International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Vol.6, Issue.2, pp.99-102, April (2019) **E-ISSN:** 2348-4519
DOI: https://doi.org/10.26438/ijsrmss/v6i2.99102

To Develop Accurate Model for Prediction of Soil Attributes with The Help of Statistical Modelling Techniques Such as Multiple Linear Regression (MLR) And Partial Least Square Regression (PLSR)

P. M. Shah[1*], D. C. Vyas[2]

[1] Department of Mathematics, Veer Narmad South Gujarat University, Surat
[2] Department of microbiology, Dolat Usha Institute of Applied Sciences, Valsad

Corresponding author: pinkyshah2302@gmail.com

Available online at: www.isroset.org

Received: 24/Mar/2019, Accepted: 19/Apr/2019, Online: 30/Apr/2019

Abstract**:** Present research paper signifies the importance of two statistical modelling techniques, Multiple Linear Regression (MLR) and partial Least Square Regression (PLSR). In the present study above mentioned methods are used to develop accurate model for prediction of soil property. MLR showes better result in comparison to PLSR in terms of Regression coefficient. In future, present models can be extended for soil data set of various other regions.

Keywords**:** soil pH, multiple linear regression, partial least square regression (PLSR).

I. INTRODUCTION

Soil attribute prediction has been practiced using various interpolation and regression techniques. The first application of interpolation and regression techniques for soil parameter estimation were based on the use of simple linear regression models between terrain attribute maps and soil parameters [1,2]. [3] Formed a generic framework for spatial prediction of soil variables based on regression-kriging. A methodological framework for spatial prediction based on regression-kriging is described and compared with ordinary kriging and plain regression was used in above mentioned research article. Earlier, [4] carried out comparison of prediction methods for the creation of field-extent soil property maps. Previously, [5] attempted the analysis of soil dataset using data mining and regression techniques. Looney and [6] used correlation coefficient with normal probability plots for soil parameter estimation.

Present study has attempted to develop an accurate statistical model for prediction of soil parameters such as Nitrogen (N), Phosphorus (P) and Potassium (K) and easy to measure soil parameter pH. For that we have successfully derived a statistical relationship between pH and N, P, K using two well-known statistical modeling technique multiple linear regression (MLR) and Partial least square regression (PLSR).

II. RELATED WORK

Earlier, a comparison of prediction methods for the creation of field-extent soil property maps was carried out [4]. [7] Building and testing conceptual and empirical models for predicting soil bulk density was carried out. Linear regression analysis technique was used by [8] to develop Model for Raft Foundation Supported on Dry Granular Soils. The major objective of the present research paper is to develop accurate model for prediction of soil attributes.

III. MATERIAL AND METHODS

Under the Soil Health Card Program of Government of Gujarat, soil samples from Valsad and Navsari Districts were collected by authorized locally trained farmers and brought for analysis to Soil Test Laboratory. From the soil samples, oil suspensions were prepared and analysis of pH, macronutrients like Phosphorus, Potassium (K), Sodium (Na) was carried out.

If possible, we intend to find the relation between the pH of soil and macronutrient content of the soil, on the basis of data availed. The details of the soil samples collected are listed in the following tables:

pH	N	P	K	pH	N	P	K
	(Kg/ ha)	(Kg/ha)	(Kg/Ha)		(Kg/ha)	(Kg/ha)	(Kg/Ha)
6.52	902	119	782	7.65	187	62.7	289
6.61	745	115	764	7.67	183	58	279
6.84	635	107	751	7.72	172	57.2	277
6.89	613	106	705	7.76	154	52	267
6.94	536	106	704	7.84	132	51.9	260
7.02	418	106	666	7.85	117	51.2	259
7.04	352	101	416	7.87	99.4	47.3	250
7.11	349	97.1	415	7.89	84.3	43.7	233
7.14	338	93.2	409	7.9	84.1	42.3	230
7.18	330	92.1	400	7.91	82.7	41.00	225
7.2	319	92.00	384	7.97	81.5	41.00	209
7.37	316	87.2	379	7.98	78.5	38.4	193
7.41	286	79.2	361	8.00	77.8	37.6	186
7.43	253	75.1	328	8.02	75.2	33.7	155
7.5	238	73.2	326	8.09	73.6	31.7	153
7.53	231	71.00	324	8.11	66.7	28.7	149
7.54	220	68.4	310	8.17	62.4	28.2	147

Table 1: Soil Details of Navsari District

A. Various regression techniques and governing equations:

1. The multiple linear equation (MLR).

The theory behind MLR has been well described in the literature and texts [9]. In multiple linear regressions, there are number of explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following figure.

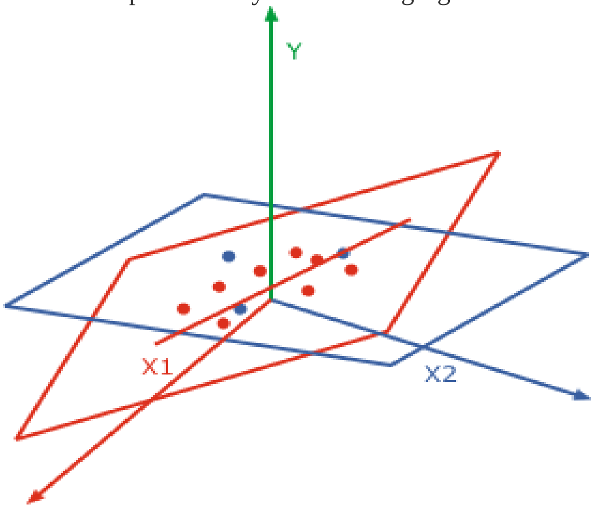

Figure.1: MLR: Regressing one Y-variable on a set of X-variables

In MLR a direct "least squares" regression is performed between the Y- and the X-matrix. In this section, the case of regression of one column vector Y, will be addressed for simplicity, but the method can readily be extended to a whole Y-matrix (as is common when MLR is applied to designed experiment data (DOE) on multiple responses. In this case one can make independent MLR models, one for each y-variable, based on the same X-matrix. The leave-one-out cross-validation method was used in the present study to select optimum model.

The following MLR model equation is just an extension of the normal univariate straight-line equation:

$$y = b_0 + b_1x_1 + b_2x_2 + ... + +b_kx_k + f$$

This can be compressed into the convenient matrix form:

$$y = Xb + f$$

The objective is to find the vector of regression coefficients *b* that minimizes *f*, the error term. This is where the least squares criterion on the squared error terms is used, i.e. find *b* so that f^Tf is minimized. MLR estimates the model coefficients using the equation:

$$b = (X^TX)^{-1}X^Ty$$

This operation involves the matrix inversion of the so-called *Dispersion Matrix* $(X^TX)^{-1}$. If any of the X-variables show any collinearity with each other i.e. if the variables are not linearly independent, then the MLR solution will not be stable (if there is a solution at all). Incidentally, this is the reason why the predictors are called independent variables in MLR; the ability to vary the X-variables independently of each other is a crucial requirement to variables used as predictors with this method. This is why in DOE; the initial design matrix is generated in such a way as to establish this independence (also called orthogonality) in the first place. MLR also requires more samples than predictors or the matrix cannot be inverted.

MLR has the following properties and behaviour:

- The number of X-variables must be smaller than the number of samples;
- In case of co linearity among X-variables, the b-coefficients are not reliable and the model may be unstable;
- MLR tends to over fit when noisy data are used.

In the present study, MLR regression analysis was performed to predict dependent variable pH of the soil using soil N, P and K concentration as the independent variable. MLR regression analysis was performed using Unscramble X (CAMO Software AS, Oslo, Norway) software. The leave-one-out cross-validation method was used in the present study to select optimum model.

2. Principles behind Multiple Linear Regression (MLR)

The basic MLR problem is an Analysis of Variance (ANOVA) problem. In ANOVA, the total variability is represented by the Total Sum of Squares (SS_T). This is defined as the squared sum of the deviations of each observation from the *Grand Mean* of the observations. The theory behind ANOVA states that SS_T can be further decomposed into two parts, a sum of squares due to regression SS_{reg} and a sum of squares due to random error SS_E. The ANOVA relationship is defined by,

$$SS_T = SS_{reg} + SS_E$$

Sum of squares due to error *SS*

The term SS_E is the term that is minimized in the least squares process. If the form of the model chosen to fit the data is correct (i.e. a linear model in this case), then the SS_E term should be normally and independently distributed

with a mean of zero and a variance s^2. In terms of the ANOVA relationship, when this term is minimized, the SS_{reg} term by definition is maximized.
Sum of squares due to error *SS*

2.2.4 Introduction to Partial Least Squares regression:
In the present study, PLS regression analysis was tested to predict dependent variable pH of the soil using soil N, P and K concentration as the independent variable. PLS regression analysis was performed using Unscramble X (CAMO Software AS, Oslo, Norway) software to determine the relative contribution of pH to the values of soil N, P and K. The leave-one-out cross-validation method was used in the present study to select optimum model.

Partial Least Squares (PLS) regression, also sometimes referred to as Projection to Latent Structures or just PLS, models both the X- and Y-matrices simultaneously to find the *latent (or hidden)* variables in X that will best predict the latent variables in Y. These PLS components are similar to principal components but will be referred to as *factors*. PLSR maximizes the covariance between X and Y. In this case, convergence of the system to a minimum residual error is often achieved in fewer factors than using PCR. This is in contrast to PCR, which first performs Principal Component Analysis (PCA) on X and then regresses the scores (T) vs. the Y data. A conceptual illustration for PLSR is shown graphically below.

3. PLSR Procedure:

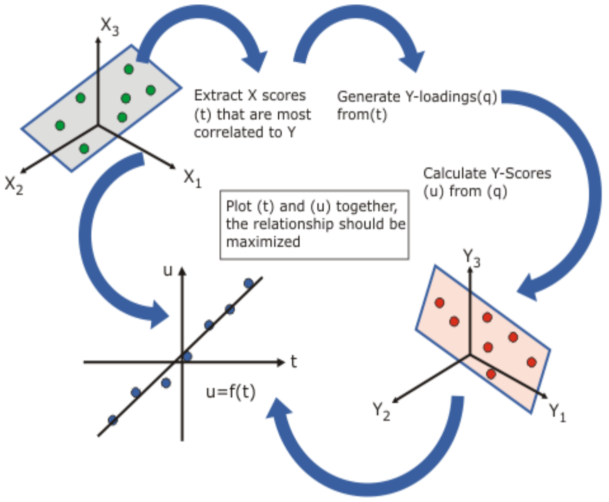
X3
X2
X1
Extract X scores (t) that are most correlated to Y
Generate Y-loadings(q) from(t)
Calculate Y-Scores (u) from (q)
Plot (t) and (u) together, the relationship should be maximized
u
t
u=f(t)
Y3
Y2
Y1

Figure.2 PLS regression carried out with one or more Y variables.

There are three algorithms available in The Unscrambler® for PLS regression such as NIPALS, Kernel PLS, and Wide Kernel PLS. In present study Kernel PLS algorithm is performed.

IV. RESULTS AND DISCUSSION

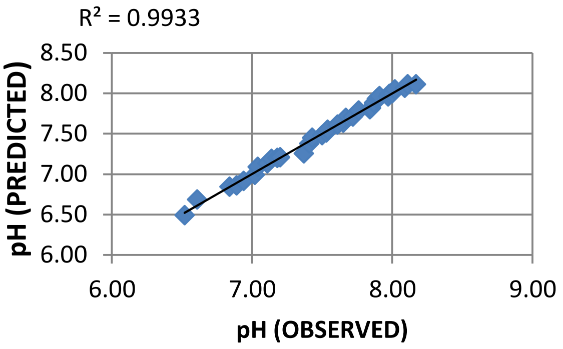

Figure: 3 Cross validation results for MLR

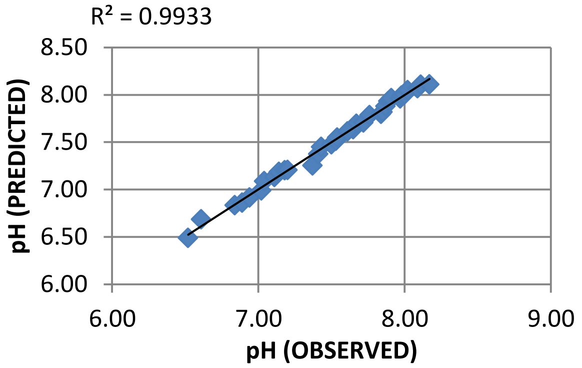

Figure: 4 Cross validation results for PLSR

Results obtained in present study is comparable with [10] where, results show that the two presented approaches provide similar capabilities to set up significant prediction models particularly for soil organic carbon and iron oxides.

Compared to other studies working in agricultural environments [11, 12], the accuracy of prediction models for both approaches developed in this study is slightly higher.

V. SUMMARY AND CONCLUSIONS

Table 2.2.1 shows various soil attributes measured from the study site. Critical observation of data revels sizable variation in the range of dataset. Mean and standard deviation of each data set represented in following table.

	pH	**N**	**P**	**K**
Mean	7.51	261.89	69.03	357.55
SD	0.44	206.01	27.73	186.05

It is important to note that both the regression modeling technique are able to extrapolate and predict pH with the help of N, P, K data set of the soil (Observe validation figure 3 and 4). However, MLR showes better result in comparison to PLSR in terms of Regression coefficient. Beta coefficients for the successful model of MLR are as follow. In future present models can be extended for soil data set of various other regions.

Intercept	8.476845
N	-0.00074
P	-0.01178
K	0.000124

REFERENCES

[1]. Gessler, P., Moore, I., McKenzie, N., Ryan, P., (1995), "Soil-landscape modelling and spatial prediction of soil attributes", International Journal of Geographical Information Systems 9 (4), 421 – 432

[2]. Moore, I., Gessler, P., Nielsen, G., Peterson, G., (1993), "Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57 (2), 443 – 452

[3]. Hengl, T., Heuvelink, G. B., & Stein. A, (2004), "A generic framework for spatial prediction of soil variables based on regression-kriging", Geoderma, 120(1-2), 75-93.

[4]. Bishop, T., McBratney, A., (2001), "A comparison of prediction methods for the creation of field-extent soil property maps", Geoderma, 103 (1 – 2), 149 – 160

[5]. Gholap, J., Ingole, A., Gohil, J., Gargade, S., & Attar, V. (2012), "Soil data analysis using classification techniques and soil attribute prediction". arXiv preprint arXiv,1206.1557.

[6]. Tranter G, Minasny B, McBratney AB, Murphy BW, McKenzie NJ, Grundy M, Brough D (2007) Building and testing conceptual and empirical models for predicting soil bulk density.

[7]. Al-Zaidee, S. R., Fadhil, A. T., & Al-Kubaisi, O. K. Using Finite Element to Modify Winkler Model for Raft Foundation Supported on Dry Granular Soils.

[8]. Looney, S., Gulledge Jr., T., (1985), "Use of the correlation coefficient with normal probability plots", The American Statistician, 39, 75 – 79

[9]. Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001). "A comparative analysis of multiple outlier detection procedures in the linear regression model", Computational statistics & data analysis, 36(3), 351-382.

[10]. Bayer, A., Bachmann, M., Müller, A., & Kaufmann, H. (2012). A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem. Applied and Environmental Soil Science, 2012.A.

[11]. Stevens, B. van Wesemael, H. Bartholomeus, D. Rosillon, B. Tychon, and E. Ben-Dor, "Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils," Geoderma, vol. 144, no. 1-2, pp. 395–404, 2008.

[12]. A. Stevens, T. Udelhoven, A. Denis et al., "Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy," Geoderma, vol. 158, no. 1-2, pp. 32–45, 2010.

AUTHORS PROFILE

Pinky. M. Shah Assistant professor (mathametics) department of mathamatics, Veer narmad south gujarat university, Surat.
Email: pinkyshah2302@gmail.com

Dhaval. C. Vyas, Assistant professor (botany) Department of microbiology, Dolat usha institute of applied sciences, valsad.
Email: vyasdhaval84@gmail.com

,