The Pharma Innovation Journal 2021; 10(5): 380-385

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2021; 10(5): 380-385

www.thepharmajournal.com
Received: 10-03-2021
Accepted: 19-04-2021

S Manishankar
Research Scholar, Department of Remote Sensing & GIS, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

R Kumaraperumal
Assistant Professor, Department of Remote Sensing & GIS, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

KP Ragunath
Assistant Professor, Department of Remote Sensing & GIS, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

Balaji Kannan
Associate Professor, Department of Soil and Water Conservation Engineering, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

Corresponding Author:
S Manishankar
Research Scholar, Department of Remote Sensing & GIS, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

Selection of environmental covariates using stepwise regression

S Manishankar, R Kumaraperumal, KP Ragunath and Balaji Kannan

Abstract
Stepwise regression is the iterative step-by-step design of a regression model requiring the option of independent variables to be used in a final model. The Stepwise regression is used to identify the most influencing variable. It is one of the statistical techniques for reducing the dimension of the data. The study was conducted in the Sarkarsamakulam block of the Coimbatore district of Tamil Nadu with 17 profile points. A total of 33 environmental covariates are used for this analysis and to make the analysis easier and accurate the covariates data has to be reduced. From the result, the least influencing variable can be omitted for further analysis. Totally ten variables were selected using this regression for further analysis. Hence it is one of the easiest methods to predict the most influencing variable using R software.

Keywords: Stepwise regression, coefficients, probability, F-statistic

Introduction
Environmental covariates, which reflect the soil formation factors, are a key approach in spatial prediction of soil properties. Climate, organism, relief or topography, and parent material are the five groups of environmental covariates classified by the CLORPT model, and these variables are briefly listed in Table 1. Climate is the most critical element in all categories. Since topography has such a significant effect on soil distribution and vegetation, it is frequently ignored as a passive component in soil formation. The environmental covariates of each group are distinct. With a wider range of environmental covariates, it is highly difficult to predict the outcomes. Stepwise regression analysis (using a F probability of 0.05 for the chosen factor) is the most widely used method for selecting the suitable regression equation, as reported by Landau and Everitt (2017) [4], and it was performed using SPSS. To identify the most influencing environmental covariates, this stepwise regression was used. It is one of the methods for reducing variables without much loss of information (Smith, 2018) [8].

Material and Methods
The study was conducted in the Sarkarsamakulam block of the Coimbatore district of Tamil Nadu. The geography of the study area lies between 1101'19.8" to 11012'21" N Latitude and 76056'31.1" to 7705'56.7" E Longitude. It covers an area of about 174.3 sq. kms. It is present at an altitude of 440 meters above Mean Sea Level (MSL). Sarkarsamakulam block comprises a total of 13 villages in it. The climate is of semi-arid and sub-tropical monsoonic type. The average annual rainfall is about 647.2 mm and it receives a major part of rainfall from North East Monsoon. The mean annual maximum and minimum temperature are 32.7^{0}C and 21.5^{0}C respectively. The major soil types are Black soil, Red soil, and Brown soil. The major soil orders in the Sarkarsamakulam block are Vertisols, Inceptisols, Entisols, Ultisol, and Alfisols these are the informations which are gathered from the Soil Survey Department of Coimbatore district.

The dataset was obtained from the Department of Remote Sensing and GIS, Tamil Nadu Agricultural University. The environmental covariates are Satellite data (Green, Blue, NIR, Red), Agro Climatic Zones (ACZ), Agro-Ecological Zones (AEZ), Western Ghats (WG), Maximum Temperature, Minimum Temperature, Climate, Rainfall, Land Use and Land Cover (LUCU), Elevation, Hill shading, Aspect, Convergence Index, General curvature, Maximum Curvature, Minimum Curvature, Profile Curvature, Tangential Curvature, Longitudinal Curvature, Plan Curvature, Total Curvature, LS factor, Slope, Mid Slope Position, Physiography, Topographic Wetness Index (TWI), Topographic Ruggedness Index (TRI), Total Catchment Area (TCA), Valley Depth, Geomorphology and Geology.

In order to identify the most influencing variables, the selected 1980 points with corresponding 33 layer staked variables are used to run the stepwise regression in SPSS software.

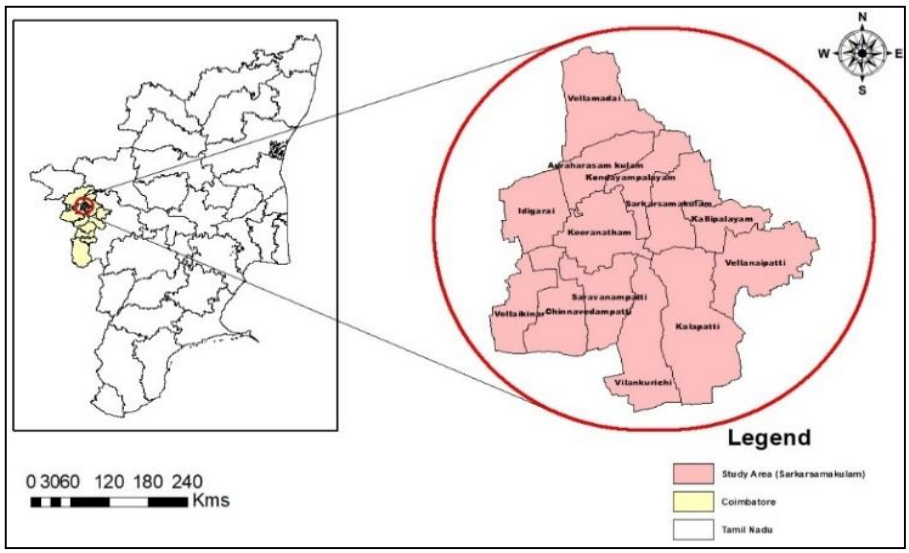

Fig 1: Location of the study area

Table 1: Parameters of Environmental Covariates

Relief/Topography		Climate	Organism	Parent material
Elevation Hill shading Aspect Convergence Index General Curvature Longitudinal Curvature LS factor Max Curvature Mid Slope Position Min Curvature Physiography Plan Curvature	Profile Curvature Slope Tangential Curvature TRI TWI TCA Total Curvature Valley Depth Western Ghats	Max Temperature Min Temperature Annual Rainfall ACZ AEZ	LULC Satellite data: NIR Green Blue Red	Geomorphology Geology

Stepwise regression

Stepwise regression is the iterative step-by-step design of a regression model requiring the option of independent variables to be used in a final model (Wilkinson and Dallal, 1981) [9]. It entails successively adding or deleting possible explanatory variables and testing for statistical significance with each iteration. Based on the test statistics and level of significance of the predicted coefficients, the variables to be added or removed are selected (McIntyre *et al.*, 1983) [6]. The availability of statistical software packages, also in models with hundreds of variables, makes stepwise regression possible.

Types of stepwise regression

Stepwise regression can be done either by evaluating one independent variable at a time and including it if it is statistically significant in the regression process or by including all potential independent variables in the model and removing non-statistically significant variables. Some use a mixture of both methods, so there are three stepwise regression methods:

- Forward Selection
- Backward Elimination
- Bidirectional Elimination

Forward selection

Forward selection starts with no model variables, measures each variable as it is applied to the model, and preserves the most statistically relevant variables, continuing the procedure until the outcomes are optimal. A forward-selection rule begins with no explanatory variables and then introduces variables, one by one, depending on the variable being the most statistically relevant, until no statistically significant variables remain (Henderson and Denison, 1989) [2].

Backward elimination

For all possible explanatory variables, a backward elimination rule begins and then discards the least statistically important variables, one by one. When each variable left in the equation is statistically important at that time the discarding stops. When there is a sufficient number of variables, backward elimination is difficult and impossible if the number of variables is greater than the number of observations. The stopping criteria differs for each method (Bendel and Afifi, 1977) [1].

Bidirectional elimination

The bi-directional elimination procedure is a combination of both forward selection and backward elimination methods that test which variables should be added or removed. The wrinkle is that the method now takes into account the statistical effects of removing variables that were previously used at any stage. For example, a variable was added in step 3 and removed in step 5, and then again added in step 8.

Working principle of stepwise regression

As its name tells how the stepwise regression works, it selects the variables in step by step manner. Stepwise regression is

used to find the lowest number of predictors that can correctly predict the dependent variable. The statistical criterion of optimizing the R2 of the included variables is used to apply variables to the regression equation one at a time (Smith, 2018) [8]. When variable is entered, the model is checked to see whether any of the included variables will improve the model if they were omitted.

When all available variables have been added, or when no statistically meaningful change in R^2 can be achieved using all of the variables not yet included, the process of adding additional variables comes to an end (Johnsson, 1992) [3]. Since variables will only be used in the regression equation if they contribute statistical importance to the analysis, all of the independent variables chosen for inclusion will have a statistically meaningful relationship to the dependent variable.

Start the test with all the predictor variables available (the' Backward: method), removing one variable while the regression model advances at a time. This strategy can be used if it has a modest number of predictor variables and want to delete only a few. The variable with the lowest 'F-to-remove' statistic is removed from the model at each step (Leigh, 1988) [5]. The statistic 'F-to-remove' is estimated as follows:

A t-statistic is computed for each variable's predicted coefficient in the model. "The t-statistic is squared, creating the statistic of "F-to-remove. Start the test with no predictor variables (the "Forward" method), incorporating as the regression model advances one at a time. Use this approach if it has large set of predictor variables. The "F-to-add" statistic is formed using the same steps above, except the method would measure the statistic for each variable not in the model. The variable is added to the model with the largest 'F-to-add' statistic.

Advantages of stepwise regression:

- It has the ability to handle vast numbers of potential predictor variables, fine-tuning the model from the available choices to pick the right predictor variables.
- It is quicker than other types of automated model selection.
- It can provide valuable information about the quality of the predictor variables to observe the order in which variables are included or excluded (Landau and Everitt, 2017) [4].

Results and Discussion

Table 2 displays the test of significance of the model using an ANOVA. The ten ANOVAs that are reported correspond to ten models. The stepwise procedure adds only one variable at a time to the model as the model is “slowly” built. At the tenth step and beyond, it is also possible to remove a variable from the model.

Table 2: Tests of Significance for Each Step in the Regression Analysis

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12802.238	1	12802.238	3.572E3	.000[a]
	Residual	4336.358	1210	3.584		
	Total	17138.597	1211			
2	Regression	14781.947	2	7390.974	3.792E3	.000[b]
	Residual	2356.649	1209	1.949		
	Total	17138.597	1211			
3	Regression	15184.724	3	5061.575	3.129E3	.000[c]
	Residual	1953.873	1208	1.617		
	Total	17138.597	1211			
4	Regression	15578.206	4	3894.551	3.013E3	.000[d]
	Residual	1560.391	1207	1.293		
	Total	17138.597	1211			
5	Regression	15968.063	5	3193.613	3.290E3	.000[e]
	Residual	1170.534	1206	.971		
	Total	17138.597	1211			
6	Regression	16879.032	6	2813.172	1.306E4	.000[f]
	Residual	259.564	1205	.215		
	Total	17138.597	1211			
7	Regression	16997.527	7	2428.218	2.072E4	.000[g]
	Residual	141.069	1204	.117		
	Total	17138.597	1211			
8	Regression	17011.836	8	2126.479	2.018E4	.000[h]
	Residual	126.761	1203	.105		
	Total	17138.597	1211			
9	Regression	17022.064	9	1891.340	1.951E4	.000[i]
	Residual	116.533	1202	.097		
	Total	17138.597	1211			
10	Regression	17026.312	10	1702.631	1.821E4	.000[j]
	Residual	112.284	1201	.093		
	Total	17138.597	1211			

In stepwise, the final model was built in ten steps; each step resulted in a statistically significant model. The degrees of freedom (df) column shows that one variable was added during each step (the degrees of freedom for the Regression effect track this for us as they are counts of the number of predictors in the model). No variables were removed from the model since the count of predictors in the model steadily increases from 1 to 10.

The deduction was verified by the display given in table 3, which tracks variables that have been entered and removed at each step. SPSS starts with zero predictors and then adds the strongest predictor, AEZ to the model if its b-coefficient in

statistically significant ($p< 0.05$, see last column). It then adds the second strongest predictor (WG). This process continues until none of the excluded predictors contributes significantly to the included predictors. AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology, Physio, Tmax, Planc, Tangc, LS, Midslope, Geology, Rainfall, LULC, Prof, Climate, TCA, Red have been entered on Steps 1 through 10, respectively, without any variables having been removed on any step.

Table 3: Variables that were Entered and Removed

Model	Variables Entered	Variables Removed	Method
1	AEZ	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	WG	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	Valley	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	ACZ	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
5	Tmin	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
6	DEM	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
7	Geomorphology	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
8	Physio	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
9	Tmax	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
10	Planc	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

The Model Summary presents the R Square and Adjusted R Square values for each step along with the amount of R Square Change was given in table 4. R is simply the Pearson correlation between the actual and predicted values. The R Square with that predictor in the model was .862. The square of the correlation between AEZ and WG ($.864^2 = .747$), and is the value of R Square Change. The R square value at the end of the model was 0.993 which means that our 10^{th} predictors account for 99.3% of the variance.

Table 4: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.864[a]	.747	.747	1.89308	.747	3572.285	1	1210	.000
2	.929[b]	.862	.862	1.39616	.116	1015.623	1	1209	.000
3	.941[c]	.886	.886	1.27179	.024	249.020	1	1208	.000
4	.953[d]	.909	.909	1.13701	.023	304.368	1	1207	.000
5	.965[e]	.932	.931	.98519	.023	401.669	1	1206	.000
6	.992[f]	.985	.985	.46412	.053	4229.080	1	1205	.000
7	.996[g]	.992	.992	.34230	.007	1011.334	1	1204	.000
8	.996[h]	.993	.993	.32461	.001	135.790	1	1203	.000
9	.997[i]	.993	.993	.31137	.001	105.497	1	1202	.000
10	.997[j]	.993	.993	.30577	.000	45.443	1	1201	.000

a. Predictors: (Constant), AEZ
b. Predictors: (Constant), AEZ, WG
c. Predictors: (Constant), AEZ, WG, Valley
d. Predictors: (Constant), AEZ, WG, Valley, ACZ
e. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin
f. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin, DEM
g. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology
h. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology, Physio
i. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology, Physio, Tmax
j. Predictors: (Constant), AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology, Physio, Tmax, Planc
Dependent Variable: Series

Table 5 shows the Coefficients table that provides the details of the entered predictor. Both the raw and standardized regression coefficients are readjusted at each step to reflect the additional variables in the model (Xu and Zhang, 2001) [10]. The b-coefficients of selected predictor are all significant and the final model states that Subgroup = 336.695 – 4.439 AEZ – 2.089 WG – 0.014 Valley + 5.021 ACZ – 15.558 Tmin – 0.078 DEM + 0.438 Geom – 0.046 Physio + 1.038 Tmax + 205.210 Planc

Table 5: The Results of the Stepwise Regression Analysis

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
10	(Constant)	336.695	4.416		76.251	.000
	AEZ	-4.439	.019	-1.174	-234.625	.000
	WG	-2.089	.031	-.236	-66.424	.000
	Valley	-.014	.000	-.125	-34.216	.000
	ACZ	5.021	.049	.643	102.945	.000
	Tmin	-15.558	.151	-.524	-103.069	.000
	DEM	-.078	.001	-.294	-68.657	.000
	Geomorphology	.438	.013	.116	33.175	.000

	Physio	-.046	.003	-.048	-14.859	.000
	Tmax	1.038	.113	.032	9.213	.000
	Planc	205.210	30.441	.017	6.741	.000

Plan curvature was the strongest predictor in this model. One of the regression assumptions is that the residuals (prediction errors) are normally distributed was given in figure 2. The scatterplot with predicted values on the x-axis and residuals on the y-axis was given in figure 3. This chart does not show much violations of the independence, homoscedasticity, and linearity assumptions. It shows a striking pattern of descending straight lines. Standardizing both variables may change the scales of our scatterplot but not its shape.

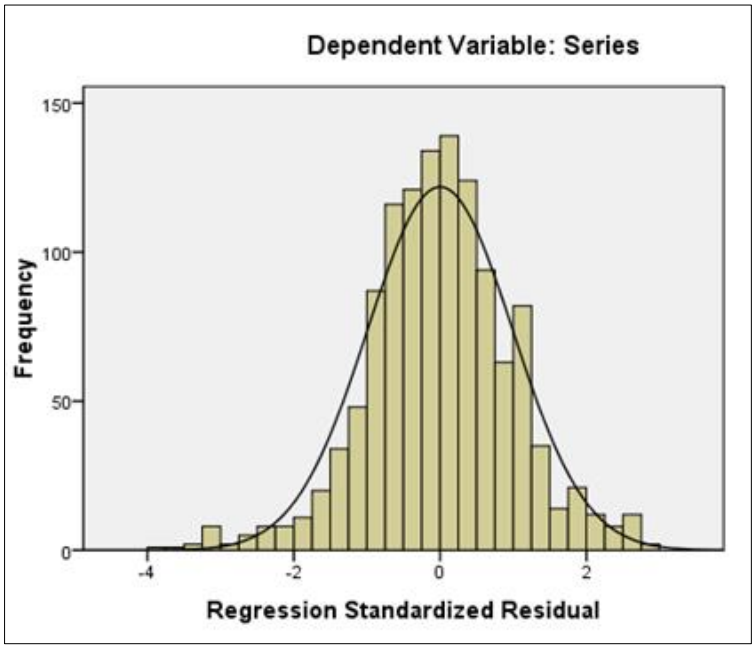

Fig 2: Histogram

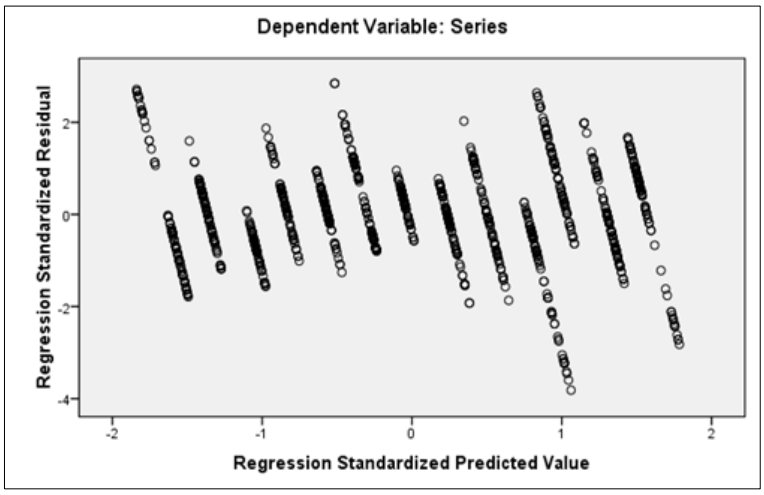

Fig 3: Scatter plot of dependent variables

Conclusion

The prediction model contained ten of thirty-three predictors and it was reached in tenth steps with no variables removed. The model was statistically significant, F (10, 1201) with $p<$ .001, and accounted for approximately 99% of the variance (R^2 = .993). Soil classes were primarily predicted by a lower level of Valley and to a higher level of plan curvature. The raw and standardized regression coefficients and their structure coefficients are shown in Table 5. Plan curvature received the strongest weight in the model followed by ACZ, Temp maximum, and DEM; temp minimum receives the lowest of the ten weights. AEZ, WG, Valley, ACZ, Tmin, DEM, Geomorphology, Physio, Tmax, Planc, Tangc, LS, Midslope, Geology, Rainfall, LULC, Prof, Climate, TCA, Red are the significant predictors selected through stepwise regression.

References

1. Bendel RB, Afifi AA. Comparison of stopping rules in forward “stepwise” regression. Journal of the American Statistical association 1977;72(357):4653.
2. Henderson DA, Denison DR. Stepwise regression in social and psychological research. Psychological Reports 1989;64(1):251-257.
3. Johnsson T. A procedure for stepwise regression analysis. Statistical Papers 1992;33(1):21-29.

4. Landau S, Everitt BS. A handbook of statistical analyses using SPSS 2017.
5. Leigh JP. Assessing the importance of an independent variable in multiple regression: is stepwise unwise?. Journal of clinical epidemiology 1988;41(7):669-677.
6. McIntyre SH, Montgomery DB, Srinivasan V, Weitz BA. Evaluating the statistical significance of models developed by stepwise regression. Journal of Marketing Research 1983;20(1):1-11.
7. Pandit V, Khairullah ZY. Stepwise regression choosing the proper level of significance. In Proceedings of the 1985 Academy of Marketing Science (AMS) Annual Conference. Springer, Cham 2015,395-398p.
8. Smith G. Step away from stepwise. Journal of Big Data 2018;5(1):1-12.
9. Wilkinson L, Dallal GE. Tests of significance in forward selection regression with an F-to-enter stopping rule. Technometrics 1981;23(4):377-380.
10. Xu L, Zhang WJ. Comparison of different methods for variable selection. Analytica Chimica Acta 2001;446(1-2):475-481.
11. Zhang Z. Variable selection with stepwise and best subset approaches. Annals of translational medicine 2016,4(7).