LINEAR REGRESSION ANALYSIS

YIELD AS A FUNCTION OF N_PI & PI_N_UPTAKE

JUNE 11, 2022

JAYA BIJESH
COSC593

Table of Contents

Introduction

The purpose of this document is to model rice yield (QuadYield) using Nitrogen rates applied at Panicle Initiation (N_PI) and Nitrogen Uptake at Panicle Initiation (PI_N_Uptake) as predictors. The statistical modelling method used is Multiple Linear Regression with QuadYield as response variable and N_PI and PI_N_Uptake as predictors.

Method

The Rice data set has 1751 observations across 6 rice growing seasons namely 2015_16, 2016_17, 2017_18, 2018_19, 2019_20 and 2020_21. For this analysis, the validation set approach has been applied on the Rice data set by holding out observations corresponding to a particular season as test data, while data points corresponding to all the other seasons were used to train the model. This analysis has been completed with each season as test data.

Three regression models have been used in this analysis as follows:

Model 1: Drill sown Reiziq

Model 2: All drill sown varieties

Model 3: All varieties across all sowing methods

The predicted vs actual values of yield have been plotted and model accuracies have been assessed for each of the models using Mean Absolute Error (MAE), Mean Squared error (MSE), Root Mean Square Error (RMSE) and R^2.

Regression Analysis

Multiple regression modelling has been done using N_PI as both a categorical variable with levels 0, 60, 90 and 120 and as a numerical variable. For each season as test data set, the abovesaid three models were run for N_PI as categorical and numeric.

N_PI as categorical predictor

The general form of the regression equation using N_PI as categorical variable is shown below:

$$QuadYield = \beta_0 + \beta_1\ PI\_N\_Uptake + \beta_2\ 60N\_PI + \beta_3\ 90N\_PI + \beta_4\ 120N\_PI + \epsilon$$

N_PI = 0 is the base level. With N_PI as a categorical predictor, regression analysis was done using validation set approach as below:

Test data set: 2015_16

The training data is constituted by the observations from all seasons except 2015_16.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 1):

Table 1: Model 1- test data 2015_16

	Coefficients	95% CI	p-value
Intercept	8.248	(7.310, 9.186)	<2e-16
PI_N_Uptake	0.022	(0.013, 0.031)	1.5e-06
N_PI60	1.024	(0.176, 1.872)	0.018
N_PI90	3.322	(0.717, 5.928)	0.013

The adjusted R^2 value of the model is 0.188, therefore only approximately 19% of the variability is explained by the model. Hence the model is not useful. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

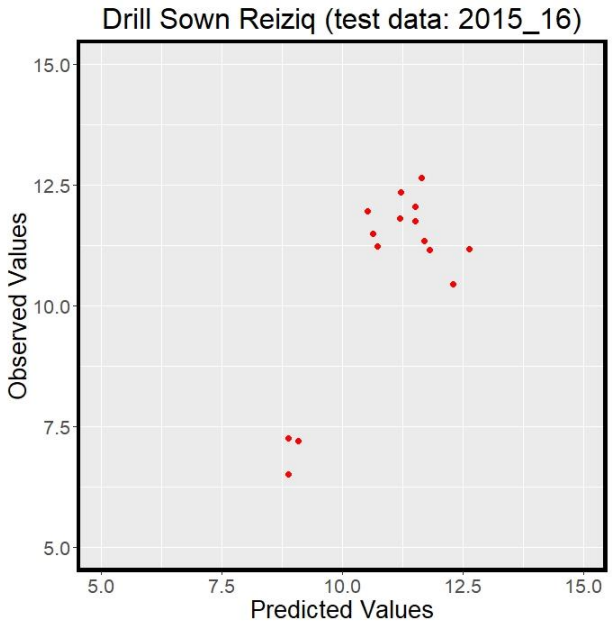

Figure 1: Model 1- test data 2015_16

The model accuracy was assessed as shown in Table 2 below:

Table 2: Model accuracy - test data 2015_16

Evaluation Metric	Value
MAE	1.1
MSE	1.6
RMSE	1.3

R^2	0.59

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 3):

Table 3: Model 2 - test data 2015_16

	Coefficients	95% CI	p-value
Intercept	7.240	(6.890, 7.591)	<2e-16
PI_N_Uptake	0.036	(0.033 0.039)	<2e-16
N_PI60	1.270	(0.943 1.597)	6.4e-14
N_PI90	3.530	(2.353 4.707)	5.7e-09

The adjusted R^2 value of the model is 0.405, therefore only approximately 41% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

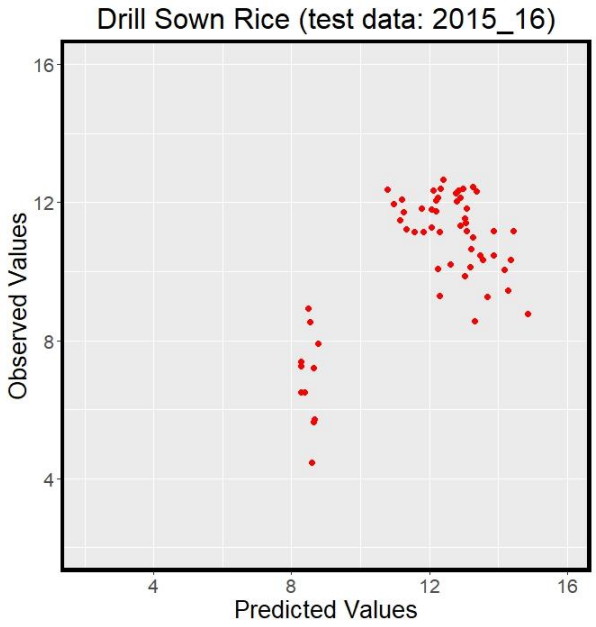

Figure 2: Model 2- test data 2015_16

The model accuracy was assessed as follows (Table 4):

Table 4: Model accuracy - test data 2015_16

Evaluation Metric	Value
MAE	1.7
MSE	5.2

RMSE	2.3
R^2	-0.22

The negative value of R2 indicates that the prediction is likely to be less accurate than the mean value of the data set over time.

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 5):

Table 5: Model 3 - test data 2015_16

	Coefficients	95% CI	p-value
Intercept	6.135	(5.854 6.416)	<2e-16
PI_N_Uptake	0.041	(0.038 0.044)	<2e-16
N_PI60	1.442	(1.172 1.712)	<2e-16
N_PI90	3.040	(2.172 3.908)	9.7e-12

The adjusted R^2 value of the model is 0.448, therefore approximately 45% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

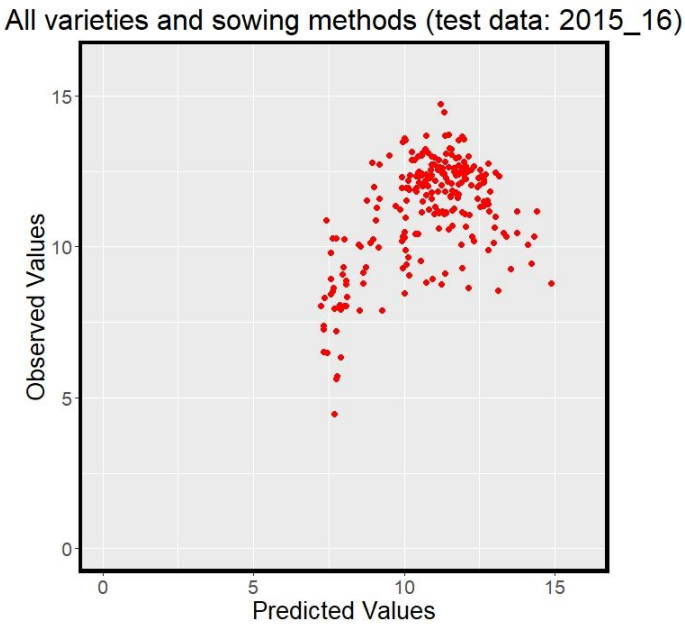

Figure 3: Model 3- test data 2015_16

The model accuracy was assessed as follows (Table 6):

Table 6: Model accuracy - test data 2015_16

Evaluation Metric	Value
MAE	1.4
MSE	3
RMSE	1.7
R^2	0.13

Summary

- Model 2 has the lowest RMSE, but with a negative R^2 value
- The highest R^2 is for model 1 (59%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2016_17

The training data is constituted by the observations from all seasons except 2016_17.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 7):

Table 7: Model 1 - test data 2016_17

	Coefficients	95% CI	p-value
Intercept	8.094	(7.125 9.06)	<2e-16
PI_N_Uptake	0.022	(0.013, 0.03)	1.6e-06
N_PI120	1.895	(-0.721 4.51)	0.154
N_PI60	0.963	(0.039 1.89)	0.041
N_PI90	3.501	(0.888 6.11)	0.009

The p-value of N_PI120 is greater than the threshold of 0.05 and the confidence interval contains 0, therefore N_PI120 is not significant. The adjusted R^2 value of the model is 0.199, therefore only approximately 20% of the variability is explained by the model. Hence the model is not especially useful. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

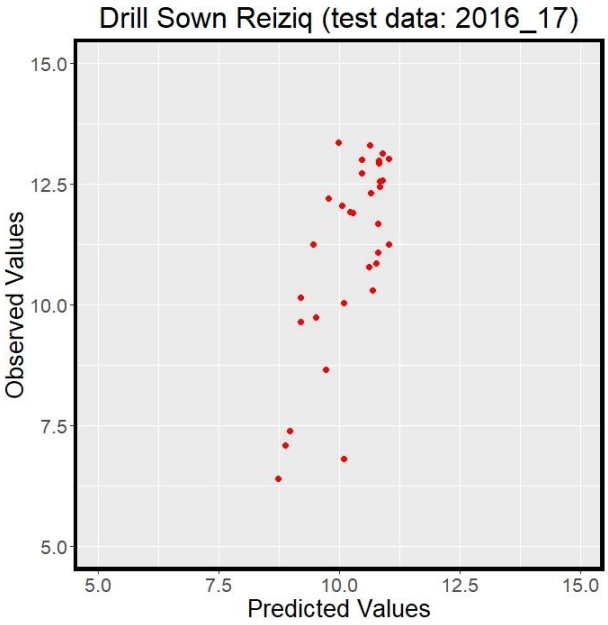

Figure 4: Model 1- test data 2016_17

The model accuracy was assessed as follows (Table 8):

Table 8: Model 1 accuracy- test data 2016_17

Evaluation Metric	Value
MAE	1.5
MSE	3.2
RMSE	1.8
R^2	0.18

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 9):

Table 9: Model 2 - test data 2016_17

	Coefficients	95% CI	p-value
Intercept	7.638	(7.240 8.036)	<2e-16
PI_N_Uptake	0.031	(0.028 0.034)	<2e-16
N_PI120	1.342	(0.106 2.577)	0.033
N_PI60	1.241	(0.869 1.613)	1.1e-10
N_PI90	3.678	(2.444 4.913)	7.4e-09

The adjusted R^2 value of the model is 0.342, therefore only approximately 34% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

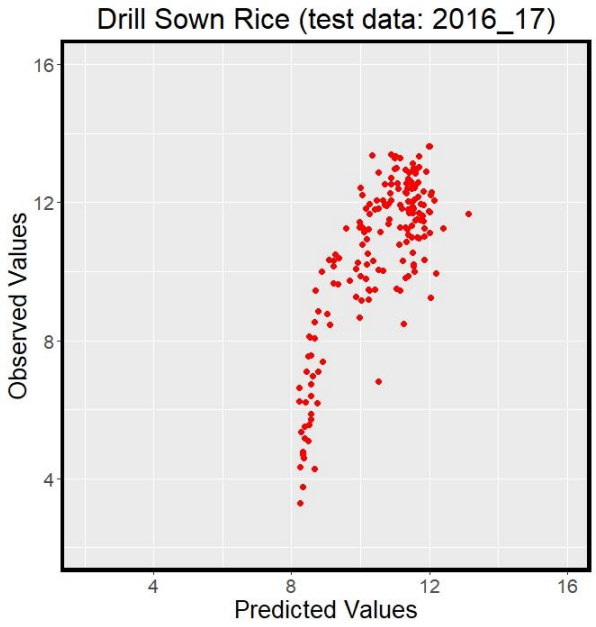

Figure 5: Model 2- test data 2016_17

The model accuracy was assessed as follows (Table 10):

Table 10: Model 2 - test data 2016_17

Evaluation Metric	Value
MAE	1.2
MSE	2.5
RMSE	1.6
R^2	0.55

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 11):

Table 11: Model 3 - test data 2016_17

	Coefficients	95% CI	p-value
Intercept	6.652	(6.363 6.941)	<2e-16
PI_N_Uptake	0.037	(0.034 0.039)	<2e-16
N_PI120	2.442	(1.818 3.066)	3.2e-14
N_PI60	1.476	(1.209 1.744)	<2e-16
N_PI90	3.062	(2.197 3.927)	5.8e-12

The adjusted R^2 value of the model is 0.407, therefore approximately 41% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

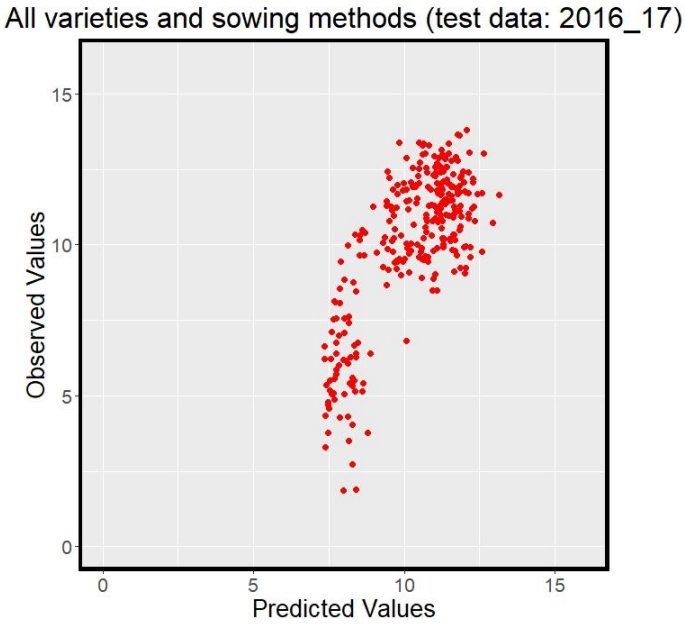

Figure 6: Model 3- test data 2016_17

The model accuracy was assessed as follows in Table 12:

Table 12: Model 3 - test data 2016_17

Evaluation Metric	Value
MAE	1.3
MSE	3
RMSE	1.7
R^2	0.68

The R2 value is 0.68, therefore 68% of the variability in yield can be explained by the predictors in this model.

Summary

- The lowest RMSE of 1.6 is for model 2.
- The highest R^2 is for model 3 (68%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2017_18

The training data is constituted by the observations from all seasons except 2017_18.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 13):

Table 13: Model 1 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	8.441	(7.510 9.372)	<2e-16
PI_N_Uptake	0.019	(0.011 0.028)	1.1e-05
N_PI120	1.741	(-0.842 4.324)	0.1848
N_PI60	1.011	(0.139 1.883)	0.0234
N_PI90	3.435	(0.853 6.016)	0.0095

The p-value of N_PI120 is greater than the threshold of 0.05 and the confidence interval contains 0, therefore N_PI120 is not significant. The adjusted R^2 value of the model is 0.169, therefore only approximately 17% of the variability is explained by the model. Hence the model is not particularly useful. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

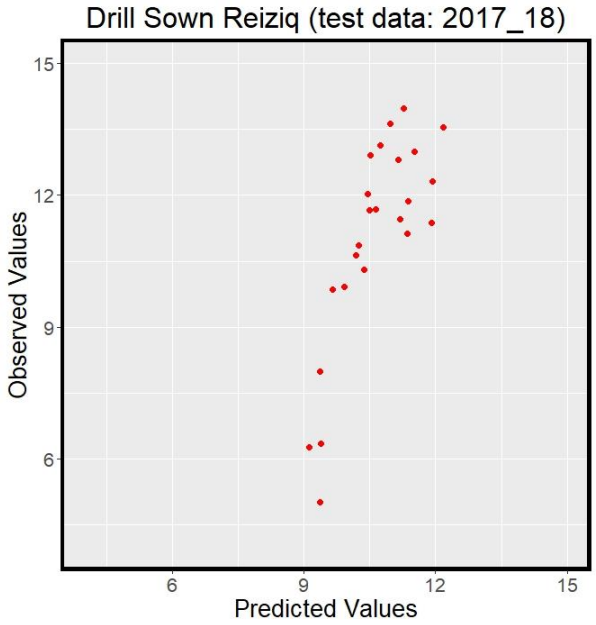

Figure 7: Model 1- test data 2017_18

The model accuracy was assessed as follows in Table 14:

Table 14: Model 1 - test data 2017_18

Evaluation Metric	Value
MAE	1.4
MSE	3.2
RMSE	1.8
R^2	0.42

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 15):

Table 15: Model 2 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	7.472	(7.10 7.840)	<2e-16
PI_N_Uptake	0.033	(0.03 0.036)	<2e-16
N_PI120	1.350	(0.13 2.567)	0.03
N_PI60	1.350	(0.99 1.708)	3.6e-13
N_PI90	3.631	(2.41 4.848)	6.9e-09

The adjusted R^2 value of the model is 0.379, therefore only approximately 38% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

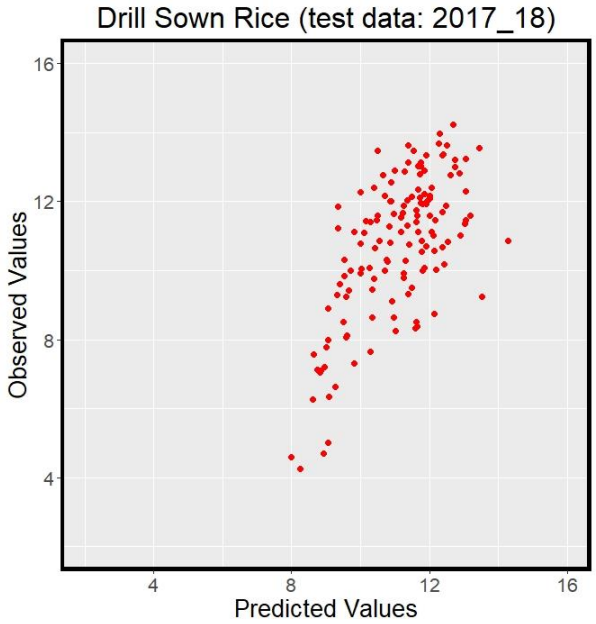

Figure 8: Model 2- test data 2017_18

The model accuracy was assessed as follows (Table 16):

Table 16: Model 2 - test data 2017_18

Evaluation Metric	Value
MAE	1.3
MSE	2.6
RMSE	1.6
R^2	0.4

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 17):

Table 17: Model 3 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	6.456	(6.167 6.746)	<2e-16
PI_N_Uptake	0.038	(0.036 0.041)	<2e-16
N_PI120	2.491	(1.871 3.111)	6.3e-15
N_PI60	1.497	(1.232 1.761)	<2e-16
N_PI90	3.035	(2.176 3.894)	6.3e-12

The adjusted R^2 value of the model is 0.424, therefore approximately 42% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

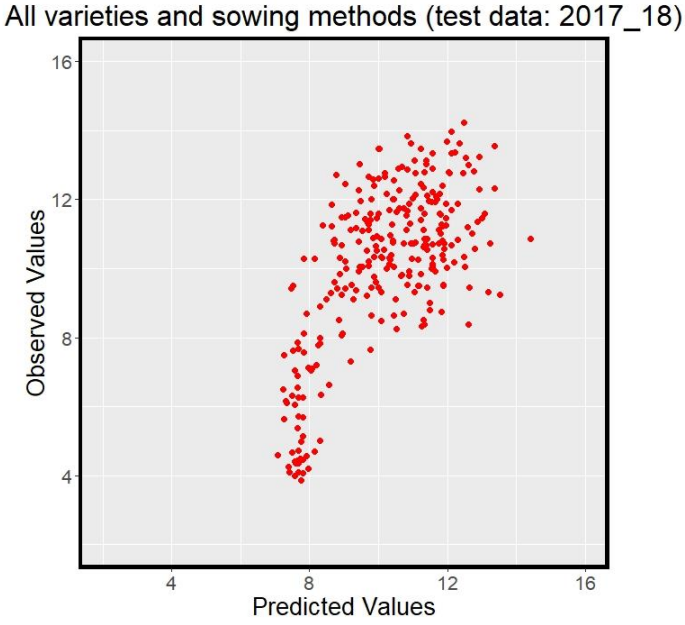

Figure 9: Model 3- test data 2017_18

The model accuracy was assessed as follows (Table 18):

Table 18: Model 3 - test data 2017_18

Evaluation Metric	Value
MAE	1.5
MSE	3.2
RMSE	1.8
R^2	0.45

The R2 value is 0.68, therefore 68% of the variability in yield can be explained by the predictors in this model.

Summary

- The lowest RMSE of 1.6 is for model 2.
- The highest R^2 is for model 3 (45%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2018_19

The training data is constituted by the observations from all seasons except 2018_19.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained as in Table 19 below:

Table 19: Model 1 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	8.321	(7.356 9.29)	<2e-16
PI_N_Uptake	0.021	(0.012 0.03)	3.8e-06
N_PI120	1.729	(-0.903 4.36)	0.196
N_PI60	0.953	(0.063 1.84)	0.036
N_PI90	3.362	(0.733 5.99)	0.013

The p-value of N_PI120 is greater than the threshold of 0.05 and the confidence interval contains 0, therefore N_PI120 is not significant. The adjusted R^2 value of the model is 0.176, therefore only approximately 18% of the variability is explained by the model. Hence the model is not especially useful. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

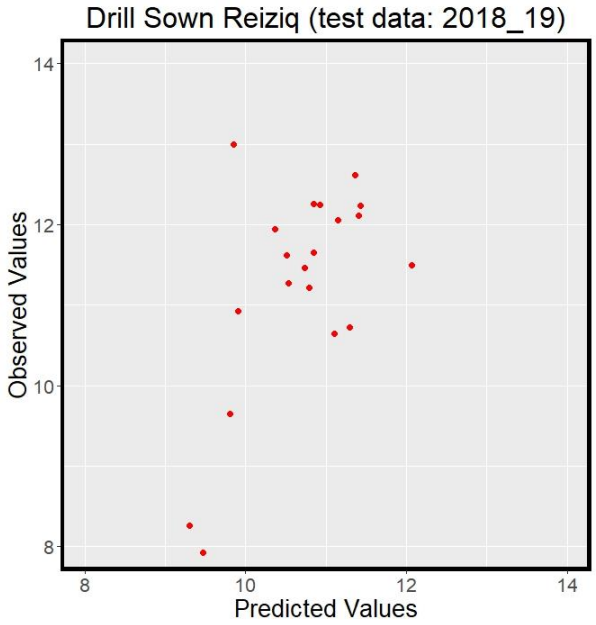

Figure 10: Model 1- test data 2018_19

The model accuracy was assessed as follows (Table 20):

Table 20: Model 1 - test data 2018_19

Evaluation Metric	Value
MAE	1.2
MSE	2
RMSE	1.4
R^2	0.49

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 21):

Table 21: Model 2 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	7.595	(7.214 7.975)	<2e-16
PI_N_Uptake	0.032	(0.028 0.035)	<2e-16
N_PI120	1.322	(0.093 2.551)	0.035
N_PI60	1.201	(0.839 1.563)	1.3e-10
N_PI90	3.637	(2.408 4.866)	9.1e-09

The adjusted R^2 value of the model is 0.351, therefore only approximately 35% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

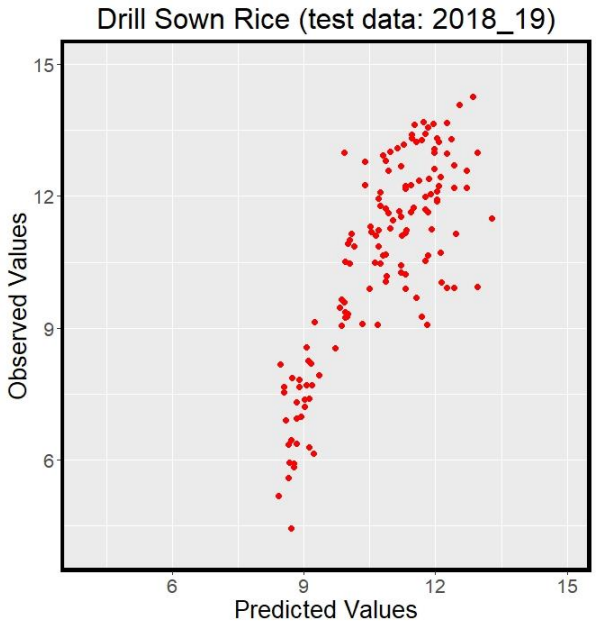

Figure 11: Model 2- test data 2018_19

The model accuracy was assessed as follows (Table 22):

Table 22: Model 2 - test data 2018_19

Evaluation Metric	Value
MAE	1.2
MSE	2.1
RMSE	1.5
R^2	0.59

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 23):

Table 23: Model 3 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	6.942	(6.668 7.216)	<2e-16
PI_N_Uptake	0.036	(0.033 0.038)	<2e-16
N_PI120	2.246	(1.659 2.833)	1.1e-13
N_PI60	1.489	(1.238 1.739)	<2e-16
N_PI90	2.915	(2.101 3.728)	3.3e-12

The adjusted R^2 value of the model is 0.417, therefore approximately 42% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

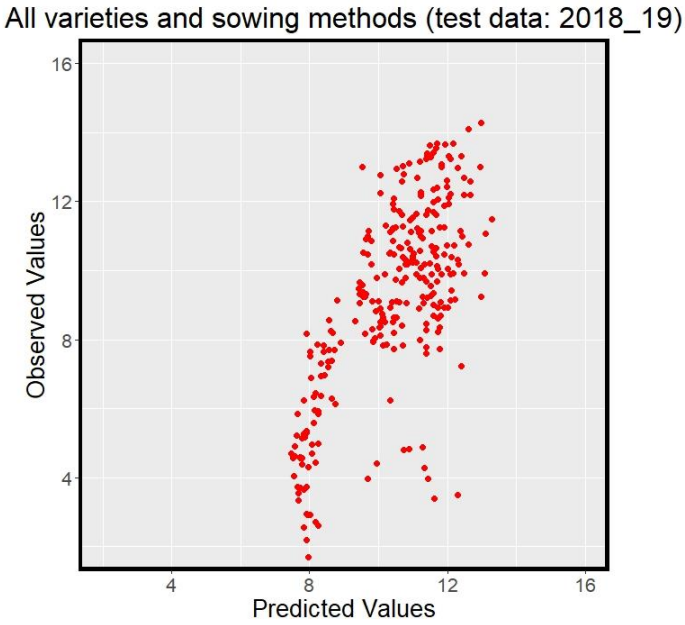

Figure 12: Model 3- test data 2018_19

The model accuracy was assessed as follows (Table 24):

Table 24: Model 3 - test data 2018_19

Evaluation Metric	Value
MAE	1.8
MSE	5.8
RMSE	2.4
R^2	0.27

The R2 value is 0.27, therefore 27% of the variability in yield can be explained by the predictors in this model.

Summary

- The lowest RMSE of 1.4 is for model 1.
- The highest R^2 is for model 2 (59%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2019_20

The training data is constituted by the observations from all seasons except 2019_20.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained as in Table 25 below:

Table 24: Model 1 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	8.385	(7.398 9.373)	<2e-16
PI_N_Uptake	0.018	(0.009 0.027)	0.00015
N_PI120	1.880	-0.679 4.440)	0.149
N_PI60	1.164	(0. 260 2.068)	0.012
N_PI90	3.612	(1.050 6.174)	0.006

The p-value of N_PI120 is greater than the threshold of 0.05 and the confidence interval contains 0, therefore N_PI120 is not significant. The adjusted R^2 value of the model is 0.164, therefore only approximately 16% of the variability is explained by the model. Hence the model is not particularly useful. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

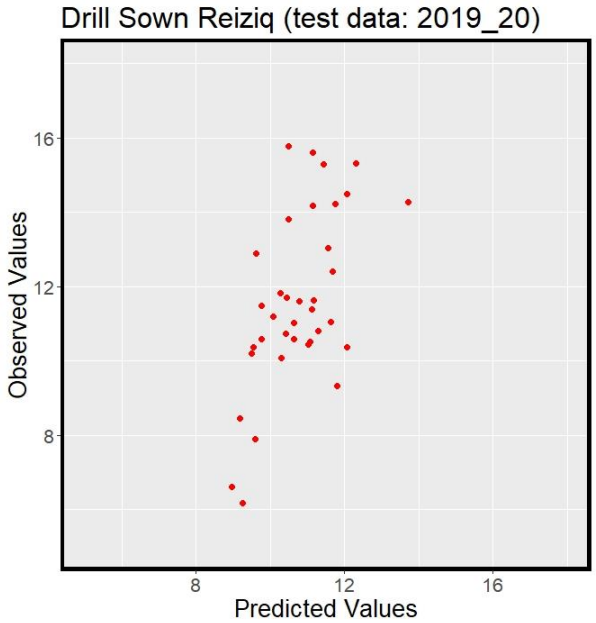

Figure 13: Model 1- test data 2019_20

The model accuracy was assessed as follows (Table 26):

Table 26: Model 1 - test data 2019_20

Evaluation Metric	Value
MAE	1.6
MSE	4.4
RMSE	2.1
R^2	0.2

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 27):

Table 27: Model 2 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	7.45	(7.070 7.825)	<2e-16
PI_N_Uptake	0.03	(0.026 0.033)	<2e-16
N_PI120	1.63	(0.475 2.776)	0.0057
N_PI60	1.34	(0.984 1.697)	4.4e-13
N_PI90	4.00	(2.843 5.147)	2.1e-11

The adjusted R^2 value of the model is 0.346, therefore only approximately 35% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

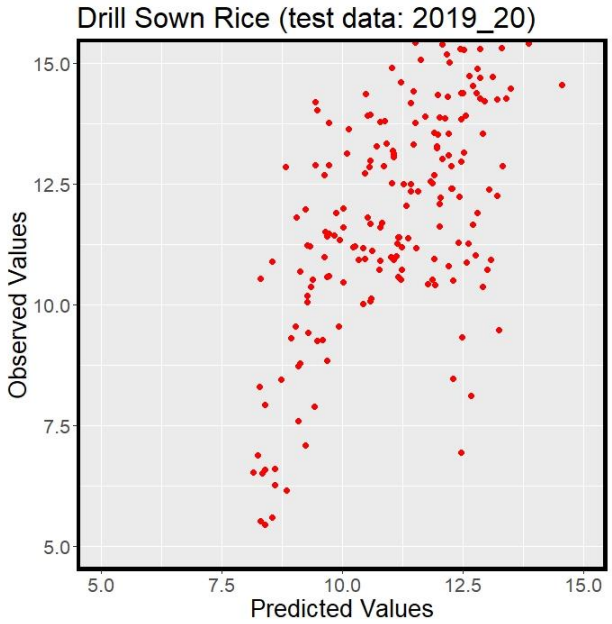

Figure 14: Model 2- test data 2019_20

The model accuracy was assessed as follows (Table 28):

Table 28: Model 2 - test data 2019_20

Evaluation Metric	Value
MAE	1.9
MSE	5.3
RMSE	2.3
R^2	0.23

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 29):

Table 29: Model 3 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	6.408	(6.137 6.679)	<2e-16
PI_N_Uptake	0.037	(0.034 0.039)	<2e-16
N_PI120	2.693	(2.104 3.283)	<2e-16
N_PI60	1.606	(1.360 1.851)	<2e-16
N_PI90	3.317	(2.497 4.137)	4.2e-15

The adjusted R^2 value of the model is 0.418, therefore approximately 42% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

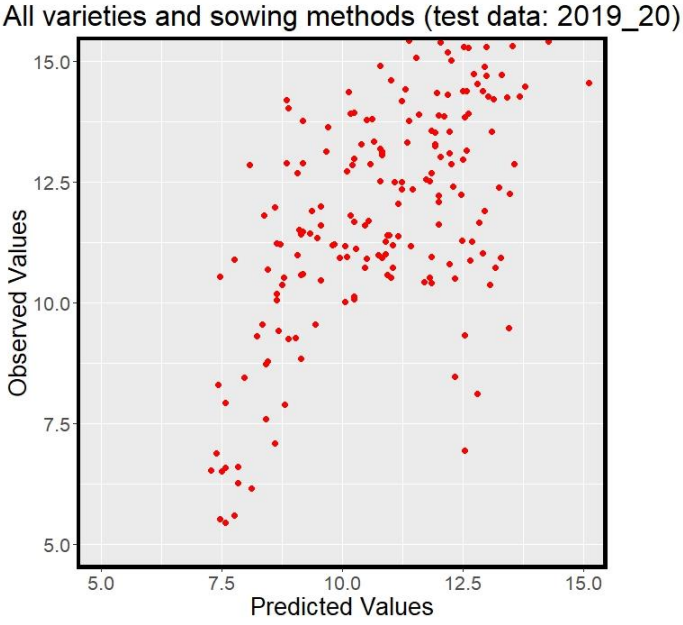

Figure 15: Model 3- test data 2019_20

The model accuracy was assessed as follows in Table 30 below:

Table 30: Model 3 - test data 2019_20

Evaluation Metric	Value
MAE	2
MSE	5.7
RMSE	2.4
R^2	0.16

The R2 value is 0.16, therefore only 16% of the variability in yield can be explained by the predictors in this model.

Summary

- The lowest RMSE of 2.1 is for model 1.
- The highest R^2 is for model 2 (23%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2020_21

The training data is constituted by the observations from all seasons except 2020_21.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 31):

Table 31: Model 1 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	7.323	(6.583 8.06)	<2e-16
PI_N_Uptake	0.033	(0.027 0.04)	<2e-16
N_PI120	1.721	(-0.109 3.55)	0.065
N_PI60	1.531	(0.905 2.16)	3.7e-06

The p-value of N_PI120 is greater than the threshold of 0.05 and the confidence interval contains 0, therefore N_PI120 is not significant. The adjusted R^2 value of the model is 0.474, therefore only approximately 47% of the variability is explained by the model. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

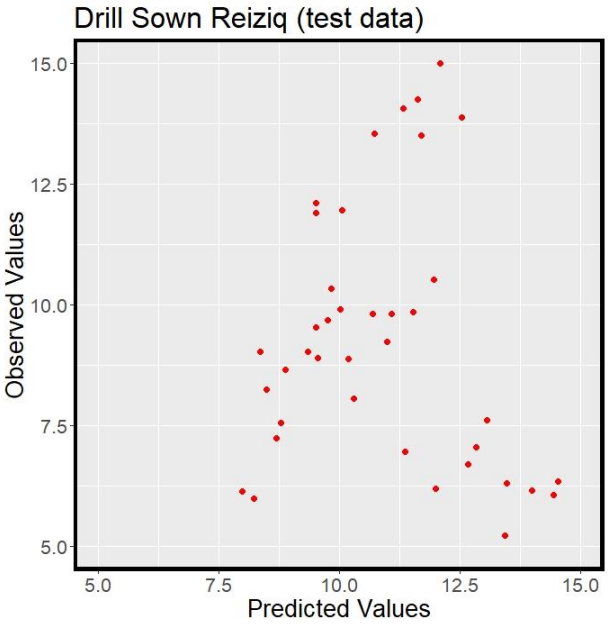

Figure 16: Model 1- test data 2020_21

The model accuracy was assessed as follows (Table 32):

Table 32: Model 1 - test data 2020_21

Evaluation Metric	Value
MAE	2.8
MSE	14
RMSE	3.8
R^2	-0.6

The negative value of R2 indicates that the prediction is likely to be less accurate than the mean value of the data set over time.

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 33):

Table 33: Model 2 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	6.901	(6.567 7.24)	<2e-16
PI_N_Uptake	0.037	(0.034 0.04)	<2e-16
N_PI120	0.543	(0.533 2.55)	0.0028
N_PI60	1.668	(1.372 1.96)	<2e-16

The adjusted R^2 value of the model is 0.498, therefore only approximately 50% of the variability is explained by the model. Hence this model is more useful than model 1. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

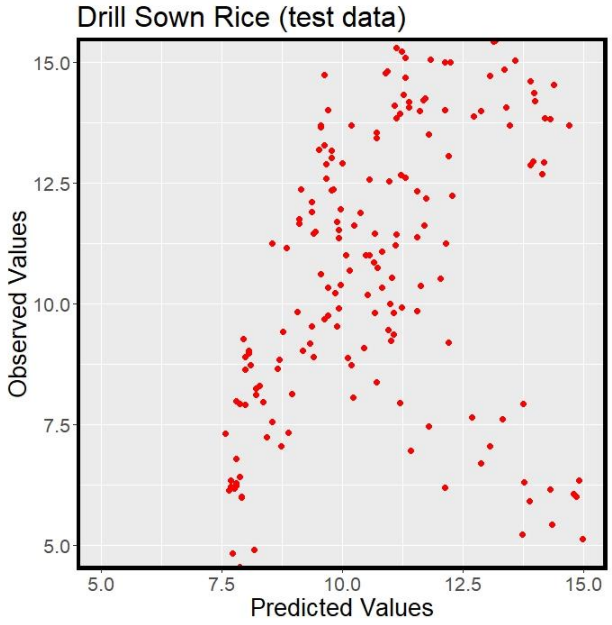

Figure 17: Model 2- test data 2020_21

The model accuracy was assessed as follows (Table 34):

Table 34: Model 2 - test data 2020_21

Evaluation Metric	Value
MAE	2.2
MSE	9.3
RMSE	3.1
R^2	0.079

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 35):

Table 35: Model 3 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	6.079	(5.809 6.350)	<2e-16
PI_N_Uptake	0.041	(0.038 0.043)	<2e-16
N_PI120	2.674	(2.109 3.240)	<2e-16
N_PI60	1.785	(1.551 2.020)	<2e-16

The adjusted R^2 value of the model is 0.481, therefore approximately 48% of the variability is explained by the model. Hence this model tends to be comparably useful to model 2. Predictions were made using the test data and the following scatterplot was generated for observed vs predicted values of yield.

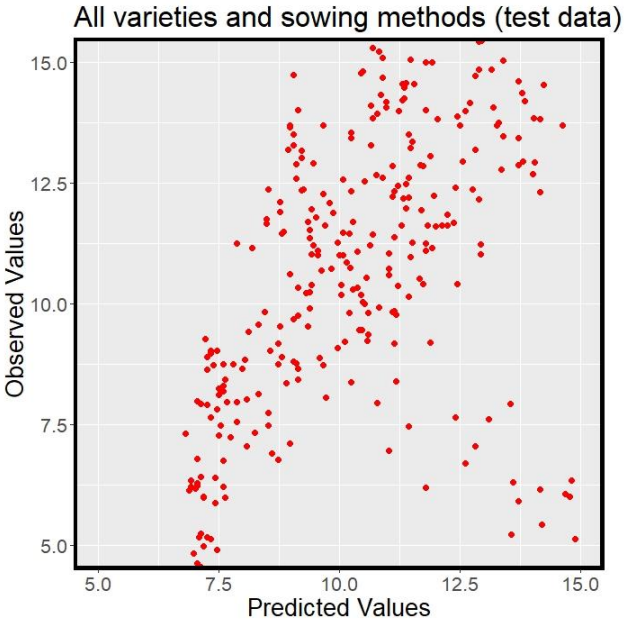

Figure 18: Model 3- test data 2020_21

The model accuracy was assessed as follows in Table 36:

Table 36: Model 3 - test data 2020_21

Evaluation Metric	Value
MAE	1.9
MSE	7
RMSE	2.7
R^2	0.26

The R2 value is 0.16, therefore only 16% of the variability in yield can be explained by the predictors in this model.

Summary

- The lowest RMSE of 2.7 is for model 3.
- The highest R^2 is for model (26%).
- The adjusted R^2 value is less than or equal to 0. 5 for all 3 models.

N_PI as numerical predictor

The general form of the regression equation using N_PI as numerical variable is as shown below:

$$QuadYield = \beta_0 + \beta_1\ PI\_N\_Uptake + \beta_2\ N\_PI + \epsilon$$

Test data set: 2015_16

The training data is constituted by the observations from all seasons except 2015_16.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 37):

Table 37: Model 1 - test data 2015_16

	Coefficients	95% CI	p-value
Intercept	8.213	(7.2746, 9.151)	<2e-16
PI_N_Uptake	0.022	(0.0134, 0.031)	1.3e-06
N_PI	0.020	(0.0074, 0.034)	0.0024

The adjusted R^2 value of the model is 0.185, therefore only approximately 19% of the variability is explained by the model. Hence the model cannot be considered much useful. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 38):

Table 38: Model 1 - test data 2015_16

Evaluation Metric	Value
MAE	1
MSE	1.4
RMSE	1.2
R^2	0.58

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 39):

Table 39: Model 2 - test data 2015_16

	Coefficients	95% CI	p-value
Intercept	7.219	(6.868, 7.570)	<2e-16

PI_N_Uptake	0.036	(0.033 0.039)	<2e-16
N_PI	0.024	(0.018, 0.029)	<2e-16

The adjusted R^2 value of the model is 0.401, therefore only approximately 40% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 40):

Table 40: Model 2 - test data 2015_16

Evaluation Metric	Value
MAE	1.7
MSE	4.8
RMSE	2.2
R^2	-0.34

The negative value of R2 indicates that the prediction is likely to be less accurate than the mean value of the data set over time.

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 41):

Table 41: Model 3 - test data 2015_16

	Coefficients	95% CI	p-value
Intercept	6.113	(5.833 6.393)	<2e-16
PI_N_Uptake	0.041	(0.039 0.044)	<2e-16
N_PI	0.026	(0.021 0.030)	<2e-16

The adjusted R^2 value of the model is 0.447, therefore approximately 45% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 42):

Table 42: Model 3 - test data 2015_16

Evaluation Metric	Value
MAE	1.3
MSE	2.8
RMSE	1.7

R^2	0.036

Summary

- Model 1 has the lowest RMSE of 1.2
- The highest R^2 is for model 1 (58%).
- Model 1 is the most accurate model.
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2016_17

The training data is constituted by the observations from all seasons except 2016_17.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 43):

Table 43: Model 1 - test data 2016_17

	Coefficients	95% CI	p-value
Intercept	8.061	(7.0934 9.029)	<2e-16
PI_N_Uptake	0.022	(0.0134 0.030)	1.1e-06
N_PI	0.019	(0.0071 0.031)	0.0021

The adjusted R^2 value of the model is 0.198, therefore only approximately 20% of the variability is explained by the model. Hence the model cannot be considered much useful. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 44):

Table 44: Model 1 - test data 2016_17

Evaluation Metric	Value
MAE	1.5
MSE	3.1
RMSE	1.8
R^2	0.2

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained as in Table 45:

Table 45: Model 2 - test data 2016_17

	Coefficients	95% CI	p-value

Intercept	7.611	(7.211 8.011)	<2e-16
PI_N_Uptake	0.031	(0.028 0.035)	<2e-16
N_PI	0.021	(0.016 0.026)	1.2e-14

The adjusted R^2 value of the model is 0.334, therefore only approximately 33% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 46):

Table 46: Model 2 - test data 2016_17

Evaluation Metric	Value
MAE	1.2
MSE	2.5
RMSE	1.6
R^2	0.55

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 47):

Table 47: Model 3 - test data 2016_17

	Coefficients	95% CI	p-value
Intercept	6.626	(6.338 6.915)	<2e-16
PI_N_Uptake	0.037	(0.034 0.040)	<2e-16
N_PI	0.024	(0.021 0.027)	<2e-16

The adjusted R^2 value of the model is 0.405, therefore approximately 41% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 48):

Table 48: Model 3 - test data 2016_17

Evaluation Metric	Value
MAE	1.3
MSE	3
RMSE	1.7
R^2	0.68

Summary

- Model 2 has the lowest RMSE of 1.6
- The highest R^2 is for model 3 (68%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2017_18

The training data is constituted by the observations from all seasons except 2017_18.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 49):

Table 49: Model 1 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	8.394	(7.4653 9.324)	<2e-16
PI_N_Uptake	0.0197	(0.0113 0.028)	6.8e-06
N_PI	0.0192	(0.0075 0.031)	0.0015

The adjusted R^2 value of the model is 0.168, therefore only approximately 17% of the variability is explained by the model. Hence the model cannot be considered much useful. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 50):

Table 50: Model 1 - test data 2017_18

Evaluation Metric	Value
MAE	1.4
MSE	3.1
RMSE	1.8
R^2	0.44

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 51):

Table 51: Model 2 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	7.449	(7.079 7.819)	<2e-16
PI_N_Uptake	0.033	(0.030 0.037)	<2e-16

N_PI	0.022	(0.017 0.027)	<2e-16

The adjusted R^2 value of the model is 0.372, therefore only approximately 37% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 52):

Table 52: Model 2 - test data 2017_18

Evaluation Metric	Value
MAE	1.3
MSE	2.6
RMSE	1.6
R^2	0.4

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 53):

Table 53: Model 3 - test data 2017_18

	Coefficients	95% CI	p-value
Intercept	6.431	(6.141 6.720)	<2e-16
PI_N_Uptake	0.039	(0.036 0.042)	<2e-16
N_PI	0.024	(0.021 0.028)	<2e-16

The adjusted R^2 value of the model is 0.422, therefore approximately 42% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 54):

Table 54: Model 3 - test data 2017_18

Evaluation Metric	Value
MAE	1.5
MSE	3.2
RMSE	1.8
R^2	0.45

Summary

- Model 2 has the lowest RMSE of 1.6
- The highest R^2 is for model 3 (45%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2018_19

The training data is constituted by the observations from all seasons except 2018_19.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 55):

Table 55: Model 1 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	8.280	(7.3169 9.24)	<2e-16
PI_N_Uptake	0.021	(0.0127 0.03)	2.5e-06
N_PI	0.018	(0.0065 0.03)	0.0026

The adjusted R^2 value of the model is 0.176, therefore only approximately 18% of the variability is explained by the model. Hence the model cannot be considered much useful. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 56):

Table 56: Model 1 - test data 2018_19

Evaluation Metric	Value
MAE	1.2
MSE	1.9
RMSE	1.4
R^2	0.51

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 57):

Table 57: Model 2 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	7.568	(7.186 7.951)	<2e-16
PI_N_Uptake	0.032	(0.029 0.035)	<2e-16

N_PI	0.020	(0.015 0.026)	1.5e-14

The adjusted R^2 value of the model is 0.343, therefore only approximately 34% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 58):

Table 58: Model 2 - test data 2018_19

Evaluation Metric	Value
MAE	1.2
MSE	2.1
RMSE	1.5
R^2	0.59

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 59):

Table 59: Model 3 - test data 2018_19

	Coefficients	95% CI	p-value
Intercept	6.919	(6.645 7.193)	<2e-16
PI_N_Uptake	0.036	(0.033 0.038)	<2e-16
N_PI	0.023	(0.020 0.026)	<2e-16

The adjusted R^2 value of the model is 0.414, therefore approximately 41% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 60):

Table 60: Model 3 - test data 2018_19

Evaluation Metric	Value
MAE	1.8
MSE	5.8
RMSE	2.4
R^2	0.27

Summary

- Model 1 has the lowest RMSE of 1.4
- The highest R^2 is for model 2 (59%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2019_20

The training data is constituted by the observations from all seasons except 2019_20.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 61):

Table 61: Model 1 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	8.324	(7.3401 9.308)	<2e-16
PI_N_Uptake	0.019	(0.0096 0.028)	8.7e-05
N_PI	0.021	(0.0095 0.033)	0.00054

The adjusted R^2 value of the model is 0.164, therefore only approximately 16% of the variability is explained by the model. Hence the model cannot be considered much useful. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 62):

Table 62: Model 1 - test data 2019_20

Evaluation Metric	Value
MAE	1.6
MSE	4.3
RMSE	2.1
R^2	0.21

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 63):

Table 63: Model 2 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	7.407	(7.028 7.787)	<2e-16
PI_N_Uptake	0.030	(0.027 0.034)	<2e-16
N_PI	0.023	(0.018 0.028)	<2e-16

The adjusted R^2 value of the model is 0.334, therefore only approximately 33% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 64):

Table 64: Model 2 - test data 2019_20

Evaluation Metric	Value
MAE	1.9
MSE	5.2
RMSE	2.3
R^2	0.24

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 65):

Table 65: Model 3 - test data 2019_20

	Coefficients	95% CI	p-value
Intercept	6.378	(6.109 6.648)	<2e-16
PI_N_Uptake	0.037	(0.034 0.040)	<2e-16
N_PI	0.026	(0.023 0.029)	<2e-16

The adjusted R^2 value of the model is 0.416, therefore approximately 42% of the variability is explained by the model. Hence this model tends to be more useful than the previous two models. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 66):

Table 66: Model 3 - test data 2019_20

Evaluation Metric	Value
MAE	2
MSE	5.7
RMSE	2.4
R^2	0.17

Summary

- Model 1 has the lowest RMSE of 2.1

- The highest R^2 is for model 2 (24%).
- The adjusted R^2 value is below 50% for all 3 models.

Test data set: 2020_21

The training data is constituted by the observations from all seasons except 2019_20.

Model 1: Drill sown Reiziq

Using drill sown Reiziq data from the training data set, a MLR model was fit on the training data and the following results were obtained (Table 67):

Table 67: Model 1 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	7.325	(6.584 8.067)	<2e-16
PI_N_Uptake	0.034	(0.027 0.040)	<2e-16
N_PI	0.022	(0.013 0.031)	3e-06

The adjusted R^2 value of the model is 0.471, therefore only approximately 47% of the variability is explained by the model. The scatterplot of observed vs predicted values is the same as model 1 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 68):

Table 68: Model 1 - test data 2020_21

Evaluation Metric	Value
MAE	2.6
MSE	13
RMSE	3.6
R^2	-0.59

The negative R-squared value means that the prediction accuracy tends to decrease compared to the average value of the data set over time.

Model 2: All drill sown varieties

Using drill sown data from the training data set for all varieties, a MLR model was fit on the training data and the following results were obtained (Table 69):

Table 69: Model 2 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	6.903	(6.566 7.239)	<2e-16
PI_N_Uptake	0.038	(0.035 0.041)	<2e-16
N_PI	0.024	(0.020 0.029)	<2e-16

The adjusted R^2 value of the model is 0.492, therefore only approximately 49% of the variability is explained by the model. Hence this model is more useful than model 1. The scatterplot of observed vs predicted values is the same as model 2 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 70):

Table 70: Model 2 - test data 2020_21

Evaluation Metric	Value
MAE	2.1
MSE	8.8
RMSE	3
R^2	0.072

Model 3: All varieties across all sowing methods

Using the entire training data set for all varieties across all sowing methods, a MLR model was fit, and the following results were obtained (Table 71):

Table 71: Model 3 - test data 2020_21

	Coefficients	95% CI	p-value
Intercept	6.084	(5.812 6.355)	<2e-16
PI_N_Uptake	0.041	(0.039 0.044)	<2e-16
N_PI	0.027	(0.024 0.030)	<2e-16

The adjusted R^2 value of the model is 0.478, therefore approximately 48% of the variability is explained by the model. Hence this model tends to be more useful than model 1 and as useful as model 2. The scatterplot of observed vs predicted values is the same as model 3 for the same test data when N_PI is categorical.

The model accuracy was assessed as follows (Table 72):

Table 72: Model 3 - test data 2020_21

Evaluation Metric	Value
MAE	1.9
MSE	6.7
RMSE	2.6
R^2	0.24

Summary

- Model 3 has the lowest RMSE of 2.6
- The highest R^2 is for model 3 (24%).
- The adjusted R^2 value is below 50% for all 3 models.

Comparison of models

When N_PI is used in the model as categorical variable, the factor level N_PI = 120 is not significant in most models. The values of regression coefficients are similar for all models for all individual test sets. Based on RMSE and R^2 values of each of the models, there is not enough evidence to suggest that there is a benefit to using separate models for each variety or sowing method over all varieties and sowing methods combined. Similarly, there is no significant difference in model results between N_PI being categorical or numerical. The linear regression model does not yield higher accuracy rates and therefore it is ideal to model the yield using more robust models such as the Random Forest model.

From the above analyses, a Multiple Linear Regression model has been fit on the full Rice data set with N_PI as categorical variable, the coefficients are as shown in Table 73 below.

Table 73: Final Model

	Coefficients	95% CI	p-value
Intercept	6.435	(6.180 6.691)	<2e-16
PI_N_Uptake	0.038	(0.036 0.041)	<2e-16
N_PI120	2.529	(1.927 3.130)	3.3e-16
N_PI60	1.553	1.320 1.787	<2e-16
N_PI90	3.082	2.246 3.917	7.1e-13

The MSE for the final model is 4.2. The adjusted R2 value is 0.432, so only 43% of the variability in rice yield can be explained by the predictors in the final model. Therefore, a more robust statistical technique needs to be undertaken to accurately model the yield.