Indexed in
Scopus Compendex and Geobase Elsevier, Chemical
Abstract Services-USA, Geo-Ref Information Services-USA,
List B of Scientific Journals, Poland,
Directory of Research Journals

International Journal of Earth Sciences and Engineering

www.cafetinnova.org

April 2015, P.P.233-237

ISSN 0974-5904, Volume 08, No. 02

Regression Model to Predict Subsurface Properties

ARULKUMARAN S[1] AND SUKUMAR S[2]
[1]*Public Works Department*
[2]*Department of Civil Engineering, Government College of Engineering, Salem, South India*
Email: *geo_arul@yahoo.com*

Abstract: Prediction of geotechnical properties of subsurface is a primary task for the construction Engineers. Mathematical models are useful to generate the data by various interpolation techniques by knowing the identified properties with relating to one another. The objective of this study is to introduce Multiple Regression (MR) analysis by MS Excel and MATLAB programming to estimate the required subsoil's properties from the measured properties. The study is carried out for Salem block of South India. The developed model allows the decision maker to utilize the available data to arrive the possible values of non-visited locations. The results implies that the predicted soil parameters values are within permissible accuracy and the model is a useful tool to plan the components of a civil engineering project at the preliminary planning stage.

Keywords: *Regression Geotechnical model, Site investigation, MATLAB, Sub soil Data.*

1. Introduction

Geotechnical Engineering is very far from being an exact science. It must be remembered that the majority of unforeseen costs associated with construction are geotechnical in nature. If a reliable estimation models could be obtained to predict subsoil properties, it will be very valuable in the preliminary stage of project planning. Computing techniques such as artificial neural networks, fuzzy inference systems, evolutionary computation, etc. and their hybrids have been successfully employed for developing predictive models to estimate the parameters. But these techniques are expensive and are not intuitive to learn. Spread sheet software is one of the powerful and practical tools for developing model since it has a good user interface and optimization capability. The better the fit, the more accurately the function describes the data. Methods of using Microsoft Excel to fit non-linear functions has the advantage that Excel is probably included in the computer package as part of Microsoft Office, and thus no additional expense is required. MATLAB is also widely used software used for analyzing and displaying data as required.

1.1. Objective

The present study is based on the following objectives,

- To develop mathematical models to predict the subsoil properties of non-visited locations using Multiple Regression analysis by MS Excel and MATLAB software.
- To compare the models accuracy in prediction by calculating coefficient of determination (R^2) of the models.

1.2. Literature Review

Hammah and Curran has studied that the site investigation seems to be the area which is best suited for introducing statistics to geotechnical engineers[1] . Angus M. Brown introduces a simple, easily understood method for carrying out non-linear regression analysis based on user input functions [2]. Many studies are conducted to predict the value of un sampled soil from the measured values. Khaled Ahmad Aali estimates the saturation percentage (SP) using measured data of clay, silt sand and organic carbon [3].Gandthimathi developed the statistical linear regression analysis of soil parameters in Coimbatore city to predict the unknown parameters for geotechnical engineering purposes [4]. Sakunthala Devi developed a relation between SPT N and other geotechnical parameters with linear and quadratic regression analysis [5]. Nath and Mukherjee developed Statistical Analysis (Multiple Linear Regression) using Microsoft Excel software for developing new correlation between soil properties. The comparative study indicates that the developed model provide a satisfactory agreement with actual values hence it may be suggested as a useful alternative to laboratory analysis for preliminary design only for Kolkata area. Furthermore a large number of sample is needed for further analysis in order to develop a equation having high co-efficient of determination for better fit [6].

#SPL02080240 Copyright ©2015 CAFET-INNOVA TECHNICAL SOCIETY. All rights reserved.

Authors used Artificial Neural Network (ANN) and other advanced techniques for developing the prediction models and compare with regression models. Nihat Sinan Isik relates the index parameters and swell index and comes out with highest coefficient of determination (R^2) equals to 0.68 using ANN model [7]. Gokcen Yonter derived the relationships between physico-chemical properties of soils and runoff using SPSS software and his resulted R^2 (Coefficient of Determination) are in the range of 0.53-0.90 [8]. Mohamed A. Shahin et al realised that despite the success of ANN in solving many complex engineering problems they suffer from some shortcomings that need further attention in the future including model robustness, transparency and knowledge extraction, extrapolation and uncertainty [9]. Teruhisa Masada relates soil properties with each other using correlation equations in SPSS software and found that the coefficient of Determination (R2) equal to 1.0 [10]. Zohreh Izadifar compares the results of Multiple Regression (MR) with techniques such as ANN and found that MR performs better than ANN in prediction of evapo transpiration [11].

Isık Yilmaz obtained equation for prediction of the swell and showed that the equation obtained through Regression analysis has a prediction performance closer to the predictions using advanced neural network models [12]. Selim altun estimate the missing data in the study region using Variogram and Kriging methods and found that the results of predictions are within acceptable limit[13]. Artur Borowie use ANN for prediction of consistency parameters (plastic limit, liquid limit) of fen soils in comparison with the standard regression analysis. On the basis of the performed analysis he stated that trained ANN are able to predict consistency parameters of fen soils with acceptable error[14].Fevzi Akbas maps the soil properties using geo statistical (Semi variogram and Kriging method) for agricultural purpose[15].

MATLAB is used in handling data and writing task specific codes for models as well as in performing statistical analysis and curve fitting works. Esteban Hormazabal Zuniga use MATLAB as a computer tool to solve geotechnical problems involved in underground mine and open pit design[16].Scott H. Brown experience in analysing a multiple linear regression model using the MATLAB script approach is that it better enables one to observe what is going on "behind the scenes" during computations.[17]. Margaret Segou describes how MATLAB can be used in scientific research for Digital Signal Processing, Data Archiving and modelling complex natural systems through Optimization [18].

Though numbers of techniques are used in prediction models, Robinson studies shown that, there is not one single prediction method that can produce chief results for the generation of continuous soil property maps all of the time for the studied data set [19].The intent of this paper is to lead the reader through MR model using MS Excel and MATLAB coding to predict the soil parameters of non-visited locations and compare their accuracy of prediction.

2. Methodology

In this study, the data from 104 soil investigation bore wells has been collected. Soils were tested for determination of grain size distribution, Atterberg limits, swell percent (Free Swell Index), Standard Penetration Test (SPT) N values and other parameters according to the procedure suggested by Indian standards. Based on the properties and testing methods, the collected soil parameters are grouped as shown in table 1.

***Table 1**: Details of collected subsoil properties*

I. Field and Index Properties		X7	% Gravel, %G
X1	Latitude	X8	% Fines (Silt and Clay),%F
X2	Longitude	**II Properties from Lab and Field Tests**	
X3	Elevation, MSL in 'm'		
X4	Depth of Exploration, D_{SF} in 'm'	X9	Liquid limit, LL in %
X5	Thickness of poor Overburden , T in 'm'	X10	Plastic limit, PL in %
		X11	Differential Free Swell, DFS in %
X6	% Sand, %S	X12	SPT-N

In multiple regression analysis, the model is of the type,

$$Y = b_1\ x_1 + b_2\,x_2 + b_3x_3 + b_4x_4 + b_5x_5 +..+ b_n\ x_{n\,+}\ a \quad (1)$$

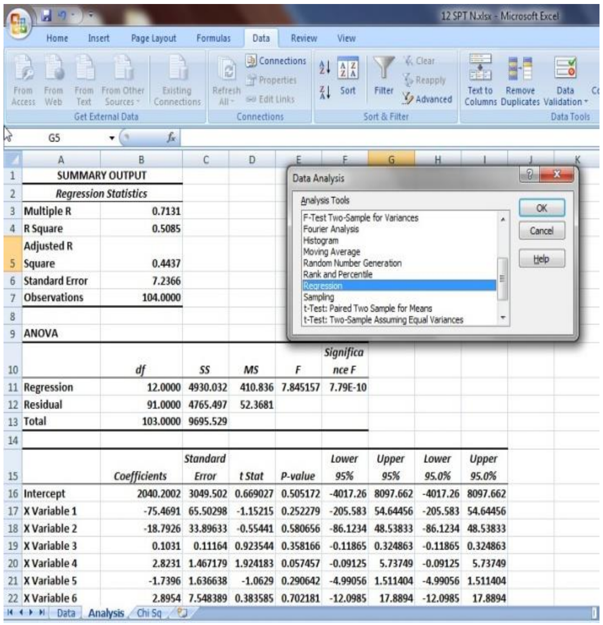
Where Y is the dependent variable, x_1, x_2 ... x_n are the independent variables, b_1, b_2... b_n are the coefficients of the respective independent variables which will be determined from the input data and "a" is a constant, where the regression line intercepts the y axis.

2.1. Data Processing and Analysis of Multiple Regressions using MS Excel

The collected parameters as listed in table 1 is correlated with each other's to derive an equation to find the parameter of interest. Most of the geotechnical design parameters are correlated with the Standard Penetration Test (SPT) N value. To find out N value, considerable resources (Man, Money and Time) are required. Deriving an equation to estimate the value of SPT N based on the collected data using a multiple regression model is explained here. Method of least square technique is used to develop this model. Data from all 104 borehole locations are analyzed and the following relationship between the SPT N values and the remaining 12 properties is derived as shown in equation (2).

$$\text{N-Value} = -52.56 * X1 - 24.34 * X2 + 0.073*X3 + 1.67 * X4 - 1.45 * X5 - 0.46 * X6 - 0.14 * X7 - 0.47 * X8 + 0.13 * X9 - 0.01 * X10 - 0.22 * X11 + 2570.7 \quad (2)$$

Thus, the obtained equation is used for the prediction of SPT N value for any location by knowing other properties without conducting elaborate field test. This will considerably save the resources in a project at preliminary stage. Similarly statistically significant and strong correlations are established among parameters. In this study, the subsoil parameters determined from field and laboratory test were correlated with other parameters using Multiple Regression analysis and the resulted R^2 are listed in table 2. Figure 1 shows the deriving of regression co efficient values using MS excel software.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7131
R Square	0.5085
Adjusted R Square	0.4437
Standard Error	7.2366
Observations	104.0000

ANOVA

	df	SS	MS	F	Significance F
Regression	12.0000	4930.032	410.836	7.845157	7.79E-10
Residual	91.0000	4765.497	52.3681		
Total	103.0000	9695.529			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2040.2002	3049.502	0.669027	0.505172	-4017.26	8097.662	-4017.26	8097.662
X Variable 1	-75.4691	65.50298	-1.15215	0.252279	-205.583	54.64456	-205.583	54.64456
X Variable 2	-18.7926	33.89633	-0.55441	0.580656	-86.1234	48.53833	-86.1234	48.53833
X Variable 3	0.1031	0.11164	0.923544	0.358166	-0.11865	0.324863	-0.11865	0.324863
X Variable 4	2.8231	1.467179	1.924183	0.057457	-0.09125	5.73749	-0.09125	5.73749
X Variable 5	-1.7396	1.636638	-1.0629	0.290642	-4.99056	1.511404	-4.99056	1.511404
X Variable 6	2.8954	7.548389	0.383585	0.702181	-12.0985	17.8894	-12.0985	17.8894

Figure 1*: Obtaining regression coefficients using Excel data analysis tool.*

2.2. Data processing and analysis using MATLAB

Matrix Laboratory (MATLAB) is numerical computing user friendly software which is developed by Mathwork. This is a language of technical computation tool useful for matrix form of analysis. MATLAB is used in handling data and writing task specific codes for models as well as in performing statistical analysis, curve fitting works and finding the statistical error .Most of the work done relies on the use of MATLAB as a common platform for preparing and processing data as well as for testing the suggested prediction models.

2.3. Estimation of Coefficient of Determination (R^2)

Coefficient of determination (R^2) which represents the fraction of the overall variance of the 'dependent' variable that is explained by the 'independent' variable. It is a measurement of how well the multiple regression line fits the data. It is calculated from the sum of the squares of the residuals and the sum of the squares of regression. The sum of the squares of the residuals captures the error between the estimate and the actual. The value of R^2 ranges from 0 to1.The calculation of R^2 using MS Excel and MATLAB are compared here. It has been seen that both the developed model has approximately equal R^2 values. Table 2 shows the value of R^2 by both models.

2.4. MATLAB Coding for R^2

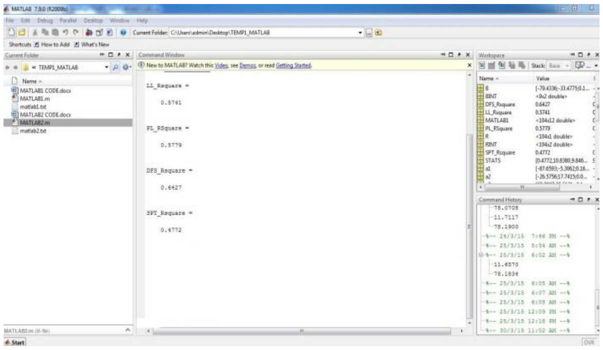
In order to calculate co efficient of determination R^2 , plotting observed and predicted curves, the coding is as follows.

```
%%% Beginning of the Program
Clear all
clc
%%%
=========================================
%%% Multi linear regression model for best fit, load
MATLAB1.txt  %%% loading the input data for
analysis
%%%
=========================================
input_data=MATLAB1(:,1:8);
output_data=MATLAB1(:,9);
x1=input_data(:,1);
x2=input_data(:,2);
x3=input_data(:,3);
x4=input_data(:,4);
x5=input_data(:,5);
x6=input_data(:,6);
x7=input_data(:,7);
x8=input_data(:,8);
y9=output_data(:,1);
[B,BINT,R,RINT,STATS] =
REGRESS(output_data,[input_data,ones(size(input_dat
a(:,1)))]);
a1=B;
LL_Rsquare=STATS(1) %%% R square Value
STATS; %%% R Square, F calue, p value, estimated
varience
 input_data=MATLAB1(:,1:8);
output_data=MATLAB1(:,10);
x1=input_data(:,1);
x2=input_data(:,2);
```

```
x3=input_data(:,3);
x4=input_data(:,4);
x5=input_data(:,5);
x6=input_data(:,6);
x7=input_data(:,7);
x8=input_data(:,8);
y010=output_data(:,1);
[B,BINT,R,RINT,STATS] =
REGRESS(output_data,[input_data,ones(size(input_data(:,1)))]);
a2=B;
PL_RSquare=STATS(1) %%% R square Value
STATS; %%% R Square, F calue, p value, estimated varience
 input_data=MATLAB1(:,1:8);
output_data=MATLAB1(:,11);
x1=input_data(:,1);
x2=input_data(:,2);
x3=input_data(:,3);
x4=input_data(:,4);
x5=input_data(:,5);
x6=input_data(:,6);
x7=input_data(:,7);
x8=input_data(:,8);
y011=output_data(:,1);
[B,BINT,R,RINT,STATS] =
REGRESS(output_data,[input_data,ones(size(input_data(:,1)))]);
a3=B;
DFS_Rsquare=STATS(1) %%% R square Value
STATS; %%% R Square, F calue, p value, estimated varience
 input_data=MATLAB1(:,1:8);
output_data1=MATLAB1(:,12);
x1=input_data(:,1);
x2=input_data(:,2);
x3=input_data(:,3);
x4=input_data(:,4);
x5=input_data(:,5);
x6=input_data(:,6);
x7=input_data(:,7);
x8=input_data(:,8);
y012=output_data1(:,1);
[B,BINT,R,RINT,STATS] =
REGRESS(output_data1,[input_data,ones(size(input_data(:,1)))]);
a4 =B;
SPT_Rsquare=STATS(1) %%% R square Value
STATS; %%% R Square, F calue, p value, estimated varience
 %%% Value calculation

y19=a1(9,1)+(a1(1,1)*x1)+(a1(2,1)*x2)+(a1(3,1)*x3)+(a1(4,1)*x4)+(a1(5,1)*x5)+(a1(6,1)*x6)+(a1(7,1)*x7)+(a1(8,1)*x8)-4;

y110=a2(9,1)+(a2(1,1)*x1)+(a2(2,1)*x2)+(a2(3,1)*x3)+(a2(4,1)*x4)+(a2(5,1)*x5)+(a2(6,1)*x6)+(a2(7,1)*x7)+(a2(8,1)*x8)+4;

y111=a3(9,1)+(a3(1,1)*x1)+(a3(2,1)*x2)+(a3(3,1)*x3)+(a3(4,1)*x4)+(a3(5,1)*x5)+(a3(6,1)*x6)+(a3(7,1)*x7)+(a3(8,1)*x8)+3;

y112=a4(9,1)+(a4(1,1)*x1)+(a4(2,1)*x2)+(a4(3,1)*x3)+(a4(4,1)*x4)+(a4(5,1)*x5)+(a4(6,1)*x6)+(a4(7,1)*x7)+(a4(8,1)*x8);
 % Plotting Observed Value Vs calculated value
x=1:104;
figure(1);
plot(x,y9,x,y19)
title('Property 1: LIQUID LIMIT')
xlabel('SITE ID')
ylabel('LIQUID LIMIT IN %')
Legend('OBSERVED','PREDICTED')
 figure(2);
plot(x,y010,x,y110)
title('Property 2: PLASTIC LIMIT')
ylabel('PLASTIC LIMIT IN %')
Legend('OBSERVED','PREDICTED')
 figure(3);
plot(x,y011,x,y111)
xlabel('SITE ID')
title('Property 3: DFS')
xlabel('SITE ID')
ylabel('DIFFERENTIAL FREE SWELL IN %')
```

***Table 2**: Value of R^2 in prediction of parameters by different analysis*

Sl No	Properties	By MS EXCEL	By MATLAB
1	Liquid Limit	0.590	0.587
2	Plastic Limit	0.588	0.591
3	Differential Free Swell	0.644	0.62
4	SPT N value	0.475	0.475


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\admin\Desktop\TEMP1_MATLAB
Shortcuts How to Add What's New
Current Folder
TEMP1_MATLAB
Name
MATLAB1 CODE.docx
MATLAB1.m
matlab1.txt
MATLAB2 CODE.docx
MATLAB2.m
matlab2.txt
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Workspace
Stack: Base
Command History
Start
OVR

```
LL_Rsquare =

    0.5741


PL_RSquare =

    0.5779


DFS_Rsquare =

    0.6427


SPT_Rsquare =

    0.4772
```

Name	Value
B	[-79.4336;-33.4775;0.1...
BINT	<9x2 double>
DFS_Rsquare	0.6427
LL_Rsquare	0.5741
MATLAB1	<104x12 double>
PL_RSquare	0.5779
R	<104x1 double>
RINT	<104x2 double>
SPT_Rsquare	0.4772
STATS	[0.4772,10.8380,9.846...
a1	[-87.6593;-5.3062;0.16...
a2	[-26.5756;17.7415;0.0...

```
78.0708
11.7117
78.1900
%-- 24/3/15  7:46 PM --%
%-- 25/3/15  5:34 AM --%
%-- 25/3/15  6:02 AM --%
11.6570
78.1834
%-- 25/3/15  6:05 AM --%
%-- 25/3/15  6:07 AM --%
%-- 25/3/15  6:08 AM --%
%-- 25/3/15 12:09 PM --%
%-- 25/3/15 12:18 PM --%
%-- 30/3/15 11:02 AM --%
```

***Figure 2**: Calculating R^2 values using MATLAB*

3. Results

The results of analysis show that the both MS Excel MATLAB models produce same Coefficient of Determination (R^2) values using MR methods in predicting the properties of soil. MATLAB have the advantage of calculating all parameters in a single programme. The authors emphasis that the procedure used here is to estimate the probable value and not for omission of detailed site investigation at any circumstances.

4. Conclusion

Based on the results the following conclusion were made,

- Multiple regressions (MR) are a powerful technique for standardizing data analysis. The results revealed that the developed models are useful in the field for practicing engineers in predicting the required geotechnical properties within permissible accuracy.
- The R^2 value calculated in this paper is designed to give the user an estimate of goodness of fit of the function to the data.

Acknowledgement

The author thanks to Editor-In-Chief Dr. D. Venkat Reddy, Professor, Department of Civil Engineering, National Institute of Technology, Surathkal, Mangalore and Dr.M .Anbarasu, Assistant professor, Govt. college of Engg.,Salem for their fruitful co-operation.

References

[1] Hammah, R.E and Curran, J.H., "Geostatistics in Geotechnical Engineering: Fad or Empowering?", *Geo Congress Conference*, Atlanta, USA, pp1-5, 2006.

[2] Angus M. Brown, "Step-by-step guide to non-linear regression analysis of experimental data using Microsoft Excel spreadsheet", *Journal of Computer Methods and Programs in Biomedicine*, 65, pp191–200, 2001.

[3] Khaled Ahmad Aali, Masoud Parsinejad and Bizhan Rahmani," Estimation of Saturation Percentage of Soil Using Multiple Regression, ANN, and ANFIS Techniques", *Journal of Computer and Information Science*, Vol 2 No 3, pp 127-136, 2009.

[4] Gandhimathi A, Arumairaj P.D, Lakshmipriya L. and Meenambal. T, "Spatial Analysis of Soil in Coimbatore for Geotechnical Engineering Purposes", *International Journal of Engineering Science and Technology*, Vol. 2(7), pp 2982-2996, 2010.

[5] Sakunthala Devi.S., and Stalin V K, "Development of soil suitability map for Geotechnical applications using GIS approach", *Indian Geotechnical Conference*, pp 797-800, 2011.

[6] Nath. S, and Mukherjee. S, "Use of empirical correlations for settlement estimation", *Proceedings of Indian Geotechnical Conference,* PP 404-410, 2013

[7] Nihat Sinan Isik , "Estimation of swell index of fine grained soils using regression equations and artificial neural networks", *Scientific Research and Essay*, Vol.4 (10), pp. 1047-1056,2009.

[8] Gokcen Yonter and Huriye Uysal, "The determination of the relationships between physical and chemical properties of soils to water erosion and crust strengths in Menemen Plain Soils, Turkey", *African Journal of Agricultural Research*, Vol. 7(2), pp 183-193, 2012.

[9] Mohamed A. Shahin, Mark B. Jaksa and Holger R. Maier, "Recent Advances and Future Challenges for Artificial Neural Systems in Geotechnical Engineering Applications", *Advances in Artificial Neural Systems*, Article ID 308239, 2009.

[10] Teruhisa Masada, "Shear Strength of Clay and Silt Embankments", Report No. FHWA/OH-2009/7, the Ohio Department of Transportation (ODOT) and the U.S. Department of Transportation, Federal Highway Administration, 2009.

[11] Zohreh Izadifar, "Modeling and Analysis of Actual Evapotranspiration using Data Driven and Wavelet Techniques", - M.S Thesis, Department of Civil and Geological Engineering University of Saskatchewan, Canada, 2010.

[12] Isık Yilmaz and Oguz Kaynar, "Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils", Expert *Systems with Applications,* Vol 38, pp5958-5966, 2011.

[13] Selim altun, Aburak Goktepe and Alper Sezer, "Geostatistical interpolation for modelling SPT data in northern Izmir", *Indian Academy of Sciences*, Vol. 38(6), pp. 1451–1468, 2013.

[14] Artur Borowiec and Krzysztof Wilk, "Prediction of consistency parameters of fen soils by neural networks", *Computer Assisted Methods in Engineering and Science*, Vol 21, pp 67–75, 2014.

[15] Fevzi Akbas, "Spatial variability of soil colour parameters and soil properties in an alluvial soil", *African Journal of Agricultural Research,* Vol 9 (12), pp 1025-1035, 2014.

[16] Esteban Hormazabal Zuniga, Pamela Ortubia Fernandez and Francisco Rovira Frez, "Some Applications to mine Geotechnical Design using MATLAB", Geotechnia, 2007.

[17] Scott H. Brown, "Multiple Linear Regression Analysis: A Matrix Approach with MATLAB", *Alabama Journal of Mathematics, 2009.*