RESEARCH ARTICLE

SPECIAL ISSUE

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05 , Issue 05S May
www.rspsciencehub.com

RSP Science Hub

http://dx.doi.org/10.47392/irjash.2023.S005

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

An Efficient Regression Method To Predict Soil pH Using RGB Values

Mithun Shivakoti[1]*, Srinivasa Reddy K*[1]*, Adinarayana Reddy*[2]

[1]*School of Computer Science & Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[2]*Department of Data Science and Artificial Intelligence, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education (IFHE), Hyderabad, Telangana, India.*

Email: mithun.shivakoti11@gmail.com

Article History

Received: 21 February 2023
Accepted: 14 March 2023

***Keywords*:**

Soil;
pH;
Prediction;
Machine Learning;
Regression;
Image processing.

Abstract

The fertility of a soil is governed by potential of Hydrogen (pH) value of the soil. This research paper presents a novel approach for predicting the pH value of a soil by using RGB (Red, Green, Blue) values of an image. The study utilizes machine learning techniques to develop a model that can accurately predict the soil pH based on the colour information captured in an image of the soil. The model was trained with a dataset containing RGB and corresponding pH value as the attributes and tested using a variety of images. Results show that the proposed model is able to predict soil pH with minimal error, demonstrating the potential for using image analysis as a practical and efficient method for soil pH determination in agriculture and soil science. With the available dataset, various regression approaches have been implemented to predict the soil pH value, and eventually the experimental results shows that the polynomial regression is the most effective method as the data is not linear for analysing this dataset.

1. INTRODUCTION

India is one of the countries that has abundant nutrients land which can be used for agriculture and farming is a time-honoured profession that has been carried out by people since the beginning of recorded history. Agriculture is increasing popularity not only in rural communities, but also among the urban people who are massively investing their time and showing interest in farming due to ever growing demand of agricultural products to serve the needs of the world's ever-increasing population. Agriculture was one of the fundamental forces that drove the industrial revolution, and the economy of a specific region. For the sake of ensuring agricultural sustainability, it is essential to gain an understanding of the long-term consequences of the many different methods of soil management and to take significant care of the soil quality.

Similar to air and water, soil is a fundamental natural resource that offers a wide range of benefits to humans in the form of commodities and services provided by ecosystems. Soil can be defined as a loose surface substance composed of rock debris and organic elements. Leaching, weathering, and the activity of microbes all work together to produce the great diversity of soil types that exist today. Each form of soil has its own set of advantages and disadvantages.

Long ago, soil was discovered, but only relatively lately did people realize how critical it is to preserve and expand the range of services provided by ecosystems. pH value of the Soil is one of the major factors to be considered before doing any cultivation (Barman). Soil pH is an important component in crop health and productivity, as well as the general health of an ecosystem. The pH value of soils can be tested to determine whether they are naturally acidic or alkaline and optimal plant growth depends on the proper pH balance, which can be determined through testing. The acidity or basicity of soil is indicated by its pH, with a pH of 7 being neutral, values less than 7 indicating acidity, and values greater than 7 suggesting basicity. A soil with 5.5 to 7.0 pH level is always good for cultivation (Barman et al.). Effective crop management and sustainable agriculture require accurate and efficient technologies for measuring and predicting soil pH.

In general, farmers took the soil samples to a laboratory that specializes in testing soil pH or consulting soil pH colour charts. Sometimes a specialist may assist the farmers in determining the pH value of the soil. However, obtaining the perspectives of experts is not always available in all situations. Again, each of these approaches requires some amount of time, work, and specialized knowledge. A soil pH chart is not an adequate method for determining the pH of soil since it requires human perception and the expertise of a trained professional. Calculating the pH value of soil in the laboratory requires the use of a soil pH meter in addition to a soil colour pH card. The technique for using a pH meter on soil has taken longer than an hour for a relatively simple soil sample. Automation is becoming increasingly prevalent in day-to-day life as a result of advances in technology and increased computer usage. The process is not only speed up as a result, but the end product is also less prone to errors. Image processing and regression are going to be the two methods that will be used to achieve the goal of determining the pH of the soil (Barman et al.).

The utilization of soil pictures, specifically through image processing and machine learning techniques is one of the promising strategies for forecasting soil pH. Regression analysis is a statistical tool for determining the relationship between one or more independent variables and a dependent variable (soil pH) (image features). By examining photographs of soil, elements such as colour, the pH value of the soil can be predicted.

Images have significant advantages over traditional approaches for predicting soil pH, such as soil sample and chemical analysis. For starters, it is a non-destructive procedure, which means that no soil is disturbed in order to collect a measurement. Second, it is a more efficient way since photographs can be used to assess large regions of soil swiftly and simply. Third, it can produce a more accurate estimate of soil pH since it considers many soil properties rather than depending on a single measurement.

In this research work, regression analysis is used to predict the value of soil pH using photographs also experiments were conducted using various regression models such as linear regression, random forest regression, decision tree regression, MLP Regression, and polynomial regression for estimating the value soil pH. For experimental work real-world data is used to evaluate the performance of these models and results are presented in this paper. The purpose of this study is to show the potential of using pictures to estimate soil pH and to give a foundation for future research in this area.

2. Literature Review

The soil pH was predicted using the RGB colour space of the photographs of the soil by [2,3, 4, 5, 6, 7, and 8]. A digital camera is used to take pictures of the soil, and then equation-1 is used to make a prediction about the soil pH.

$$Feature\ of\ Soil\ (pH\ Index)\ = Red/Green/Blue \qquad (1)$$

The researchers Abu et al. (Abu, Nasir, and Bala) developed an expert system that makes use of fuzzy logic to regulate soil pH. During the procedure, adjusted the pH level of soil was adjusted in order to allow the farmers to replace the fertilizer and guarantee that the plant would have a high quality. An approach that illustrates the soil's physical characteristics was presented by Babu and colleagues (Babu and Pandian). They implemented the fractal dimension technology through the usage of LabView.

In order to test the model, a 24-bit colour photographs are used as input and then transformed those images to 8-bit, and then extracted the features by using an equation that was suggested by Kumar et al. (Kumar et al.). Aziz et al. (Aziz, Ahmed, and Abraham) employed the same red, green, and blue values of images that were given by Kumar et al. (Kumar et al.), and they fed those values into the neural network for the purpose of training and testing, for which an accuracy rate of 80% was achieved by making use of ten hidden neurons in the hidden layer. Garibashvili and Mahantesh S.D et al (Gurubasava) determined the pH value of the soil based on the average of the RGB values of the photographs of the soil. The average values of RGB were compared to the actual pH of the soil, as well as their prediction of it. Barman et al. (Barman et al.) devised a method for predicting the soil pH by employing HSV colour image processing and regression techniques like linear, logarithmic, exponential, and quadratic predicted the soil pH in addition to the computation of the hue, saturation, and value of the soil pictures.

Sagar et al. (Sagar et al.) proposed a method for predicting the pH of soil in which photos are taken with the assistance of a camera and raspberry pi, and the pH of the soil is estimated by utilizing an algorithm applied to the obtained image after the image has been processed. Aside from soil pH, additional soil variables, such as soil moisture (Sagar et al. Matei et al. Pandey et al. Taheri-Garavand, Meda, and Naderloo), prediction of Azotobacterial population in soil (Ebrahimi et al.), soil mapping (Barman et al.), and soil organic matter (Mohan, Mridula, and Mohanan Ayoubi et al.) determined with the help of machine learning algorithms. The authors of these studies presented an analysis of the relationship between various soil parameters using ANN and regression. As a result of the relevant literature, it has come to our attention that methods of machine learning can be utilized to forecast several soil parameters, one of which is soil pH.

Anup Vibhute et al. (Vibhute and Koli) collected soil samples from different soil fields, took three photos of each sample, and then pre-processed the colour photographs to reduce noise. The results of their research may be found in the paper. Following the extraction of the various colour characteristics, such as RGB, Lab, HSV, and GLCM, a correlation was observed between the calculated features and the pH values. As a result, the regression analysis that produced the best results, with a low Mean Square Error (MSE) and a high R^2 value, was computed.

3. Proposed System

A. Dataset description

Dataset has been taken from Kaggle (ROBERT) in the form of a Comma Separated Value (CSV) file, which has values of pH for various combinations of Red(R), Green(G), Blue(B) values of an image. As it is an image the values of R, G, B would be in the range of 0-255. This dataset consists of 653 rows and there are 4 attributes, namely blue, green, red and label(pH). Table 1 shows the sample RGB values of the data set.

TABLE 1. Sample RGB values of the data set

Red(R)	Green (G)	Blue(B)
231	27	36
250	84	36
255	164	37
255	205	22
221	223	38
148	214	29
76	181	0
0	156	13
0	166	92

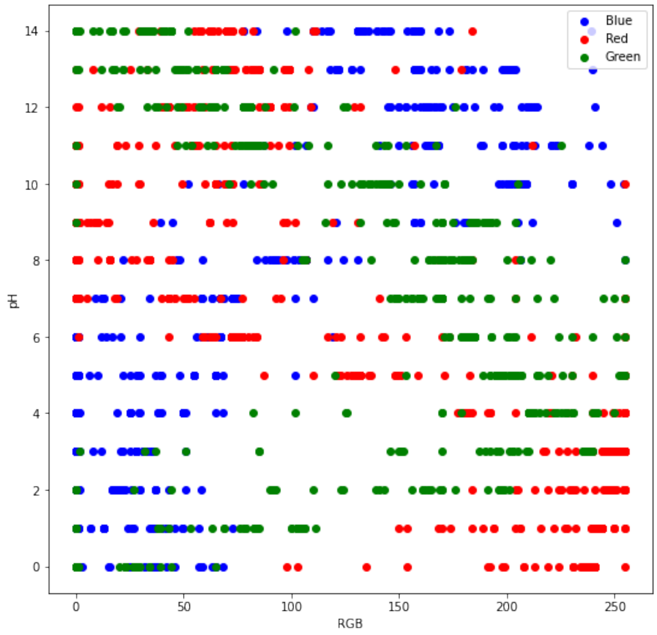

FIGURE 1. Scatter plot of RGB values and the corresponding pH values

Figure 1 depicts the scatter plot Red, Green and Blue values, in the range of 0-255 and their corresponding pH (0-14) in the dataset, it can be easily inferred from the plot that the datapoints aren't linear, rather widely spread across length and breadth of the plot.

B. Methodology

Linear Regression

For this, we employed a technique known as linear regression analysis. Linear regression, often known as the linear model, a statistical technique for predicting a numerical outcome variable (y). At least one predictor variable must be used in order to make this forecast (x). The goal is to provide an expression for the value of y as a function of x. Once a statistically sound model has been built, it may be used to forecast the future using the updated x values.

Random Forest Regression

Random Forest Regression algorithms are a type of Machine Learning technique that employs the usage of numerous random decision trees, each of which has been trained on a subset of data. The usage of several trees offers the algorithm stability and decreases variance. Because of its capacity to operate effectively with big and diverse datasets, the random forest regression technique is a popular model. Each tree is created by the algorithm using a distinct sample of input data. A different sample of characteristics is chosen for splitting at each node, and the trees operate in parallel with no interaction. The forecasts from each tree are then averaged to give a single outcome, which is the Random Forest prediction.

Decision Tree Regression

A Decision Tree Regressor is a type of supervised machine learning algorithm that is used for regression tasks. It creates a model in the form of a tree structure, with internal nodes representing feature(s) on which the data is split and leaf nodes representing the output. The algorithm recursively splits the data into subsets based on the feature that results in the highest reduction in impurity (e.g., variance or mean squared error). The final output of the tree is the average target value of the training samples in the corresponding leaf node. Decision tree regressors are simple to understand and interpret and can handle both linear and non-linear relationships between features and target.

Polynomial Regression

Polynomial Regression is a form of regression analysis in which the relationship between the independent variables and dependent variables are modelled in the nth degree polynomial. Polynomial Regression is a subcategory of Linear Regression, in which a curvilinear relationship between the dependent and independent variables is assumed to exist between the data points and the polynomial equation that is fit to the data. When there is no linear association between the variables, polynomial regression is the method that is utilized.

MLP Regression

A Multi-Layer Perceptron (MLP) Regressor is a type of artificial neural network that is used for supervised learning tasks, specifically for regression problems. It consists of multiple layers of artificial neurons, with the input layer receiving the input data, and the output layer providing the predicted output. The layers in between, called hidden layers, are used to learn complex representations of the input data. The MLP regressor is trained using a variant of the backpropagation algorithm, which adjusts the weights of the neurons in each layer in order to minimize the difference between the predicted and actual output.

Root Mean Squared Error

The Root Mean Square Error (RMSE), also known as the root mean square deviation, is a popular statistic used to evaluate the accuracy of a forecast shown in equation 2. Using Euclidean distance, it demonstrates how far the predicted values deviate from the actual values.

Here, we calculate the residual (the difference between the prediction and the truth) for each data point, then the norm of the residuals, then the mean of the residuals, and finally the square root of the mean to get the root mean squared error. The RMSE should be kept as low as possible for the best models.

$$RMSE = \sqrt{\sum_{i=1}^{N} \parallel y(i) - \hat{y}(i) \parallel \hat{2}/N} \tag{2}$$

where N is the number of data points, y(i) is the ith measurement, and ŷ(i) is its corresponding prediction.

Coefficient Of Determination(R^2)

It is a representation of the squared correlation between the known values of the outcomes that have been observed and the values that have been predicted by the model as shown in equation 3. The better the model, the higher the R^2 value should be. This correlation is represented as a value between 0.0 and 1.0. A value of 1.0 indicates a perfect fit and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the calculation fails to accurately model the data at all.

$$\gamma = n(\sum_{xy}) - (\sum_x)(\sum_y)/ \sqrt{[n\sum_{\widehat{x2}} - (\sum_x\widehat{)2}][n\sum_{\widehat{y2}} - (\sum_y\widehat{)2}]} \tag{3}$$

Where,

n = Total number of observations, Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value, Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value, Σy^2 = Sum of the Squares of the Second Value

The coefficient of determination = (correlation coefficient)^2 = R^2

Firstly, the dataset was imported followed by divided the dataset into train and test datasets in the ratio of 0.8: 0.2 (80% train and 20%test dataset). Much of data pre-processing has not been done as the dataset seemed to be perfect without any null values nor any unnecessary attributes. The obtained data are evaluated both qualitatively and quantitatively. The resulting values are also graphically illustrated.

Secondly, the predictions were done using multiple regression models namely Linear Regressor, Polynomial Regressor, Random Forest Regressor, Decision Tree Regressor and MLP Regressor, further the results have been graphically represented.

Various parameters have been customized, during the training of the model. They include setting maximum depth for Random Forest Regressor has been set to 2, degree of 4 for polynomial regression and finally for MLP Regressor maximum iterations were set to 50, early stopping being true and solver being 'Limited-memory Broyden–Fletcher–Goldfarb–Shanno', (which is usually used optimization algorithm for non-linear optimization problems). For both Linear Regression and Decision Tree Regressor default parameters have been used and no customizations have been done. The obtained RMSE and R2 score values for various regression models are shown in Table2.

From the Table 2 it can be analysed that the highest score was achieved by Polynomial Regressor (95.60%) followed Decision Tree Regressor (94.84%) and Random Forest Regression (92.77%) being the 3 best performing models while MLP Regression and Random Forest Regression has achieved R^2 score of 89.24% and 75.26%. Figure 2 and Figure 3 graphs represent the RMSE and R^2 score of selected regression models.

TABLE 2. Obtained RMSE and R^2 score values for various regression models

Model	RMSE	R^2 Score
Linear Regression	2.21	0.75
Random Forest Regression	1.19	0.92
Polynomial Regression	0.93	0.96
Decision Tree Regression	1.01	0.94
MLP Regression	1.46	0.89

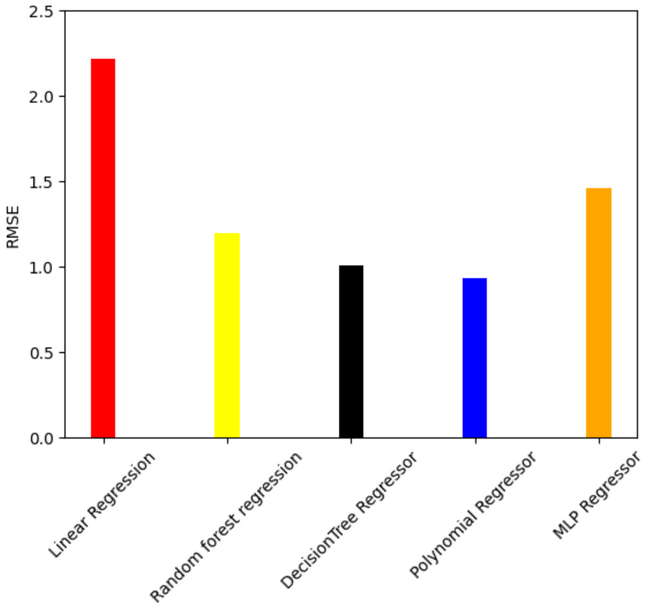

FIGURE 2. RMSE of proposed models

From figure 3, the Decision Tree Regression and Random Forest Regression had an edge over Linear Regression which is because the data is not linear and can be easily inferred from figure 1, with the data being widely scattered, these algorithms outperformed linear regression. Table 3 represents the comparison of the proposed models with the existing models.

4. CONCLUSION

The utilization of soil images, specifically through image processing and machine learning techniques is one of the promising strategies for forecasting the pH value of soil. Regression analysis is used in

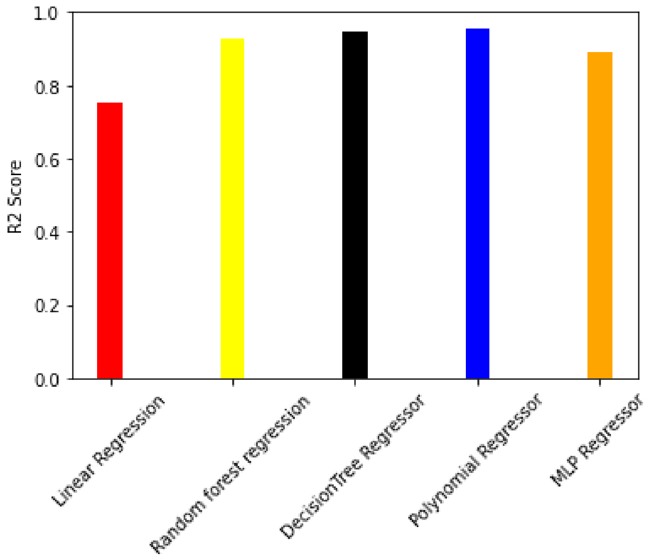

FIGURE 3. R^2 score of proposed models

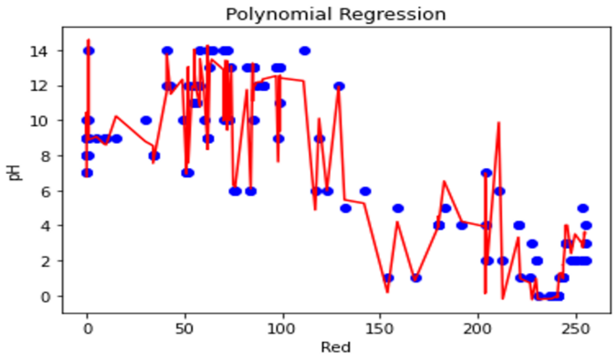

FIGURE 4. Polynomial regression graph between pH and attribute 'Red'

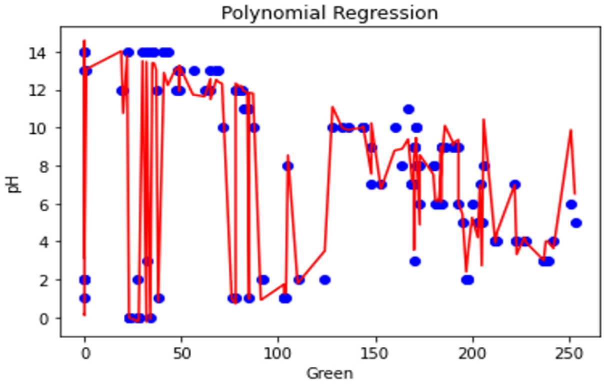

FIGURE 5. Polynomial regression graphbetween target and attribute 'Green'

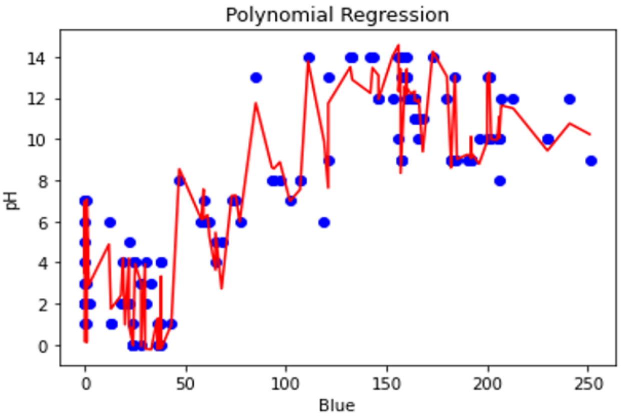

FIGURE 6. Polynomial regression graphbetween pH and attribute 'Blue'

TABLE 3. Comparison of proposed models R^2 value with the existing models

S.N	Title	R^2
1.	Soil pH Determination Using Mobile Phone Captured Image	0.6
2	Predication of soil pH using HIS colour image processing and regression over Guwahati, Assam, India	0.86
3.	Prediction of Soil pH using Smartphone based Digital Image Processing and Prediction Algorithm	0.94
4.	Determine the pH of Soil by using Neural Network Based on Soil's Colour	0.80
5.	Proposed Model	0.96

this paper for determining the relationship between one or more independent variables and a dependent variable (soil pH) (image features). Based on the findings the Polynomial regression model is the most effective. Linear regression is not possible to use on these data because it can't be linearly separated and attempting to use the linear regression with this dataset resulted in the model not functioning very well. In addition, Polynomial regression was producing encouraging findings, with a root mean squared error of 0.93 being the absolute minimum. As a result, polynomial regression followed by random forest regressor are effective because it accurately predicts the pH value of soil. Figures 4,5 and 6 are the polynomial regression graphs (between R, G and B attributes of the dataset with respect to pH) which clearly depicts how good the polynomial regression model fits with the data.

The novelty of this research work lies in the usage of R, G, B values separately for training and using an image during testing, where the R, G, B values of the image are extracted and sent to the model for pH predictions, whereas in the traditional pH prediction methods include both training and testing of model with image dataset. The proposed method has the potential to reduce the cost and time required for traditional soil testing methods. The study proposes a non-invasive method to predict soil pH using RGB values which can be easily measured by a camera which is also efficient and provides accurate results. Application incorporating this model is compatible to run on mobile devices which would make it easier for the users, especially farmers with minimal technical knowledge, as it is matter of only clicking the image of soil in the app and submitting it for the pH predictions. Limitations do include the sample size used in the study. As it is relatively small which may limit the generalizability of the findings. The future scope of the current work will be focusing on improvement of the pH predictions by applying deep learning techniques.

References

Abu, M A, E M M Nasir, and C R Bala. "Simulation of Soil PH Control system using Fuzzy Logic Method". *International Journal of Emerging Trends in Computer Image & Processing* 1 (2014): 15–19.

Ayoubi, Shamsollah, et al. "Application of Artificial Neural Network (ANN) to Predict Soil Organic Matter Using Remote Sensing Data in Two Ecosystems". *Biomass and Remote Sensing of Biomass* (2011): 978–953.

Aziz, M M, D R Ahmed, and B F Abraham. "Determine the pH of Soil by using Neural Network Based on Soil's Colour"". *International Journal of Advanced Research in Computer science and Software Engineering* 6 (2016): 51–54.

Babu, C S M and M Pandian. "Determination of Chemical and Physical Characteristics of Soil using Digital Image processing". *International Journal of Emerging Technology in Computer Science & Electronics* 2 (2016): 331–335.

Barman, Utpal. "Prediction of Soil pH using Smartphone based Digital Image Processing and Prediction Algorithm". *JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES* 14.2 (2019). 10.26782/jmcms.2019.04.00019.

Barman, Utpal, et al. "Predication of soil pH using HSI colour image processing and regression over Guwahati, Assam, India". *Journal of Applied and Natural Science* 10.2 (2018): 805–809. 10.31018/jans.v10i2.1701.

Ebrahimi, Mitra, et al. "Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses". *Computers and Electronics in Agriculture* 140 (2017): 409–421.

Gurubasava, Mahantesh S D. "Analysis of Agricultural soil pH using Digital Image Processing". *International Journal of Research in Advent Technology* 6 (2018): 1812–1816.

Kumar, Vinay, et al. "Determination of soil pH by using digital image processing technique". *Journal of Applied and Natural Science* 6.1 (2014): 14–18. 10.31018/jans.v6i1.368.

Matei, Oliviu, et al. "A Data Mining System for Real Time Soil Moisture Prediction". *Procedia Engineering* 181 (2017): 837–844. 10.1016/j.proeng.2017.02.475.

Mohan, R R, S Mridula, and P Mohanan. "Artificial Neural Network Model for Soil Moisture Estimation At Microwave Frequency"". *Progress In Electromagnetics Research M* (2015): 175–181.

Pandey, Abhishek, et al. "Artificial neural network for the estimation of soil moisture and surface roughness". *Russian Agricultural Sciences* 36.6 (2010): 428–432.

Sagar, S, et al. "Moisture And pH Detection Using Sensors And Automatic Irrigation System Using Raspberry Pi Based Image Processing". *International Journal of Engineering Technologies and Management Research* 5 (2018): 153–157. 10.5281/zenodo.1202079.

Taheri-Garavand, Amin, Venkatesh Meda, and Leila Naderloo. "Artificial neural Network−Genetic algorithm modeling for moisture content prediction of savory leaves drying process in different drying conditions". *Engineering in Agriculture, Environment and Food* 11.4 (2018): 232–238. 10.1016/j.eaef.2018.08.001.

Vibhute, Anup and M S Koli. “Soil pH parameter Estimation Using Image Processing and Regression Analysis”. *Journal of Xi’an University of Architecture & Technology* XII (2020).

Embargo period: The article has no embargo period.

To cite this Article: Shivakoti, Mithun, Srinivasa Reddy K, and Adinarayana Reddy. “**An Efficient Regression Method To Predict Soil pH Using RGB Values**.” International Research Journal on Advanced Science Hub 05 .05S May (2023): 35–42. http://dx.doi.org/10.47392/irjash.2023.S005