION

An analysis of the available historical data indicates following relationship between the crop yield and the physical factors.

### 2.1 Variation of Crop Yield with Rainfall

The rainfall is one of the important factors that affect the crop production. Figure 1 represents a scatter plot of the total seasonal yield given in kg per hectare and the total seasonal rainfall (July – September). It is observed that for extremely low rainfall and

## 2. Neural Network-Based Forecasting Models

### 2.1. Nonlinear Autoregressive Neural Network (NARNN) Model

Nonlinear autoregressive (NAR) is a widely used statistical forecasting model for time series [15, 16]. The forecasting model takes the form as follows:

$$y\,(t) \approx f(y(t-1)\,, y\,(t-2)\,, \ldots, y\,(t-d))\,, \qquad \_\_\_\_\_(1)$$

where $y(t)$ is the forecasted output and $f$ is an unknown function of $d$ previous known outputs. Traditionally, function $f$ is determined by statistical optimization processes, such as the minimum mean square method.

The feed forward neural network has been used to establish NAR models, in which the traditional function $f$ is replaced by a number of neurons that work together to implicitly approximate the same functionality as

$$y\,(t) \approx f(y(t-1)\,, y\,(t-2)\,, \ldots, y\,(t-d)) = \sum_i b_i \psi\left(\sum_{j=1}^{d} a_{ji} y(t-j)\right), \qquad \_\_\_\_\_(2)$$

where $\psi$ is the transfer functions $a_{ji}$; denotes the input-to-hidden layer weights at the hidden neuron j; and $b_i$ is the hidden-to-output layer weight.

This is a time-delay and recurrent neural network model. The input is the known time series which is fed to the hidden layer as input according to the number of time delay. This model is visually illustrated in Figure 1.

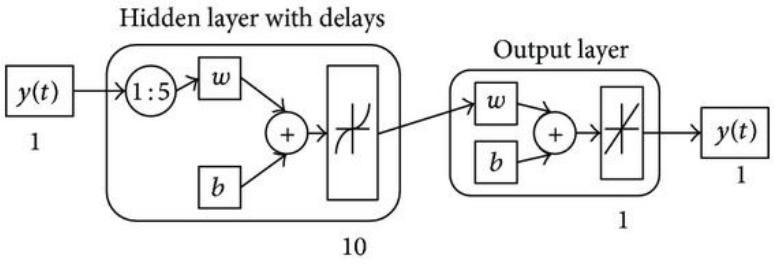



**Figure 1: Structure of NARNN with 10 hidden neurons and 5 delays.**

### 2.2. Nonlinear Autoregressive with External Input Neural Network (NARXNN) Model

Nonlinear autoregressive with external input (NARX) is a modified NAR model by including another relevant time series as extra input to the forecasting model, which can be expressed as

$$y\,(t) \approx f(x(t-1)\,, x\,(t-2)\,, \ldots, x\,(t-d)\,, y\,(t-1)\,, y\,(t-2)\,, \ldots, y\,(t-d))\,, \qquad \_\_\_\_\_(3)$$

where x(t) is the external input to the forecasting model with the same number of time delay as y(t).

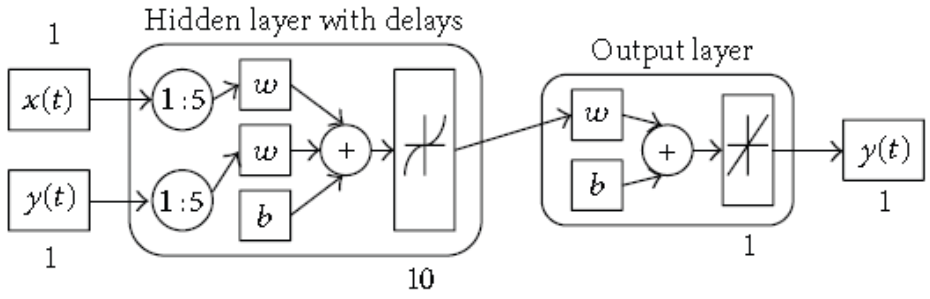



**Figure 2: Structure of NARXNN with 10 hidden neurons and 5 delays.**

Similarly the feed forward neural network is able to establish NARX models, which can be expressed as

$$y(t) = \sum_{i} c_i \psi \left( \sum_{j=1}^{d} (a_{ji} x(t-j) + b_{ji} y(t-j)) \right), \qquad \text{_____(4)}$$

where $\psi$ is the transfer functions; $a_{ji}$and $b_{ji}$ denote the inputto- hidden layer weights at the hidden neuron $j$; and $ci$ is the hidden-to-output layer weight.

This time-delay recurrent neural networkmodel uses two known time series as independent inputs to the hidden layer according to the same number of time delay. This model is visually illustrated in Figure 2.

Spatial Feedforward Neural Network Forecasting Model. Multilayer perceptron (MLP) model belongs to feedforward neural networks. In terms of functionality, MLP has no difference from the neural networks used in both NARNN andNARXNNmodels if the input is time series. Additionally MLPs have been proven to be able to approximate any continuous function by adjusting the number of nodes in the hidden layer, with numerous cases of successful applications. Figure 3 illustrates the general structure of a three-layer MLP with one hidden layer of $L$ nodes, a $p$-dimensional input vector X, and a $q$-dimensional output vector Y. The relationship between the input and output components for this MLP can be generally expressed as

$$y_k = \varphi\left(\sum_{j=1}^{d} a_{kj}\psi\left(\sum a_{ji}x_i\right)\right),$$

where $\varphi$ and $\psi$ are the transfer functions; $aji$ denotes the input-to-hidden layer weights at the hidden neuron $j$; and $b_{kj}$ is the hidden-to-output layer weights at the output unit $k$.

There are at least two relevant time series used in the NARXNN model, the internal series $[t]$ and external series $[t]$. Time series analysis emphasises on the appearance of consecutive events. However, for example, in crop yield forecasting, the current plantation area should have a much higher impact on the forthcoming crop yield than the historic yields of any past years.

Treating crop yield and plantation area as a correlated pair, MLPs have been used to approach the nonlinear relation that may exist between the two sequences in a correlated "spatial" manner, rather than a correlated temporal mode. This has resulted in some encouraging outcomes.

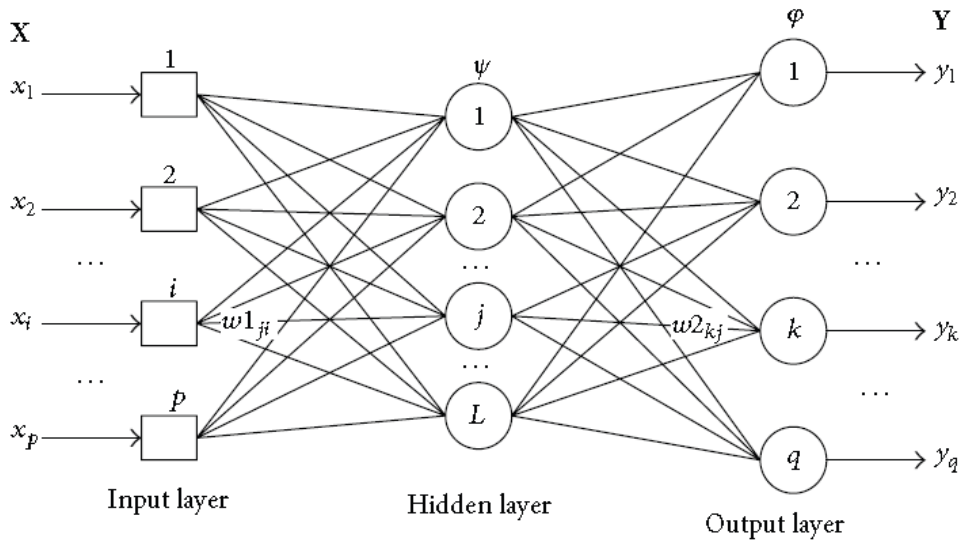
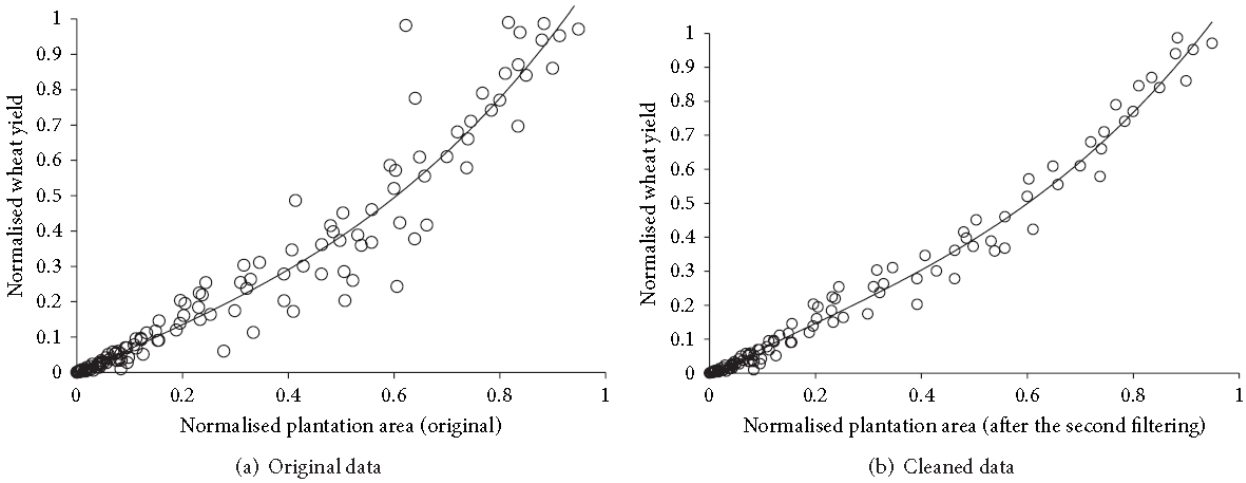



**Figure 3:Three layer multilayer perceptron (MLP) neural network.**




(a) Original data (b) Cleaned data

**Figure 4: Correlations between normalised wheat yield and plantation area.**

## 3. CROP DATA FOR NEURAL NETWORK SIMULATION

Historic Crop Data. The Queensland historic wheat plantation area in hectare and wheat yield in ton from 1861 to 2007 are extracted from the report of Australian Bureau of Statistics [25], which gives a total of 135 entries over the past 147 years. Both plantation area in hectare and wheat yield in ton are listed in their approximated absolute values each year in the original data. We normalise both factors with their ceilings in the order of millions. The ceiling for plantation area is 13million hectares and that for wheat yield is 20million tons. Plot of these two normalised factors is shown in Figure 4(a). After two rounds of outlier detection and exclusion, a third order polynomial correlation has been defined as

$$w = 0.8197a^3 - 0.5102a^2 + 0.8511a - 0.0073, (6)$$

where $w$ represents the normalised annual wheat yield and $a$ is the normalised plantation area. This correlation fits the filtered data well (Figure 4(b)) with a coefficient of 0.9904. This nonlinear correlation indicates that, through properly training, a neural network system can be used to approach such nonlinear relation between the crop production and plantation area.

Data for Training and Testing Neural Network Models. Neural network training requires a sufficient amount of data for achieving a high reliability. For MLPs, since temporal factor does not play any role in correlation, this correlation can be used to generate more data without changing the general trend. By keeping the same pattern, a moving window operator with different sizes is repeatedly applied to these cleaned data so as to generate more entries to fill the gaps where the original entries are scarce. The final data to train and test the selected neural networks are compiled by mixing the cleaned and regenerated entries together.

For both NARNN and NARXNN models, such data expansion cannot be applied because the training data must be a sequence ordered by time. Among the 135 datasets, the first three are with an interval of 5 years and thus excluded from both NARNN and NARXNN training and testing. Time series data for NARNN model is the normalised annual wheat yield by the corresponding plantation area because NARNN

**Table 1: Statistical results of neural network training and testing.**

| | Training | | | Testing | | |
|---|---|---|---|---|---|---|
| | *N* | MSE | *R* | *N* | MSE | *R* |
| MLP (100) | 329 | 0.0001 | 0.9996 | 40 | 0.0001 | 0.9943 |
| MLP (200) | 329 | 0.0001 | 0.9998 | 40 | 0.0001 | 0.9981 |
| NARNN (5 d) | 105 | 0.1035 | 0.9682 | 20 | 0.0231 | 0.8800 |
| NARNN (10 d) | 105 | 0.0069 | 0.9753 | 20 | 0.0188 | 0.9298 |
| NARXNN (5 d) | 105 | 0.0174 | 0.908 | 20 | 0.0528 | 0.6945 |
| NARXNN (10 d) | 105 | 0.0271 | 0.8574 | 20 | 0.0628 | 0.6801 |

takes only one time series as the input. To some extent, this normalised series actually absorbs the effect of plantation area on crop yield into the forecasting. For NARXNNs, the normalised annual wheat yield by wheat ceiling is the internal input [$t$] and the normalised plantation area by plantation ceiling is the external input [$t$]. If thenormalisedwheat yield by plantation area is used as the internal input [$t$], the effect of plantation area on wheat yield will be doubly accounted in the forecasting.

## 4. RESULTS OF NEURAL NETWORK SIMULATION

Two MLP models are used for training and simulation. By running the process ten times with both the 100-node hidden-layer MLP and 200-node hidden-layer MLP, the simulations produce a highly satisfactory average outcome for both training and testing (Table 1). The difference is that the latter achieves a slightly lower MSE and a higher correlation than the former but both show a high consistency between the results of training and simulation or testing.

Two NARNN models are also used in training and simulation. Both models have ten hidden neurons but with 5 delays and 10 delays, respectively. Running ten repetitions for each model has resulted in a fairly satisfactory outcome on average shown in Table 1. Since the model changes data partition for training and testing dynamically between separate runs, a highly satisfactory outcome from training does not always produce a highly satisfactory outcome from testing. In general, the result of simulation is always inferior to that of training, with the 10-delaymodel performing slightly better than the 5-delay model. Similar trends are

found from the results of training and testing the two NARXNN models, whose performance is even worse than that of NARNNs (Table 1).

**DISCUSSION**

In terms of consistency between the performances of training and testing using the same model, MLPs are able to achieve the highest consistency and hence produce the best simulation results among the three forecasting models. This is mainly because the data used to train MLPs have been subject to outlier cleaning, which means the abnormal wheat yields outside the statistical trend have no impact on the training and testing. In addition, without the temporal constraint, the expanded dataset ensures that the MLPs are adequately trained and tested with multiple crossing validations. Since the original data have been cleaned, in theory the MLPs should only be effective for crop forecasting of any "normal" year.

NARNNs exhibit a highly satisfactory performance in training but the simulation is highly dependent on the selection of testing dataset; hence, the range of forecasting error is large. This indicates that a well trained NARNN model is not able to produce consistently accurate forecasting. This inconsistency between the training and testing is clearly illustrated in Figure 5.Our experiments also show that changes in number of hidden neuron and length of delay (>3) for NARNN do not make significant improvement to the performance of forecasting. Although the NARNNs are not consistent in forecasting, they use the whole data without excluding "abnormal" datasets in both training and testing. This is a complement to MLPs to some extent.

NARXNNs exhibit similar inconsistent patterns between training and testing but even worse than NARNNs (Figure 6). This may be caused by the double impacts on forecasting exerted by the "anomalies" in both the wheat yield series and plantation area series, which were not excluded through data cleaning.

Data cleaning for both NARNN and NARXNN is very challenging since both models use time series as the input. Excluding some temporal events will leave irregular gaps in the time series, which in turn influences the training and testing. The other possible reason that contributed to the inconsistency between the training and testing of both NARNN and NARXNN may be the inadequacy of the historic data in this case. Since we cannot artificially create extra yearly crop yields, like using interpolation to generate extra spatial datasets [9, 26], using time series based NN models to forecast crop yield may be immature at this stage.

**CONCLUSION**

The spatial NN model is able to predict the wheat yield with respect to a given plantation area with a high accuracy, compared with the temporal NN models such as NARNN and NARXNN. However, the high accuracy of the spatial NN model in crop yield forecasting is only applicable to the forecasting of crop yield within normal ranges because the model is trained using the cleaned and expanded data following a third-order polynomial trend between

the crop yield and plantation area. NARNNs may be used as a complementary means to the MLPs due to its usage of the whole data. Users must be cautious when using either NARNN or NARXNN for crop yield forecasting due to the inconsistency between training and forecasting.

In the future, other factors, such as efficiency of irrigation systems, variations in rainfall and temperature, quality of crop seeds, topographic attributes, soil quality and fertilisation, and disease occurrences, should be incorporated in forecasting model building and simulation.

**REFERENCES**

1. G. Hoogenboom, "Contribution of agrometeorology to the simulation of crop production and its applications," Agricultural and Forest Meteorology, vol. 103, no. 1-2, pp. 137–157, 2000.
2. Y. W. Jame and H. W. Cutforth, "Crop growth models for decision support systems," Canadian Journal of Plant Science, vol. 76, no. 1, pp. 9–19, 1996.
3. C. A. Campbell, R. P. Zentner, and P. J. Johnson, "Effect of crop rotation and fertilization on the quantitative relationship between spring wheat yield and moisture use in southwestern Saskatchewan," Canadian Journal of Soil Science, vol. 68, no. 1, pp. 1–16, 1988.
4. W. W. Guo and H. Xue, "An incorporative statistic and neural approach for crop yield modelling and forecasting," Neural Computing and Applications, vol. 21, no. 1, pp. 109–117, 2012.
5. M. M. Li, W. Guo, B. Verma, K. Tickle, and J. O'Connor, "Intelligent methods for solving inverse problems of backscattering spectra with noise: a comparison between neural networks and simulated annealing," Neural Computing and Applications, vol. 18, no. 5, pp. 423–430, 2009.
6. E. Bojórquez, J. Bojórquez, S. E. Ruiz, and A. Reyes-Salazar, "Prediction of inelastic response spectra using artificial neural networks," Mathematical Problems in Engineering, vol. 2012, Article ID 937480, 5 pages, 2012.
7. Y. E. Shao, "Prediction of currency volume issued in Taiwan using a hybrid artificial neural network and multiple regression approach," Mathematical Problems in Engineering, vol. 2013, Article ID 676742, 9 pages, 2013.
8. E. Bojórquez, J. Bojórquez, S. E. Ruiz, and A. Reyes-Salazar, "Prediction of inelastic response spectra using artificial neural networks," Mathematical Problems in Engineering, vol. 2012, Article ID 937480, 5 pages, 2012.