International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 4, No. 2 (Nov-2017)

Comparative Study of District Level Rice Production under Artificial Neural Network and Multiple Linear Regression Model

Neeraj Tiwari[1] and Ankuri Agarwal[1]

[1]Kumaun University, SSJ Campus, Almora- 263601, Uttarakhand, India

Received May 14,2017, Revised September 18, 2017, Accepted October 15, 2017, Published November 1, 2017

Abstract: Using Artificial Neural Network (ANN) methodology a district level prediction of rice production in Uttarakhand state of India has been proposed. The proposed model has been empirically compared with the existing Multiple Linear Regression (MLR) Model. Analysis of data obtained from a survey carried out from Directorate of Economics and Statistics, Uttarakhand, India revealed that ANN methodology performs better as compared to MLR model in terms of R- square value, Mean Square Error (MSE) value and Root Mean Square Error (RMSE) values. This approach does not require any additional survey or conducting extra crop cutting experiments for crop production estimate at the district level.

Keywords: ANN Model , MLR Model, Small area estimation.

1. INTRODUCTION

Small Area Estimation (SAE) techniques assist state and local governments in making various decisions such as how to allocate resources in small areas. This technique overcomes the problem of small sample sizes to produce small area estimates and also improves the quality of the direct survey estimates obtained from the sample in each small area. SAE became a sound statistical procedure to create the estimates and an evaluating system to ensure the reasonable estimates for small areas.

Various researchers have attempted to deal with the different techniques of small area estimation. Earlier reviews on the topic of small area estimation focused on demographic methods for population estimation in post censual years. Purcell and Kish [13] reviewed demographic methods as well as statistical methods of estimation for small domains. An excellent review provided by Zidek [10] introduced a criterion that can be used to evaluate the relative performance of different methods for estimating the populations of local areas. McCullagh and Zidek [14] elaborated this criterion more precisely. Reviews by Rao [9] and Chaudhuri [1] covered more recent techniques as well as traditional methods of small area estimation. Bellow and Lahiri [11] discussed an empirical best linear unbiased prediction approach to small area estimation of crop parameters assuming linear mixed model. Chandra et al. [8] employed small area estimation technique to derive model based estimates of proportion of poor households at district level in the state of Uttar Pradesh in India. Elazer [4] used the Lagrange Multiplier method to adapt the PQL (Penalised quasi likelihood) approach applied to random effect logistic models. Tzavidis et al. [12] described an application of small area estimation to agriculture business survey data by using empirical best linear unbiased predictor (EBLUP) and model based direct estimators. Chen and Lahiri [15] discussed an alternative method to overcome the inefficiency of design-based methods for making inferences about small area proportions for rare events. Sud et al. [16] demonstrated an application of spatial information to estimate the production at district level for the state of Uttar Pradesh in India.

The present study was carried out to describe computer-based SAE approach for estimating the production of Rice at District level. In this paper we propose a statistical model known as Artificial Neural Network Model that link the variable of interest with auxiliary information available from various sources. The performance of the proposed model is compared with the existing Multiple Linear Regression (MLR) model by making a comparison between predicted and measured values of the variable of interest. Results of the study suggest that the proposed model appears to perform better than the MLR model. In section 2, preliminaries about MLR model are discussed. In Section 3, we discussed the

E-mail address: kumarn_amo@yahoo.com, Ankuri.stats21@gmail.com

basics about ANN model. Section 4 proposes a new Artificial Neural Network (ANN) model for small area estimation. The model specifications and analysis on the basis of an empirical study are discussed in this section. In section 5, the comparison of ANN model with MLR model is carried out to demonstrate the utility of the proposed model. Concluding remarks and future avenues of research are presented in Section 6.

2. MULTIPLE LINEAR REGRESSION (MLR) MODEL

Multiple regression analysis is a multivariate statistical technique used to examine the relationship between a single dependent variable and a set of independent variables, whose values are known to predict the single dependent variable. Several MLR methodologies for estimation of crop production at Block/Panchayat level have been made. Panse et al. [17] introduced two phase sampling to provide crop yield estimates. Singh [6] introduced the method of double sampling in agriculture. Stasny et al. [7] presented a technical report on county estimates of wheat production in Ohio state university, Ohio. Sisodia and Singh [3] developed an estimator for crop production at block level using crop cutting and other related information at district level. Sisodia and Singh [2] described scale down approach using MLR model to obtain block level estimates from the district level crop production as follows:

Let the model is

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \,, \qquad (1)$$

where Y_i is the crop production in the i^{th} year (i=1, 2, ..., n), X_{ij} is the value of the j^{th} auxiliary variable (j=1, 2, ..., p) in the i^{th} year, $\beta' = (\beta_0, \beta_1, ..., \beta_p)$ is the vector of unknown parameters and ε_i is error term. It is assumed that ε_i follows normal distribution with mean 0 and variance σ^2. Let the fitted model be denoted as

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \hat{X}_{ij} \,, \qquad (2)$$

Where $\hat{\beta}_j's$ are the least square estimate of $\beta_j's\ (j = 0, 1, 2, ..., p)$ and $\hat{Y}_i$ is the estimated crop production in the ith year.

Following Montgomery and Peck (1982), the sum of square due to regression is decomposed to define weight for the jth auxiliary variable on the basis of its relative contribution in the model, given by

$$w_j = \frac{sum\ of\ \ square\ due\ to\ \ j^{th}\ \ auxiliary\ \ \text{variable}}{sum\ of\ \ square\ due\ to\ regression}$$

Using these weights, an estimator of crop production for a block in a given year is constructed as follows:

$$\hat{Y}_q = \left[\sum_{j=1}^{p} w_j \ x_j(q) \right] \hat{\overline{Y}} \,; q = 1, 2, ..., Q$$

where Q is the total number of blocks in the given district and $x_j(q)$ is the value of the jth auxiliary variable in the q^{th} block in the given year and $\hat{\overline{Y}}$ is the estimated average yield of a crop based on the fitted model (2) in the same year .

In order to find out the relative contribution of individual auxiliary variables, the auxiliary variables are included in the model depending upon the order of magnitude of correlation coefficients between auxiliary variables and crop production (Y). On the basis of their relative contribution, the values of weights were calculated using SPSS16.0.

3. ARTIFICIAL NEURAL NETWORK (ANN) MODEL

The ANN is an information-processing system that consists of a graph representing the processing system as well as various algorithms that assess the graph. To develop ANN model, the data set consists of input variables and an output variable. The input variable represents the related parameters that influence the production of the commodity under study. In order to get the best outcome from the network model, the data has to be expressed in a deterministic way by dividing the collected data into three parts, namely,

(i). Training set: Used to adjust the weights on the neural network.

(ii). Validation set: Used for the verification of the network.

(iii). Test set: Used for the prediction, from where the prediction accuracy is determined.

The training data presented to an ANN are randomly partitioned into two samples, one for adjusting the weight in the ANN model and another for testing the accuracy of the ANN model during the training process. The testing sample is used to prevent over fitting, which occur if an ANN begins to model sample specific characteristics of the training data that are not representative of the population from which the data was drawn.

In ANN approach, the model is based on the system of human neurology and power of neural network lies in the fact that it has an ability to represent both linear and nonlinear relationships and to learn these relationships directly from the data being modeled. The word network in ANN model refers to the interconnection between the neurons. These neurons act like parallel processing units. An artificial neuron is a unit that performs a simple operation on its inputs and imitates the function of neurons. The mathematical form of this model is represented as,

$$y_k = \sum_{j=1}^{m} x_j \; w_{kj} + b_k \; ,$$

where, y_k = response variable
x_j = auxiliary variable
b_k = bias term
and w_{kj} = synaptic weight of k^{th} neuron of j^{th} auxiliary variable.

Multilayer feed forward artificial neural networks are multivariate statistical models used to relate m predictor variables to n response variables. The model has several layers, each consisting of either the original or some constructed variables. The most common structure involves three layers: the input layer consisting of the original predictors, the hidden layer comprising of a set of constructed variables, and the output layer made up of the responses. Each variable in a layer is called a node. The transformation function is either sigmoid or linear, and is known as activation function or transfer function. ANN model includes a set of synapses that can be characterized by a weight. The weight associated with each synapse is known as synaptic weight.

4. MODEL SPECIFICATION AND ANALYSIS: AN EMPIRICAL STUDY

Time series data on production of rice, area under rice, irrigated area under rice and fertilizer consumption pertaining to the period 1990-91 to 2002-03 for Uttarakhand state of India were taken from the Bulletin of Agricultural statistics, published by the government of Uttarakhand, India. The district wise data on auxiliary variables in Uttarakhand state of India were also taken for the year 2000-01 from the Bulletin of Agricultural statistics, published by government of Uttarakhand, India. The purpose of this study is to model the time series data by using statistical technique and the neural networks and then compare the results of these two techniques.

The proposed ANN was trained using different sets of data. The detailed case processing summary of the ANN is presented in Table 1.

TABLE 1. MODEL SUMMARY

Training	Sum of Squares Error	.084
	Relative Error	.012
	Stopping Rule Used	Maximum number of epochs (100) exceeded
	Training Time	00:00:00.008
Dependent Variable: y		

During the training phase, the system adjusts its connection/weights strength in favor of the inputs that are most effective in determining a specific output.

Various numbers of hidden layers and numbers of processing units in hidden layers of the neural network topology were tested to find out the right combination of processing units and a number of hidden layers in order to solve the problem. The ANN model was developed by using statistical software SPSS 16 with two hidden layers. Linear trend gives the value of R square as 0.98. General Network information of the ANN (2, 3) model is shown in Table 2.

TABLE 2. GENERAL NETWORK INFORMATION OF ANN (2, 3) MODEL

Input Layer	Factors	1	TA
		2	IA
		3	F1
	Number of Units[a]		45
Hidden Layer(s)	Number of Hidden Layers		2
	Number of Units in Hidden Layer 1[a]		2
	Number of Units in Hidden Layer 2[a]		3
	Activation Function		Sigmoid
Output Layer	Dependent Variables	1	y
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares
a. Excluding the bias unit			

For developing this model sigmoid function is used as an activation function for hidden layer H1 and H2 and Identity function for the output layer.

In ANN (2, 3) model, three parameters i.e. total area (TA), irrigated area (IA), and fertilizer used (F1), are included in the study to predict the production of rice for small area. The relative importance of these three auxiliary variables that influence the production of rice is calculated by using statistical software SPSS-16 and is presented in the Fig 1 and the numerical value is presented in the Table 3.

TABLE 3. INDEPENDENT VARIABLE IMPORTANCE

	Importance	Normalized Importance
TA	.307	86.9%
IA	.354	100.0%
F1	.339	95.9%

From Table 3, it is clear that irrigated area under rice production (IA) has maximum importance as compare to other auxiliary variables.

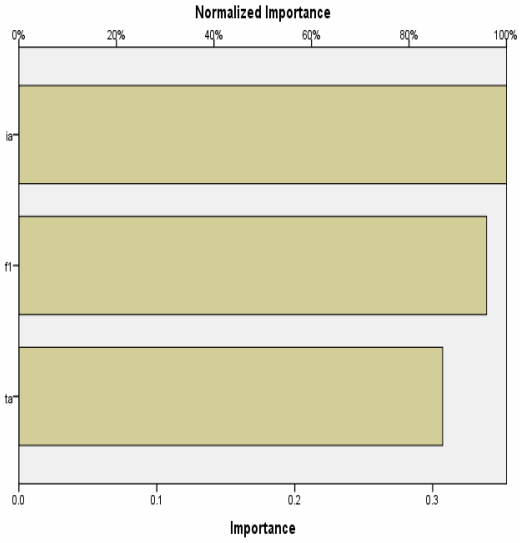

Figure 1. Independent Variable Importance

Irrigated area under Rice production shows 100% normalized importance in the proposed ANN model.

5. COMPARISON OF PROPOSED ARTIFICIAL NEURAL NETWORK MODEL WITH MULTIPLE LINEAR REGRESSION MODEL FOR PRODUCTION OF RICE AT DISTRICT LEVEL

If Y, x_1, x_2 and x_3 denote respectively the actual rice production, area under production, irrigated area and fertilizer used, then the MLR model for estimating the production of rice based on given auxiliary variables was obtained as follows-

$$Y = 15203.886 + 2.341\ x_1 - 0.704\ x_2 + 0.001\ x_3$$

$$(R^2 = 0.529,\ n = 13,\ p = 0.028)$$

The null hypothesis for the model is that the coefficients equal to zero (no effect). A predictor that has a low p-value is a meaningful addition to the MLR model because changes in the variable values of predictor are related to changes in the response variable. For testing null hypothesis p-value or probability value is used in order to quantify the idea of statistical significance of each term. The p-values are represented in Table 4.

TABLE 4. P-VALUES IN THE MLR MODEL

Variables in the MLR Model	Intercept	Total Area (TA)	Irrigated Area (IA)	Fertilizer Used (F1)
p-values	0.046	0.001	0.029	0.032

From Table 4, it is clear that the p-value for each term is <0.05, which indicates that the null hypothesis is rejected at 5% level of significance. This indicates that the predictors (Total area, irrigated area, fertilizer used) used in the MLR model effects the production of Rice.

The values of Mean Square Error (MSE), Root Mean Square Error (RMSE) and R^2 for this model are given in Table 5.

TABLE 5. MLR MODEL TO ESTIMATE THE PRODUCTION OF RICE AT DISTRICT LEVEL

Model	**MSE**	**RMSE**	**R^2**
MLR	1.394E9	37336.31	0.529

The value of multiple correlation coefficient was found to be significant at 5% level of significance. The value of R^2 was obtained as 0.529, indicating 52.9% variation in production of rice due to the joint effect of all the three auxiliary variables included in the study.

On the basis of data related to the production of rice and various auxiliary variables, the production of rice at district level was estimated using Artificial Neural Network Model and Multiple Regression Model. The actual and the estimated values of rice production in year 2001 using the proposed ANN model and Multiple Linear Regression Model are represented in Table 6.

TABLE 6. ACTUAL AND ESTIMATED VALUES OF RICE PRODUCTION IN THE YEAR 2001

2001							
Sl.No	District	Dependent Variable	Independent Variable			Two Estimates for Rice Production	
		Actual Rice Production (Y)	Area Under Production (x_1)	Irrigated Area (x_2)	Fertilizer Used (x_3)	Estimated Rice Production using MLR Model	**Estimated Rice Production using ANN Model**
1	Chamoli	13503	13364	1352	27	24936.9	**9355.001**
2	Dehradun	24396	21616	11724	2235	31692.36	**21625.7**
3	Haridwar	45520	22633	20334	15696	51604.39	**45972.7**
4	Pauri	25051	9307	6905	49	51496.18	**22582.35**
5	Rudraprayag	11419	12750	2351	25	21095.21	**8875.147**
6	Tahari	20753	10316	7176	71	29553.5	**15548.7**
7	Uttarkashi	14861	20696	4896	90	23761.5	**11833.69**
8	Almora	25969	14239	5656	136	46981.04	**20164.8**
9	Bageshwar	17553	11458	5424	141	32577.19	**15214.24**
10	Champawat	13187	15546	2118	109	25843.04	**10330.39**
11	Nainital	37644	23070	13446	3217	36841.05	**25495.37**
12	Pithodaghar	28672	94753	4501	175	52072.11	**20964.4**
13	Udham Singh Nagar	289134	11108	94193	42182	226787.2	**184159.7**

From Table 6, it is clear that the estimated values of the rice production using ANN model (given in bold letters in the last column of the Table 6) are quite closer to the actual value(given in the third column of the Table 6) as compared to the estimated values obtained through the MLR model. To make a numerical comparison between the two models, we study the performance of the proposed ANN model using R^2, Mean Square Error (MSE) and Root Mean square error (RMSE) values. The values of these three statistical indexes for comparing the performance of ANN (2, 3) Model with the MLR Model are presented in Table 7.

TABLE 7. VALUES OF STATISTICAL INDEXES

Model	MSE	RMSE	R^2
ANN Model	2.598E6	1611.831	0.989
MLR Model	1.394E9	37336.31	0.529

From Table 7, we observe that the MSE and RMSE for the ANN model are smaller than those for the MLR Model. Also the determination coefficient R^2 for ANN model is around 0.989. From the present study of the production of Rice at district level using state level data shows that the analysis using neural network yields better results than those obtained through the regression method.

Various number of hidden layers and number of processing units in hidden layers of the Neural Network topology were tested to find out the right combination of processing units and number of hidden layers in order to solve the problem with acceptable training times. This varied number of hidden layers and processing units give rise to different models. The experiment was started with small number of processing units in two hidden layers H1 and H2 (1 processing unit in each hidden layer) and then by increasing this number by growing method. ANN (2, 3) with R^2 value as 0.989 gives the best fit. The predicted results of two neurons in H1 and three neurons in H2 are presented in Fig 2. In Multiple Linear Regression analysis, stepwise method was used to find the important characters contributing to production of rice at district level in Uttarakhand, India. Among all the parameters, the total area under production plays an important role in district level rice production. The R^2 curve for MLR model is represented in Fig 3.

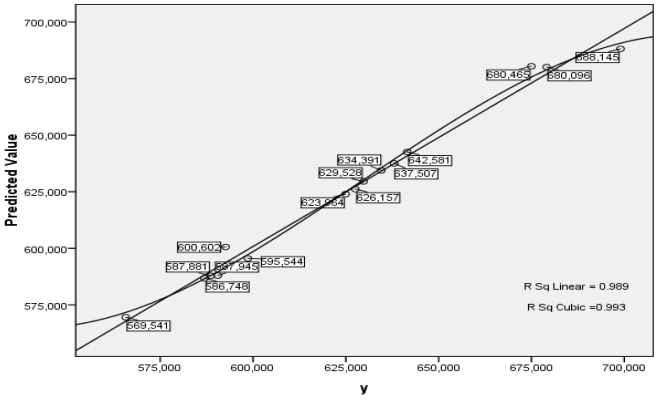

Figure 2. ANN R- Square value curve

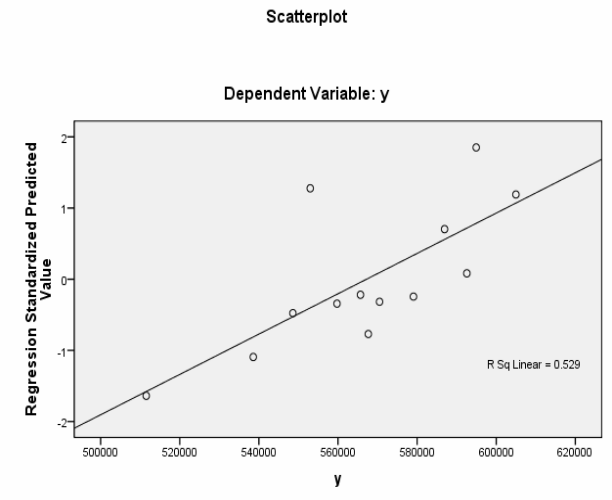

Figure 3. MLR R- Square value curve

The present study indicates ANN model provide better small area estimates than MLR model. The ANN model may be recommended for use in small area estimation.

6. CONCLUSION AND DISCUSSION

In this paper we demonstrate an application of small area estimates using ANN model. Time series data on production of rice, area under rice, irrigated area under rice and fertilizer consumption pertaining to the period 1990-91 to 2002-03 for Uttarakhand state of India were taken from the Bulletin of Agricultural statistics, published by the government of Uttarakhand. The district wise data on auxiliary variables in Uttarakhand state were also taken for the year 2000-01 from Agricultural bulletin, published by government of Uttarakhand, India.

It has been concluded that the estimates of district level rice production obtained from the MLR model are far away from their actual district level rice production as compare to the estimated production obtained through the proposed ANN model. The ANN models are now widely used for obtaining estimates in different fields, such as image classification for small areas, function and fitness approximation for small areas and data processing including clustering, etc. This paper demonstrates the utility of ANN model for small area estimation. On the basis of the results obtained through the empirical study, it may be concluded that the ANN model performs better than the existing MLR model for small area estimation.

ACKNOWLEDGMENTS

The authors are grateful to the editor and a referee for his/her constructive comments, which have led to considerable improvement in presentation of this manuscript.

REFERENCES

[1] A. Chaudhuri, "Small domain statistics: a review, Technical report ASC/92/2," Indian Statistical Institute, Calcutta,1992.

[2] B.V.S. Sisodia, and A. Singh, "On small area estimation Techniques- An application in agriculture," Indian Society of Agricultural Statistics, New Delhi, vol. 66(3), pp. 391-400, 2012.

[3] B.V.S. Sisodia, and A. Singh, "On small area estimation- An empirical study," Indian Society of Agricultural Statistics, New Delhi. vol.54(3), pp. 303-316, 2001.

[4] D. Elazar, "Small area estimation under additive constraints to published direct survey estimates, "Australian Bureau of Statistics, vol. 66, pp. 15-30, 2012.

[5] D. C. Montgomery, and E. A. Peck, "Introduction to Linear Regression Analysis," John Wiley & Sons, Canada, 1982.

[6] D. Singh, "Double sampling and its application in agriculture, Contributions in statistics and Agricultural Sciences. Dr. V.G lecture," Indian Society of Agricultural Statistics, New Delhi, 1968.

[7] E. A Stasny, P. K Goel, and D. J Rumsey, "County estimates of wheat production. Technical Report 458, Department of Statistics," The Ohio state University, Columbus, Ohio, 1991.

[8] H. Chandra, "Small area estimation with binary variables," Indian Agricultural Statistics Research Institute New Delhi, vol. 64, pp.367-374, 2010.

[9] J. N. K. Rao, "Synthetic estimators, SPREE and best model based predictors," Conference on Survey Research Methods in Agriculture, **1-16**. U S. Dept. Agriculture, Washington, D C, 1986.

[10] J. V. Zidek, "A review of methods for estimating the populations of local areas," Technical Report 824, University. British Columbia, Vancouver, 1982.

[11] M. E. Bellow, and P. S. Lahiri, "An empirical best linear unbiased prediction approach to small area estimation of crop parameters," National Agricultural Statistics Service, University of Maryland, 2011.

[12] N. Tzavidis, R. Chambers, N. Salvati, and H. Chandra, "Small Area Estimation in Practice: An Application to Agricultural Business Survey Data," Journal of the Indian Society of Agricultural Statistics, vol.66, pp. 213-228, 2012.

[13] N. J. Purcell and L. Kish, "Estimation for small domain," Biometrics, vol.35, pp. 365-384, 1979.

[14] P. McCullaghand Zidek, "Regression methods and performance criteria for small area population estimation," Small Area Statistics, Wiley, New York, vol. 62, pp. 74, 1987.

[15] S. Chen , and P. Lahiri, "Inference on Small Area Proportions, Joint Program of Survey methodology," University of Maryland, College Park, USA. 2012.

[16] U.C. Sud, K. Aditya, and H. Chandra, "District Level Crop Yield Estimation under Spatial Small Area Model," vol. 69(1), pp. 49-56, 2015.

[17] V.G. Panse, M. Rajogopalam, and S. Pillai, "Estimation of crop yields for small areas," Biometrics, vol. 66, pp. 374-388, 1966.