Pak. J. Agri. Sci., Vol. 52(1), 239-256; 2015
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
http://www.pakjas.com.pk

COMPARISON OF REGRESSION MODELS TO PREDICT POTENTIAL YIELD OF WHEAT FROM SOME MEASURED SOIL PROPERTIES

Muhammad Naveed Aman[1,*] and Aman Ullah Bhatti[2]

[1]FAST National University of Computer and Emerging Sciences, Peshawar, Pakistan
[2] University of Agriculture Peshawar, Pakistan
[*]Corresponding author email, s email :drbhattises@gmail.com

Knowledge of potential yield of wheat is imperative for site specific fertilizer management. Data collected from field trials conducted on wheat in Khyber Pakhtunkhwa (KPK) Province, Pakistan was used to predict potential yield of wheat. Various regression models were employed to get the predictions. A complete diagnostic analysis of the residuals of each model is presented. Multiple regression models give us limited prediction power for our data. Models like Classification and Regression Trees (CART) and Random Forests are also explored. The models created are compared on the basis of predictive power and miss-classification error rates. Our results revealed that Random Forests give us very good results if yield is divided into three categories.
Keywords: Regression models yield potential soil properties

INTRODUCTION

Yield potential of a crop is not known under certain soil and environmental conditions, which can be of great use in formulating fertilizer requirements. To get full benefits from various technological inputs such as fertilizers, improved crop varieties, pesticides, and better agricultural practices, the soils must be managed according to the yield potential of a crop on a particular soil. Potential yield of crops can be predicted using different variables such as soil properties and weather data. Crop yield is affected by various factors: i) nutrients availability in the soil, ii) inputs that are under the discretionary control of the farmer such as variety, crop rotation, and weed control, iii) soil properties that are known or can be measured but which are not under the control of the farmer, and iv) climatic factors that are not known with certainty and can not be controlled by the farmer such as rainfall.

Many research workers established relationship between soil properties and wheat yields, and used these empirical relationships to determine potential wheat yield (Legget, 1959; Pawson et al., 1961; Khan and Akbar, 1990; Burleigh et al., 1991; Bhatti and Mulla, 1992; Bakhsh et al., 1994; Bhatti et al., 1997; Bhatti et al., 1998a, b, c).

Two review articles dealing with the inclusion of soil fertility variables in response analysis have been published by Nelson et al. (1985), and Nelson (1987). Sain and Jaurgui (1993) also developed a flexible functional form model for deriving fertilizer recommendations using soil variables, previous crop and rainfall data (Mombeila et al., 1981). Makowski et al. (2001) also used different statistical methods predicting responses to applied nitrogen and calculating optimal nitrogen rates..

In the previous work, multiple regression models were developed to predict wheat yield from soil data obtained from various experimental sites [Bhatti et al., 1998c] However these lack a complete diagnostic analysis, and do not consider the other prediction and classification models such as Classification and Regression Trees (CART) and Random Forests. Keeping in view the of importance of knowledge of potential wheat yield for specific site fertilizer management, the present study was carried out to compare various regression models to predict potential wheat yield from some measured soil properties.

MATERIALS AND METHODS

Yield data from 55 simple fertilizer trials conducted on farmers' fields using Pirsabak-85 wheat variety in different irrigated areas of Khyber Pakhtunkhwa province of Pakistan (Table 1) were used for modeling.

Table 1. Detail of Fertilizer Trials

District	Number of Trials		
	1992-93	1993-94	Total
Mardan	5	6	11
Swat	7	6	13
D.I. Khan	11	11	22
Peshawar	3	-	3
Kohat	3	-	3
Bannu	3	-	3
Total	32	23	55

Two fertilizer rates were used: 120-90-60 and 60-45-0 kg N-P_2O5-K_2O ha^{-1} respectively. The area of each field trial was

1000 m^2. Grain yield was recorded on hectare basis in each trial after threshing of wheat.

Soil data obtained from these trials included organic matter (0.09 to 1.7 %), soil pH (5.97 to 8.88), lime content (3.1 to 24.2 %), AB-DTPA extractable P (1.6 to 15.6 mg kg^{-1}), K (42 to 250 mg kg^{-1}, and total mineral soil N (5.6 to 53.76 mg kg^{-1}).

Different regression models viz., Multiple regression, CART and Random Forest Models (Kunter et al., 2004) were developed to predict [Bhatti et al., 1998c]ict and classify yield based on measured soil properties. All the Regression models presented in this paper have been created using the open source software *R* Ver. 2.10. The graphs and table presented have also been created using *R*.

RESULTS AND DISCUSSION

Residual analysis of previous models: In this Section we will present the residual analysis of the model presented in Bhatti et al., (1998c) which is given in equation 1.

Equation 1

We shall call this model as *Model 1*. The R^2_{adj} for Model 1 is approx. 25 %. Some of the Diagnostic plots are shown in Fig. 1.

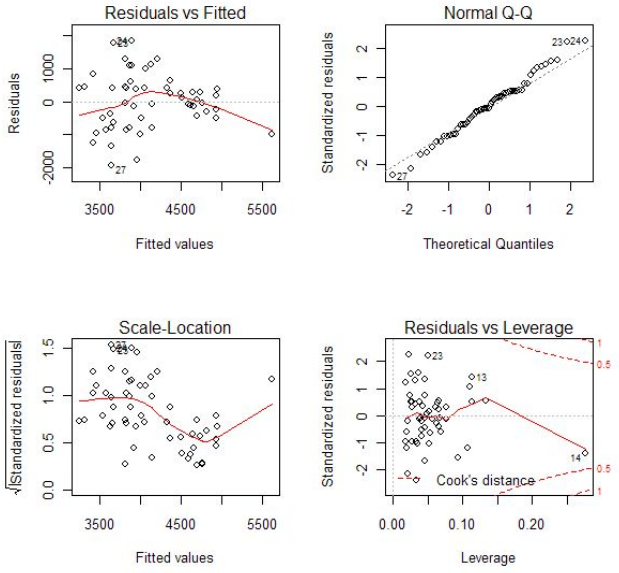

Figure 1Diagnostic Plots for Model 1

Figure 1 and some other statistical tests (results not shown),reveal no significant anomalies in the model in terms of normality of error terms, and constancy of error variance. To get an insight into the independence of error terms, the time series plot and the Auto-correlation plot for the residuals is shown in figure 2. Figure 2 shows that there is some autocorrelation among the residuals. To get further insight the Durbin-Watson test was performed on the residuals which gives us a test static of 1.3585, which at a significance level of 0.05 % proves that the residuals are auto-correlated. This autocorrelation leads to a number of problems including inefficient regression coefficients, inaccurate estimates of error variance etc. The auto-correlation is usually due to an important predictor that is missing in the model, we will explore additional predictor options in section 2.

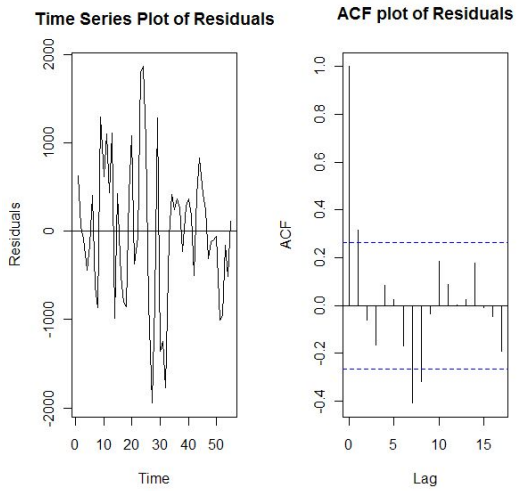

Figure 2 . TS Plot and ACF Plot of Residuals for Model 1

A partial F-test was carried out to see if the other predictors can be dropped from the model and it was observed that the other predictors can be dropped from the model. The partial F-test is used to test the hypothesis of whether the interaction terms are significant or not. The test indicated that the interaction terms are significant and can not be ignored. So based on this result, the interaction term i.e., lime content* soil pH was included in the model.The interaction, and polynomial terms are also explored in section 2.

The variance inflation factors for the coefficients are given in Table 1, which shows that there is no significant multicolinearity among the predictors.

Table 2 Variance Inflation Factors for Model 1

	Lime Content	Soil PH
VIF	1.035	1.035

This diagnostic analysis shows that model 1 is doing a poor job in terms of predictions (25 %) as well as in terms of inferences.

An improved multiple regression model: In this section we develop a new model based on the data given in Bhatti et al., [1998c]. The new model has better prediction power and inference capabilities than model 1.

Added variable plots for Model 1 were created, figure 3 shows some of these plots.

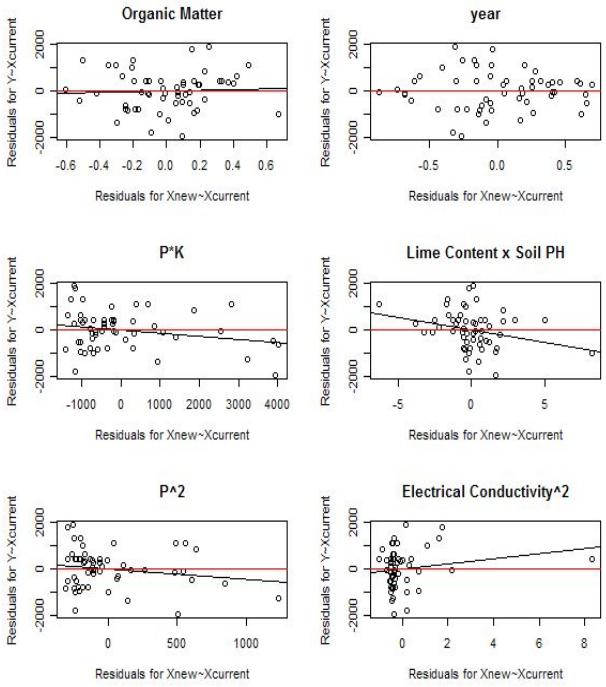

Figure 3. Added Variable Plots for Model 1

The added variable plots show that an interaction term of *Lime Content* and *Soil pH*, seems to have some useful additional information. To confirm this result a partial F-Test was performed using the extra sum of squares. The hypothesis that were tested are

H_0 = *Coefficient of Interaction term is zero*

H_1 = *Coefficient of Interaction term is not zero*

The Partial F-test concludes H_1 (Calculations not shown). Thus this shows that an interaction regression model should be used. Due to the interaction term observations that are centered around their mean were used i.e , where x_i is the *i*th centered observation. Centering is used as it reduces the multicollinearity substantially.

Adding the interaction term (Lime Content x Soil PH) to model 1 leads to the regression relation given in equation 2.

Equation 2

We shall call this model as *Model 2*. Model 2 has an R^2_{adj} of approximately 30 %. Some of the diagnostic plots for Model 2 are shown in Fig. 4.

The residuals Vs Fitted plot in Fig. 4 shows that the variance of error terms is constant. A Levene test on the residuals gives us a p-value of 0.9755. Figure 4 and the Levene test show that the equal-variance assumption seems reasonable for these data.

A normal q-q plot for the residuals is shown in Fig. 4. The Anderson Darling (AD) normality test on the residuals gives us a p-value of 0.8992. Figure 4 and the AD normality test show that the normality assumption for the residuals is reasonable for model 1.

The variance inflation factors for the predictors of Model 2 are given in table 2, which shows that there is no significant multicollinearity among the predictors.

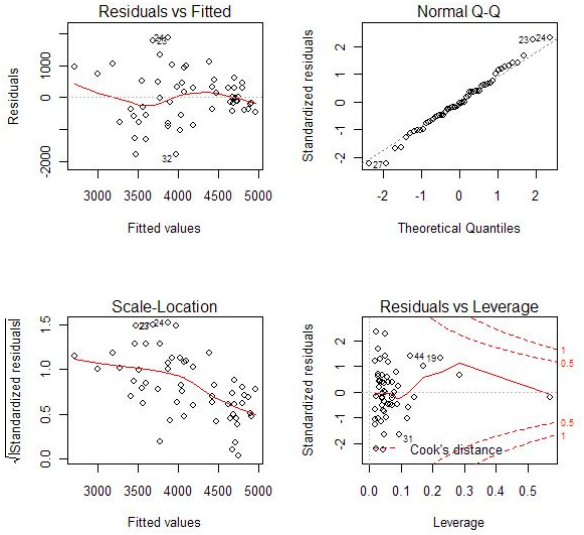

Figure 4. Diagnostic Plots for Model 1

Table 3 .Variance Inflation Factors for Model 2

	Lime Content*	Soil PH*	Lime Content* x Soil PH*
VIF	1.122	1.116	1.198

The times series plot and ACF plot for the residuals of model 2 are shown in Fig. 5.

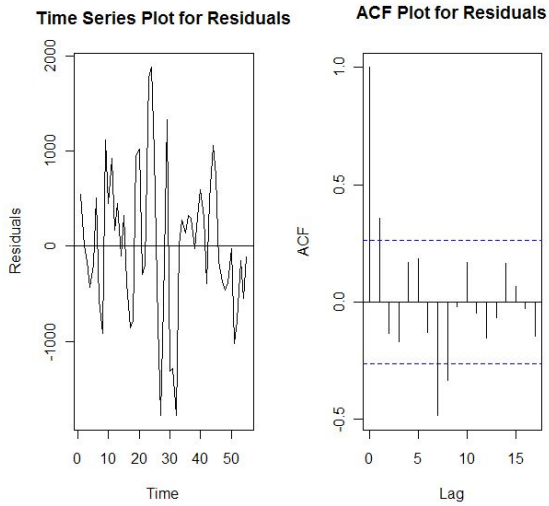

Figure 5 TS and ACF Plot for Residuals of Model 2

Fig. 5 shows that there is still autocorrelation among the residuals. Although we are concerned with accurate predictions rather than inferences, we further looked into the data for some extra information which can solve the problem of auto-correlation and give us better predictions. The data

points have been collected from seven cities of Khyber Pakhtunkhwa province. Furthermore, six indicator variables were included to get more insight of the effect of cities on the regression relation. Significant improvement was observed just by using one indicator variable. The indicator variable is set to 1 if the data point was collected from the city Kohat, otherwise it is set to 0. We call this indicator variable as Kohat, which is defined as follows

The new regression relation by adding the indicator variable *Kohat* is given in equation 3.

Equation 3:

Model 3 has an R^2_{adj} of approximately 45 %. Thus, a significant improvement in the predictive power of our model was observedl. Some of the diagnostic plots for Model 3 are shown in Fig. 6.

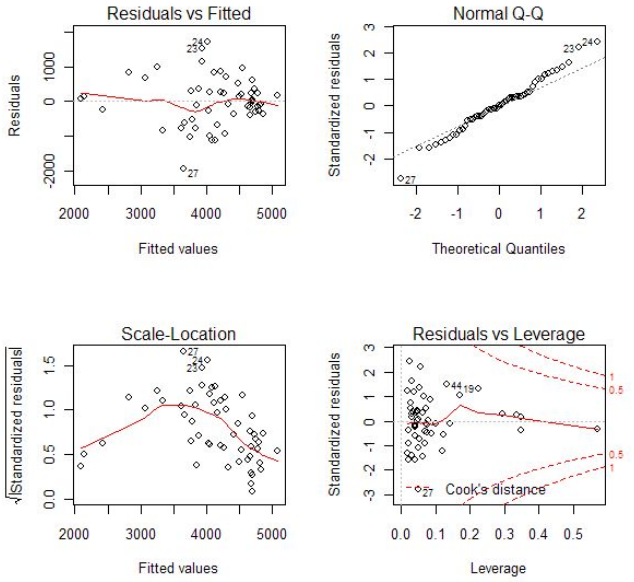

Figure 6. Diagnostic Plots for Model 3

The residuals Vs Fitted plot in Fig. 6 shows that the variance of error terms is constant. A Levene test on the residuals gives us a p-value of 0.9955. Figure 6 and the Levene test show that the equal-variance assumption seems reasonable for the residuals.

A normal q-q plot for the residuals is shown in figure 6. The Anderson Darling (AD) normality test on the residuals gives us a p-value of 0.623. Figure 6 and the AD normality test show that the normality assumption for the residuals is reasonable for model 1.

The variance Inflation factors for the predictors in Model 3 are given in table 4, which shows that there is no significant multicollinearity among the predictors.

Table 4 .Variance Inflation Factors for Model 3

	Lime Content*	Soil PH*	Kohat	Lime Content* x Soil PH*
VIF	1.217	1.120	1.092	1.198

The Time Series Plot and ACF plot for the residuals of model 3 are shown in Fig. 7.

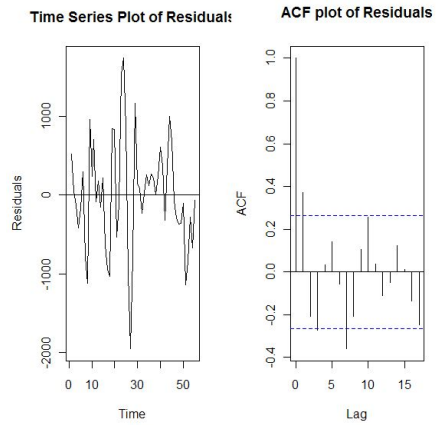

Figure 7 TS and ACF plots for Model 3

Figure 7 shows that adding the indicator variable *Kohat* did not help in terms of resolving the autocorrelation problem.

To compare Model 1 with Model 3 and to validate the models, the data set is divided into a training data set and a validation data set. Sixty seven percent of the observations were randomly selected as the training data set and rest as the validation data set. The results are shown in table 5.

Table 5. Comparison of Model 1 and Model 3

Statistic	Model 1 Training Data Set	Model 1 Validation Data Set	Model 3 Training Data Set	Model 3 Validation Data Set
b_0	9511.39	9988.58	4246.96	4361.10
$s\{b_0\}$	2030.61	5326.55	121.30	252.73
b_1	-67.54	-65.26	-64.89	-49.40
$s\{b_1\}$	23.50	42.00	21.54	47.93
b_2	-590.17	-643.63	-774.47	-608.34
$s\{b_2\}$	268.18	692.74	245.96	663.62
b_3	--	--	-1669.57	-1942.66
$s\{b_3\}$	--	--	535.42	1020.71
b_4	--	--	-102.22	-287.72
$s\{b_4\}$	--	--	49.76	279.19
R^2_{adj}	0.2714	0.1025	0.4513	0.2467
RMSE	831.5	943.6	721.6	864.5
*MSPR	471009.3	--	359227.4	--

*The Mean Squared prediction error which is defined as follows. Where: Y_i is the value of the response variable in the *i*th validation case is the predicted value for the *i*th validation case based on the model-building data set n^* is the number of cases in the validation data set.

Model 3 gives us much better prediction power than the model put forward by Bhatti et al. (1998c). The results reveal that there is still some important predictor missing in the model. Inclusion of rain fall data in the future may result in getting better results.

Classification and Regression Trees: This section explores the classification and regression trees (CART). Classification and regression trees were used to classify the Yield into three categories (High, Medium, and Low) based on the following:

High: Yield > 4000

Medium: 3000 < Yield ≤ 4000

Low: Yield ≤ 3000

The training and validation data sets of section 2 are used to create the classification trees. A classification tree using the training data set is shown in figure 8. We shall call this model as *Tree 1*.

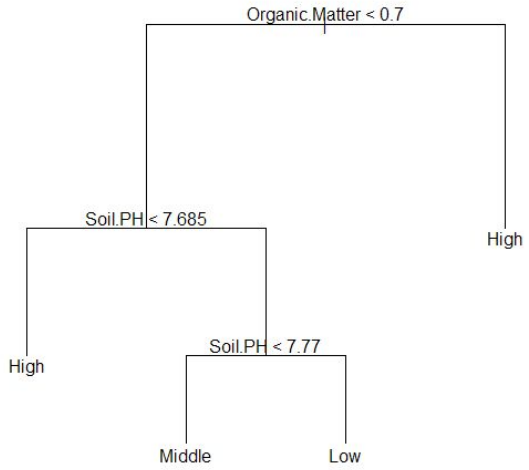

Figure 8 Classification Tree – Tree 1

The confusion matrix for *Tree 1* is given in table 6.

Table 6. Confusion Matrix for Tree 1 on Training data set

		Predicted			
		Low	Medium	High	Error Rate
True	Low	3	0	1	25%
	Medium	2	5	0	28.5%
	High	2	0	23	8.0%

Table 7. Confusion Matrix for Tree 1 on Validation data set

		Predicted			
		Low	Medium	High	Error Rate
True	Low	1	3	1	80.0%
	Medium	3	0	0	100.0%
	High	2	1	8	27.3%

Table 6 shows that the three classes have unbalanced miss-classification error rates. After applying the model on the validation data set, very high miss-classification error rates for "Low" and "Medium" classes were observed as shown in Table 7. To give importance to the Minority classes i.e., "Low" and "Medium", a weighted classification tree was created with the following weights

High = 1, Medium = 2, Low = 2

A Weighted classification tree (Tree 2) using the training data set is shown in figure 9.

Table 8. Confusion Matrix for Tree 2 on Training data set

		Predicted			
		Low	Medium	High	Error Rate
True	Low	4	0	0	0.0%
	Medium	1	5	1	28.5%
	High	3	1	21	16.0%

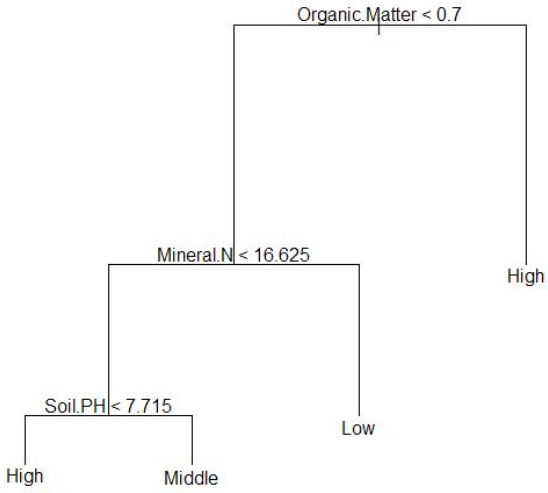

Figure 9. Weighted Classification Tree – Tree 2

The Confusion matrices for Tree 2 are given in table 8 and 9. These tables show a significant improvement over *Tree 1*.

Table 9. Confusion Matrix for Tree 2 on Validation data set

		Predicted			
		Low	Medium	High	Error Rate
True	Low	2	2	1	60.0%
	Medium	1	2	0	33.3%
	High	4	2	5	54.5%

These results show that a weighted CART is doing a better job. The Weighted CART using the complete data set is shown in figure 10.

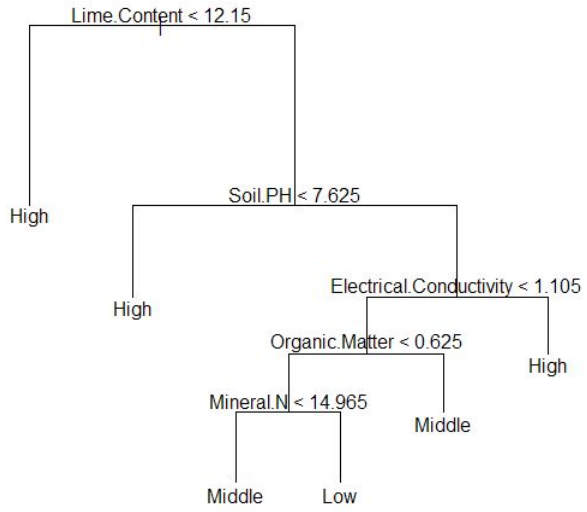

Figure 10

For Tree 3, a plot of the fitted Vs Actual is shown in figure 11. The Standard error for this regression tree is 535.43. It was observed that the regression tree is doing a poor job in predicting the Yield.

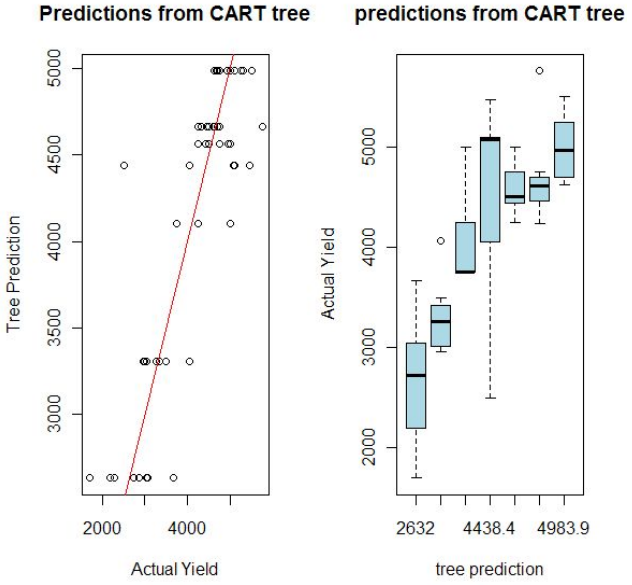

Figure 11. Fitted Vs Actual for Tree 3

We can conclude that classification and Regression trees are not doing a good job.

Random Forests: In this section it was examined how Random Forests perform in predicting/classifying our response variable – Yield.

Plain Random Forests for classification gave us greatly unbalanced miss-classification error rates (results not shown), for this reason, weighted Random Forests (*WRF*) was developed. We used 500 trees and 3 variables are sampled at each split of the tree. The confusion matrix for *WRF* is given in table 10. The OOB error rate for *WRF* is 5.45 %.

Table 10 .Confusion Matrix for WRF

		Predicted			
		Low	Medium	High	Error Rate
True	Low	7	2	0	22.2%
	Medium	1	9	0	10.0%
	High	0	0	36	0.0%

A Random Forest for Regression was developed and termed as Reg_RF. This model uses 200 trees and 2 variables are sampled at each split of the trees. The plot for Fitted Vs Actual for Reg_RF is shown in Figure 12.

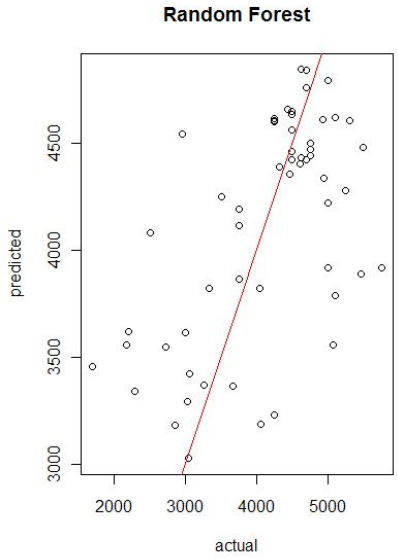

Figure 12. Fitted Vs Actual for Reg_RF

The standard error for Reg_RF is 780.41. Figure 11 shows that Reg_RF is doing a poor job in terms of prediction.

We can conclude from this section that using Random Forests for classification is giving us good results; however using Random Forests for regression is giving us poor results.

Comparison of the three Methods: This section compares the models created in sections 2, 3, and 4 using Multiple Regression, CART and Random Forests respectively. Initial comparison of the models was based on prediction power, and then based on classification.

Model 3 created in section 2 was the best model in terms of prediction using multiple regression, we compare model 3 with the regression models created in section 3, and 4 using CART and Random Forests. The plots for the Fitted Vs Actual for these three models are shown in Fig. 13.

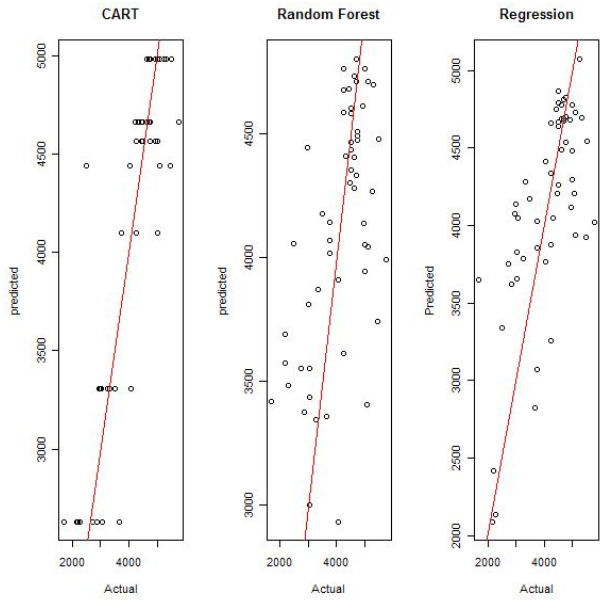

Figure 13. Comparison of Three Regression models

Figure 13 shows us that Multiple Regression is doing a better job than CART and Random Forests in terms of predictions. The standard errors for the three methods are given in table 11.

Table 11.Standard Errors of the Three Regression Models

	Multiple Regression (Reg 3)	CART (Tree 3)	Random Forests (Reg_RF)
RMSE	725.7	535.44	794.10

Models created in section 5 and 6 based on CART and Random Forests to classify Yield into the three classes are compared in table 12.. The miss-classification error rates using for the two models are given in table 12.Table 12. Comparison of CART and Random Forests

Table 12.

	Miss-Classification Error Rates	
	CART (Tree 2)	**Random Forests (WRF)**
Low	22.2%	22.2%
Medium	100.0%	10.0%
High	8.3%	0.0%

Table 12 shows that Random Forests give us better results in terms of classification.

For efficient nutrient management , knowledge of yield potential of wheat at a particular site is very important. Mullen et al. (2003) suggested from their results that recognizing yield potential of crop is very important for obtaining a response to N fertilization. Similarly, Fowler (2003) observed that for high yield potential , N fertilizer rate was very high. In the present study, various regression models were compared with Forest Random model for determining yield potential of wheat for different sites using soil data. The Random Forest model was found better which can be used for this purpose. Yield potential of wheat has been very successfully used for fertilizer management of wheat in a spatially variable large field (Mulla *et al.*, 1992; Bhatti et al., 1998a) as well as for formulating site- specific N rate for a particular site (Bhatti et al., 1998c). Our next step will be to develop a computer program for determining N rate for a particular site using soil data and potential yield.

Conclusion: Based on the poor predictions by Regression and the preceding discussion/results we conclude that using Random Forests to classify Yield into three categories gives us very small misclassification error rates and the best results.

REFERENCES

Bakhsh, A., A.H.Gurmani, A.U.Bhatti, and H.Rehman. 1994. Effect of various combination of N, P and K on the yield of wheat in Rod-kohi areas of D.I.Khan Division. Pak. J. Soil Sci. 9(1-2): 43-47.

Bhatti, A.U., M. Afzal and Farmanullah. 1997. Effect of slope position on soil properties and wheat yield. J. Engg. & Applied Sci. 16(2):45-50.

Bhatti, A.U., F. Hussain, Farmanullah, and M.J. Khan. 1998a. Use of spatial patterns of soil properties and wheat yields in geostatistics for determination of fertilizer rates. Comm. Soil Sci. Plant Anal. 29:509-522.

Bhatti, A.U., M. Khan, K.S. Khurshid and Farmanullah. 1998b. Site specific determination of N rates for rainfed wheat using available soil moisture. Pak.J. Arid Agri. 1(1):11-18.

Bhatti, A.U., M.Khan, K.S. Khurshid and Farmanullah. 1998c. Site specific determination of N rates for irrigated wheat based on soil properties. In: Proceedings of Symposium on Nutrient Management for Sustainable Agricultural Growth at NFDC Islamabad, Dec. 8-9,1997:131-137.

Bhatti, A.U., and D.J. Mulla. 1992. Modelling relationship between soil properties and wheat yields in dryland areas. Sarhad J. Agri. 8(5): 579-586.

Burleigh, J.R., C.F. yamoah, J.L. Regas, and Val. J. Eylands. 1991. Analysis of factors related to wheat yields on fields in the Buberuka lands of Rawanda. Agron. J. 83: 625-630.

Fowler, D.B. 2003. Crop nitrogen demand and grain protein concentration of wheat. Agron. J. 95:260-265.

Khan, B.R. and G. Akbar. Sailaba (Rod-Kohi). Agriculture in Balochistan. Rod-Kohi Agriculture, Problems and Prospects Symposium, Agri. Res. Institute D.I. Khan Nov. 27-29. P.85-94.

Kunter, M.H., C.J. Nechtsheim, and J. Neter. 2004. Applied Linear Regression Models. 4th Ed. McGraw-Hill/Irwin New York N.Y.

Legget, G.E. 1959. Relationship between yield, available moisture, and available nitrogen in Eastern Washington dryland areas. Washington Agricultural Experiment Station Bull. 609. Washington State University, Pullman, Wa, U.S.A.

Makowski, D., D. Wallach, and J.M. Meynard. 2001. Statistical methods for predicting responses to applied nitrogen nd calculating optimal nitrogen rates. Agron. J. 93:531-539.

Mombela, F.A., J.J. Nicholaides III, and L.A. Nelson. 1981. A method determine the appropriate mathematical form for incorporating soil test levels for recommendation responses. Agron. J. 73:937-941.

Mulla, D.J., A.U. Bhatti, M.W. Hammond, and J.A. Benson. A comparison of winter wheat yield and quality under uniform versus spatially variable fertilizer management. Agri. Ecosystem and Environ. 38:301-311.

Mullen, R.W., K.W. Freeman, W.R. Raun, G.V. Johnson, M.L. Stone, and J.B. Solic. 2003. Identifying an in-season response index and the potential to increase wheat yield with nitrogen. Agron. J. 95:347-351.

Nelson, L. A. 1987. Role of response surfaces in soil test calibration. P. 31-40. In: J. R. Brown (ed.). Soil testing: Sampling, correlation and interpretation. SSSA, Madison, W.I.

Nelson, L. A., R.D. Voss, and J. Pesek. 1985. Agronomic and statistical evaluation of fertilizer response. P.53-90. In: O.P. Englestad (ed.). Fertilizer Technology and Use. 3rd ed. SSSA, Madison, W.I.

Pawson, W.W., O.L. Bough Jr., J.P. Swanson, and G.M. Harner. 1961. Economics of cropping systems and soil conservation in the Palouse. Agri. Exp. Stn. Of Idaho, Oregon, and Washington, and A.R.S. U.S.A. Bull 2.

Sain, G.E. and M.A. Jaurgui. 1993. Deriving fertilizer recommendations with a flexible functional form. Agron. J. 85:934-937.