Journal of Indian Geomorphology
Volume 12, 2024 ● ISSN 2320-0731
Indian Institute of Geomorphologists (IGI)

Detection of Multicollinearity among Soil Geomorphological Variables: Study on Southwestern Hills of West Bengal

Arindam Sarkar
Department of Geography, Purash Kanpur Haridas Nandi College, Howrah, West Bengal - 711410
E-mail: arindam.srkr1@gmail.com (Corresponding author)

Abstract: *Multicollinearity analysis is a popular statistical technique that deals with situations where multiple variables are linearly correlated. This technique can easily detect the presence of a strong correlation. The multicollinearity algorithm's primary outputs are the correlation matrix, route-square, tolerance, and variance inflation factor. The present study is concentrated on the Southwestern hills (Ajodhya, Garpanchakot, Biharinath, and Susunia) of West Bengal. Seventy-seven soil samples are collected from the top of the hills to the foothill pediment areas according to changes in elevation and slope. The location of sample points was determined using a GPS handset. The elevation of the sample points has been measured using an altimeter and further rechecked by overlay operation on the SRTM digital elevation model. A clinometer compass has been used to determine the slope of the sample locations. Soil and geomorphology are deeply correlated with each other. Two data sets with standard geomorphological and different soil variables have been employed for the study. Soil variables were determined in the laboratory using different soil sample analysis methods. This research article attempts to find multiple correlations among the soil and geomorphological variables as they are closely related. Strong multiple linear correlations have been detected between the variables of sand, silt, clay, organic carbon, and organic matter through route square, tolerance, and inflation factor values. Multicollinearity can be resolved in different ways, and principal component analysis is one of them.*

Keywords: Multicollinearity, Ajodhya, Garpanchakot, Biharinath, Susunia.

Introduction

Geomorphologists are currently researching to quantify the rates of bedrock weathering, a crucial process that significantly impacts the Earth's surface (Richter *et al.*, 2019). This topic warrants extensive discussion among soil scientists to understand its implications further through different quantitative techniques. Multicollinearity is associated when more than two explanatory variables are linearly correlated in a multiple regression model. Multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated (Daoud, 2017). The term multicollinearity denotes a linear relationship among some or all predictor variables of

the regression model. If there is no linear relationship between predictor variables, they are considered orthogonal (Jensen and Ramirez, 2012). The term multicollinearity is introduced in economics by the economist Ranger Frisch. Multicollinearity may be a problem in the regression model because we cannot distinguish individual effects between the independent and dependent variables. In regression analysis, there are many assumptions about the model, namely, multicollinearity, nonconstant variance (non-homogeneity), linearity, and autocorrelation (Osborne and Waters, 2002). Multicollinearity occurs when two or more independent variables in a regression model are deeply correlated (Stephanie, 2015). Multicollinearity reveals that an independent variable can be predicted from another independent variable in a regression model. A fundamental assumption in the multiple linear regression model is that the rank of the matrix of observations on explanatory variables is the same as the number of explanatory variables (Shalabh, 2019). Despite this, the classical linear regression is associated with another assumption — there is no relation among independent variables in a regression model. However, the results of several studies indicate that the predictor variables are nearly perfectly correlated. If one or more assumptions are violated, the model is no longer reliable and is unacceptable in estimating the population parameters (Daoud, 2017). Multicollinearity analysis has mainly been used in economics (Ando and Modigliani, 1963; Alesina and Dollar, 2000; Mela and Kopalle, 2002; Alvi and Senbeta, 2012; Baharumshah *et al.*, 2017; Baidoo *et al.*, 2018; Daniel *et al.*, 2021), earth science (Box *et al.*, 1989; Aerts and Chapin, 1999; Sandra, 2002; Brooks *et al.*, 2006; Avitabile *et al.*, 2012; Ayanu *et al.*, 2012; Mercy *et al.*, 2016; Kim, 2019) and data science (Belsley *et al.*, 1980; Adeboye, 2014).

Multicollinearity study has excellent opportunities to identify the relative importance of independent variables when explaining the variation caused by the dependent variables. When multicollinearity is present in the data set, the confidence interval of the coefficient and statistics tends to become very wide and very small. Rejection of the null hypothesis is uncompromising when multicollinearity is present in the data set. Because multicollinearity is associated with a lesser p-value, the data set has enough evidence to conclude that the null hypothesis is true. In that case, the significant value of that test remains more extensive than the predestined significance value. The data set does not have enough proof to reject the null hypothesis, which means the result is not statistically significant. Multicollinearity is a problem because it can detect the statistical significance of an independent variable. In the present study soil, geomorphological samples have been collected using a random stratified method, so there are some possibilities, like inaccurate use of dummy variables and repetition of the same type of variable, and variables are highly correlated. In this regard, the multicollinearity test can get the perfect solution. The present study is mandated to spot multicollinearity among the soil geomorphological variables better, to understand the significance of the independent variable and its result. It is possible because the degree of multicollinearity significantly impacts the p-value and coefficient result. Multicollinearity is the most popular statistical tool often used in regression analysis and statistical analysis when dealing with an extensive, diverse database and one has some desired output.

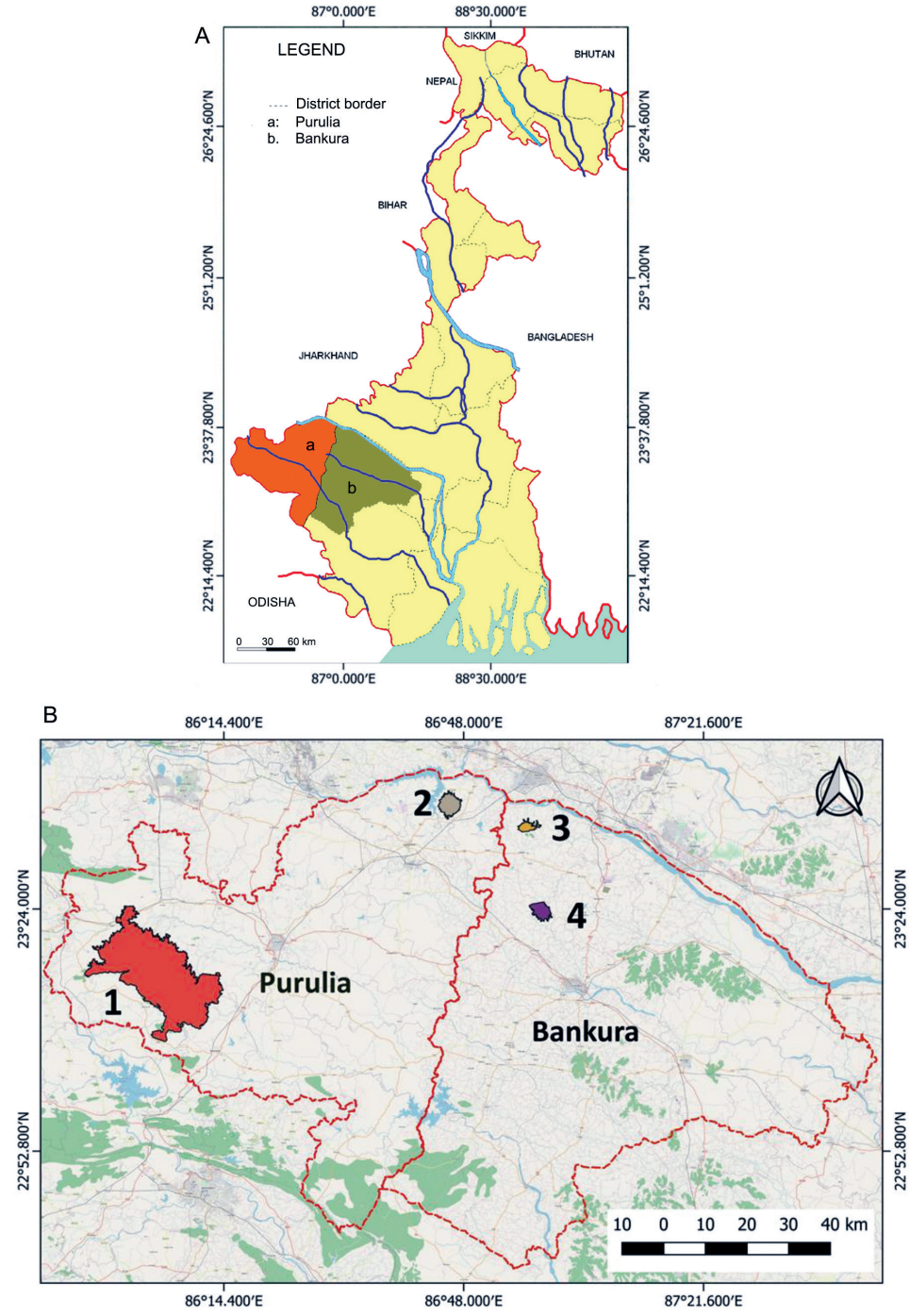

Figure 1. (A) represents the location of Purulia and Bankura districts within the administrative boundary of West Bengal, and (B) shows the location of Ajodhya (1), Garpanchakot hills (2) in Purulia district and location of Biharinath (3), Susunia (4) hills in Bankura district.

Soil geomorphology is a scientific study of soil character according to the change of geomorphological parameters such as elevation and slope (Sarkar, 2019a). The evolution and development of soil geomorphology as a new branch of geomorphology is associated with the emergence and blooming of geomorphology and pedology (Zinck *et al*., 2016). Several studies have been conducted on geology (Coenders and Saez, 2000; Kroll and Song, 2013; Bager *et al.*, 2017), soil (De, 1972; Das *et al.*, 2010), geomorphology (Dasgupta, 2015) and soil geomorphology (De, 1984a; De, 1984b; De, 1987; De and Chatterjee, 2009; Sarkar and Das, 2018; Sarkar, 2019a; Sarkar, 2019b; Sarkar, 2019c; Sarkar, 2021) in the study area.

It is observed that, there is a lack of scientific articles on statistical data analysis, like multicollinearity. The present study aims to detect, fix, and give some solutions regarding multicollinearity among the soil geomorphological data set. Soil geomorphological data sets are associated with several independent factors with high correlation. That is why a multicollinearity study is very significant for research in soil geomorphology. The present study is concentrated in the southwestern hills of Ajodhya, Garpanchakot, Biharinath, and Susunia of Purulia and Bankura districts of West Bengal (Fig. 1 and 2).

Data source and methodology

Two different sets of data of soil geomorphological variables have been introduced for the present study (Fig. 3). The first set of data is associated with 2 geomorphological variables (elevation and slope), 7 sediment and soil variables (gravel, very coarse sand, coarse sand, moderate sand, fine sand, very fine sand, silt and clay). The second data set contains the same geomorphological variables with 8 soil variables (sand, silt, clay, hygroscopic moisture, organic carbon, organic matter, pH, and salinity). Total 77 soil samples (Fig. 2) have been collected from the top of the hills to the foothill areas according to changes in the range of elevation and slope following a random stratified method. GPS determines the locations of soil samples. Elevation and slope of the sample point have been measured by altimeter as well as clinometer and rechecked by overlay of GPS waypoint on SRTM digital elevation model (SRTM1N22E085V3, SRTM1N22E086V3; SRTM1N22E087V3, SRTM1N23E085V3, SRTM1N23E086V3, SRTM1N23E087V3; Date of Accusation on 11/02/2000; Spatial resolution: 1-Arc second, Source. https://earthexplorer.usgs.gov/).

Data behaviour, geographical background

In the present study, more than one variable of soil geomorphology is considered as an independent variable. The assumption has been made depending on the variables' association, structure, and behaviour. Variables are linearly correlated (Sarkar and Das, 2018; Sarkar, 2019a; Sarkar, 2019b; Sarkar, 2019c; Sarkar, 2021). Here, equal importance is asserted for each variable. The topography of the area is well understood from the mean elevation (251.519 m) and mean slope (10°) of the sample points. The standard deviation (Table 1) of elevation and slope indicates high variation in topography of the area from where samples are collected. This area is dominated by moderate sand, gravel, and coarse sand (Table 1). This sand-dominated area is characterised by less soil hygroscopic moisture, moderate soil organic matter, and stable pH conditions (Table 1).

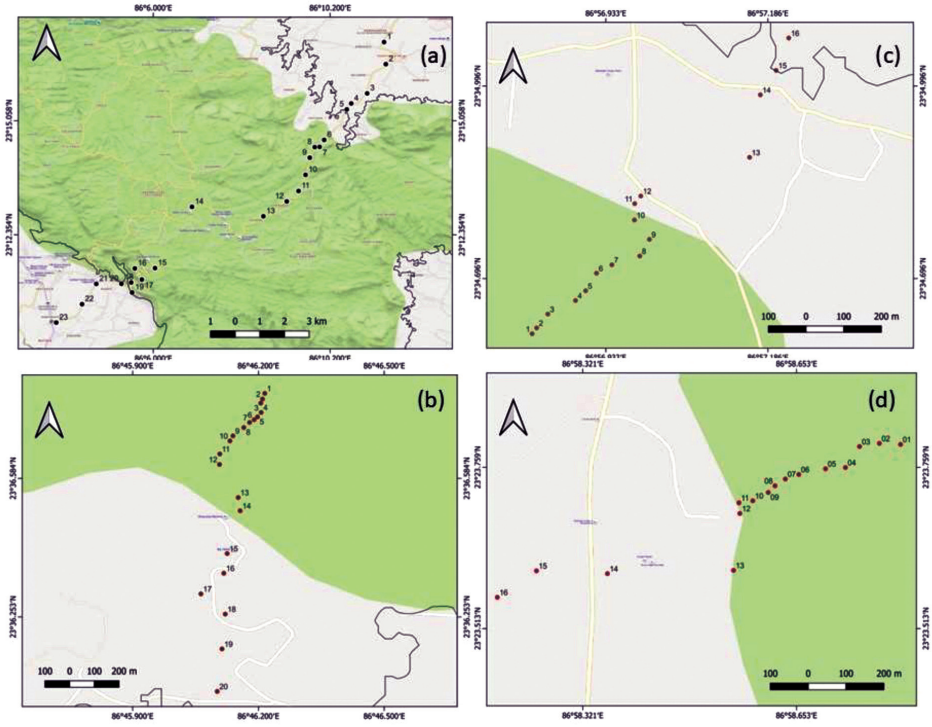

Figure 2. A broad view of sample locations in the study area. Location of samples in the foothills and the hilly parts of Ajodhya (a), Garpanchakot (b), Biharinath (c), and Susunia (d).

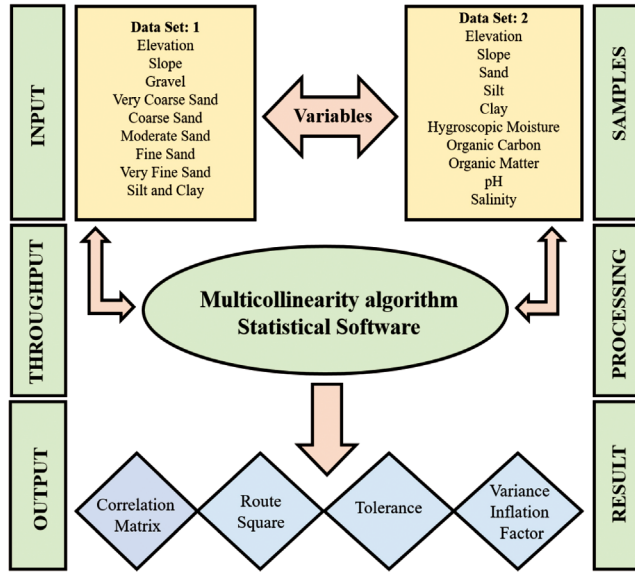

Figure 3. Methodological flow chart.

Table 1. Summary statistics for the study area. It provides a concise overview of the main characteristics of the dataset, which are the key features of the area under consideration in the present research.

Variable	Minimum	Maximum	Range	Mean	Standard Deviation
Data Set: I					
Elevation (m)	125.000	562.000	437.00	251.519	106.975
Slope (degrees)	2.500	32.500	30.00	10.799	8.092
Gravel (gm)	0.000	880.600	880.60	195.888	236.686
Very coarse sand (gm)	2.825	191.300	188.60	55.832	39.416
Coarse sand (gm)	17.360	233.000	215.64	99.728	51.849
Moderate sand (gm)	107.700	781.400	673.70	345.575	167.021
Fine sand (gm)	23.400	541.900	518.50	174.986	100.129
Very fine sand (gm)	2.797	122.600	119.81	49.440	26.560
Silt and clay (combined) (gm)	0.220	43.000	42.78	5.400	6.432
Data Set: II					
Elevation (m)	125.000	562.000	473.00	251.519	106.975
Slope (degrees)	2.500	32.500	30.00	10.799	8.092
Sand (%)	37.500	86.250	48.75	68.468	11.885
Silt (%)	4.500	36.750	32.25	17.250	7.089
Clay (%)	4.000	41.000	37.00	14.286	8.055
Hygroscopic moisture (%)	0.040	5.485	5.44	1.536	1.318
Organic carbon (%)	0.390	19.600	19.21	4.689	5.278
Organic matter (%)	0.608	33.712	33.10	8.000	9.116
Soil pH	2.540	7.720	5.18	6.784	0.752
Salinity (µ/cm)	27.600	188.310	160.71	61.016	34.127

Elevation and slope of sample points range between 125–562 m and 2°–32° respectively. Maximum variation has been detected in the samples, such as gravel, moderate sand and fine sand (Table 1), as well as the percentage of clay, organic matter, and soil salinity (Table 1). The Eigen value of elevation (variable 1), slope (variable 2), and sand (variable 3) correspond to high percentage (38.24%, 24.55%, 11.53%) of variance (Sarkar, 2019a). Geomorphological variables like elevation and slope, as well as soil variables like sand, have been considered as principal soil geomorphological variables in this area (Sarkar and Das, 2018; Sarkar, 2019a; Sarkar, 2019b; Sarkar, 2019c; Sarkar, 2021).

The southwestern hills of Ajodhya, Garpanchakot, Biharinath, and Susunia are located in the basement of the eastern extending part of the Chhotanagpur plateau (Mahapatra, 2008; Bandyopadhyay *et al.*, 2014). This part depicts nearly four to five thousand million years of geological succession (Bhattacharya *et al.*, 1985). The formation of the peninsular shield was evidenced during the Archean age, with the oldest granite and gneiss rock composition (GSI, 1999). The present landscape of this area is the result of the climatological and geological processes which have been controlling its morphological evolution throughout the geological past (Sarkar,

2019a). This area is associated with several topographic expressions of its geomorphic evolution. Significant slope break is present from west to east with gently undulating topography, occasional hillocks and an elevation range between 650 m to 50 m above the mean sea level (Sarkar, 2019a). This is significant evidence of the peneplanation process working over this area (Sarkar, 2019a). The soil depth in the foothill pediment area is relatively thick (Sarkar, 2019a). Transported soil is present in the foothill pediment area (Das *et al.*, 2010; Ghosh, 2013). The area is entirely dominated by sandy loam soil texture, followed by loamy sand, which is usually made of sand along with lower amounts of silt and clay (Sarkar, 2019b).

Table 2a. Correlation matrix table showing the correlation coefficients in the dataset I, revealing the strength and direction of the relationship between multiple variables.

Variables	Elevation (m)	Slope (in °)	Gravel (%)	Very coarse sand (%)	Coarse sand (%)	Moderate sand (%)	Fine sand (%)	Very fine sand (%)	Silt and clay (%)
Elevation (m)	1	0.057	0.105	0.167	–0.222	–0.221	–0.159	–0.009	–0.004
Slope (in°)	0.057	1	0.293	0.000	–0.205	–0.258	–0.050	–0.145	–0.182
Gravel (%)	0.105	0.293	1	0.184	–0.450	–0.622	–0.259	–0.096	–0.178
Very coarse sand (%)	0.167	0.000	0.184	1	0.209	–0.291	–0.534	–0.288	0.139
Coarse sand (%)	–0.222	–0.205	–0.450	0.209	1	0.570	–0.326	–0.389	0.212
Moderate sand (%)	–0.221	–0.258	–0.622	–0.291	0.570	1	0.150	–0.222	0.012
Fine sand (%)	–0.159	–0.050	–0.259	–0.534	–0.326	0.150	1	0.682	–0.058
Very fine sand (%)	–0.009	–0.145	–0.096	–0.288	–0.389	–0.222	0.682	1	0.323
Silt and clay (%)	–0.004	–0.182	–0.178	0.139	0.212	0.012	–0.058	0.323	1

Table 2b. Correlation matrix table showing the correlation coefficients in the dataset II, revealing the strength and direction of the relationship between multiple variables..

Variables	Elevation (m)	Slope	Sand	Silt	Clay	Hm (%)	OC (%)	OM (%)	Soil pH	Salinity (µ/cm)
Elevation (m)	1	0.057	0.213	–0.118	–0.210	–0.032	–0.364	–0.371	0.028	–0.133
Slope (in °)	0.057	1	0.194	0.031	–0.314	–0.128	0.083	0.086	0.253	0.001
Sand (%)	0.213	0.194	1	–0.751	–0.814	–0.120	–0.048	–0.049	–0.119	–0.010
Silt (%)	–0.118	0.031	–0.751	1	0.229	0.106	0.280	0.281	0.137	0.040
Clay (%)	–0.210	–0.314	–0.814	0.229	1	0.085	–0.175	–0.174	0.055	–0.020
HM (%)	–0.032	–0.128	–0.120	0.106	0.085	1	0.439	0.437	0.075	–0.068
OC (%)	–0.364	0.083	–0.048	0.280	–0.175	0.439	1	1.000	0.125	–0.042
OM (%)	–0.371	0.086	–0.049	0.281	–0.174	0.437	1.000	1	0.126	–0.039
Soil pH	0.028	0.253	–0.119	0.137	0.055	0.075	0.125	0.126	1	–0.091
Salinity (µ/cm)	–0.133	0.001	–0.010	0.040	–0.020	–0.068	–0.042	–0.039	–0.091	1

Hygroscopic moisture (HM), Organic carbon (OC), Organic matter (OM)

Result and discussion: Multicollinearity

Correlation matrix

Multicollinearity affects only the specific independent variables that are deeply correlated to each other. It also affects the coefficient and p-values but does not influence the predictions, precisions of the prediction, and the goodness of fit (Frost, 2021). A significant correlation coefficient in the predictor variable's correlation matrix detects multicollinearity among the data set. The correlation coefficient between two variables will be near unity when multicollinearity is a severe issue between any two predictor variables. A correlation matrix has been used to identify a correlation or bi-variate relationship between two independent variables. Multicollinearity must exist when one independent variable is correlated with one or a linear combination of multiple independent variables. Multicollinearity is best to use in prediction. It can identify structures within the data and conduct operational discussions. This method is suitable for use to avoid numerical problems during mathematical counting.

According to the correlation matrix table (Table 2a, 2b, 2c), 70 combinations of bivariate relationships have been detected among the variables. Only one combination (slope and very coarse sand) has been identified which has no relation. Although it is not valid in reality, it has appeared here for some unknown reason. Despite that, 60% of combinations show very weak relationship (positive and negative) among them. Weak relationships (positive and negative) have been identified between 25.71% of the combinations and 8.58% of combinations are moderately correlated with each other. Fine sand and moderate sand have shown a strong positive relation. Silt and sand, gravel and moderate sand show a robust negative relation. A strong positive relationship between organic carbon and organic matter has been identified. The test result (Table. 2a, 2b, 2c) detects a strong linear relationship among the variables. High degree of interrelationship and inter-association among the variables in the data set is well explained by multicollinearity analysis.

A high negative correlation has been observed between gravel and moderate sand, very coarse sand and fine sand, moderate sand and very coarse sand, fine sand and very coarse sand, and between clay and sand (Table. 2a, 2b, 2c). The increased amount of gravel with a decrease in moderate sand and simultaneous increase in the amount of fine sand with a decreased amount of fine sand signify that this area is characterised by active denudation process to achieve peneplanation. Positive and negative correlations among the variables can also set a strong argument about the active operation of geomorphic processes and deep interaction between the different sub-systems of the earth like lithosphere, hydrosphere, pedosphere, atmosphere, and biosphere. Significant amount of gravel and sand indicates active mechanical weathering and extreme climate over this area. Correlations between elevation and gravel, elevation and sand, slope and gravel, slope and sand provide strong evidence supporting the active operation of the surface removal process. Elevation and gravel are positively correlated, indicating high gravel concentration in the high elevated areas. A high elevated area with steep slope does not support a thick soil profile which can support a deep-rooted plant system. Hence, grasses on a shallow soil profile cover this area, which cannot provide much protection from mechanical weathering.

Table 2c. The relationship and correlation degree using absolute r value and Evans' modified correlation degree (1996).

Very strong relation	Strong relation	Moderate relation	Weak relation	Very weak relation	No relation	Very weak relation	Weak relation	Moderate relation	Strong relation	Very strong relation
Negative						Positive				
(–) 1.00	(–) 0.8	(–) 0.6	(–) 0.4	(–) 0.2	0	(+) 0.2	(+) 0.4	(+) 0.6	(+) 0.8	(+) 1.0
0.81–1.0	0.61–0.8	0.41–0.6	0.21–0.4	0.01–0.2		0.01–0.2	0.21–0.4	0.41–0.6	0.61–0.8	0.81–1.0
	Si vs Sa G vs MS	G vs MS G vs CS VCS vs FS	E vs CS E vs MS E vs Cl E vs OC E vs OM Sl vs C G vs FS CS vs FS CS vs VFS MS vs VFS	E vs FS E vs VFS E vs S&C Cl vs OC Cl vs OM Cl vs S HM vs S OC vs S OM vs S pH vs S Sl vs HM Sl vs CS Sl vs FS Sl vs VFS Sl vs S & C G vs VFS G vs S & C VCS vs MS VCS vs VFS FS vs S & C	Sl vs VCS	E vs Sl E vs G E vs VCS E vs pH Si vs HM Si vs pH Si vs S OC vs pH HM vs pH OM vs pH Sl vs Sa Sl vs Si Sl vs OM Sl vs OC Sl vs S Cl vs HM Cl vs pH G vs VCS VCS vs S & C VCS vs CS MS vs FS MS vs S & C	E vs Sa Si vs Cl Si vs OC Si vs OM Sl vs G Sl vs pH CS vs S & C VFS vs S & C	HM vs OC HM vs OM CS vs MS	FS vs VFS	OC vs OM
Total (∑) (70)/ Percentage (100)										
00	02/2.86%	03/4.29%	10/14.29%	20/28.57%	01/1.43%	22/31.43%	08/11.43%	03/4.29%	01/1.43%	00

Slope (Sl), Gravel (G), Very coarse sand (VCS), Sand (Sa), pH, Coarse sand (CS), Moderate sand (MS), Fine sand (FS), Very fine sand (VFS), Silt and clay (S& C), Silt (Si), Clay (C), Hygroscopic moisture (HM), Organic carbon (OC), Organic matter (OM), Salinity (S).

Status of a linear relationship

If the R^2 value is 1, then it is evident that there is a robust linear relationship between the dependent variable (y) of the data set and the explanatory variable (x). A high R^2 value reveals that the independent variable thoroughly explains all variations of the dependent variable. A strong linear relationship is present in the case of gravel, coarse sand, moderate sand and fine sand (Table 4a), elevation, sand, silt, clay, organic carbon, and organic matter (Table. 4b). Moderate linear relationship has been identified in very coarse sand, silt and clay (Table.4a), slope and hygroscopic moisture (Tabler.4b). Elevation shows a strong linear relationship (Table 4b), indicating elevation change is closely related to the change of other variables in the data set. A variable with strong linear relationships can be considered as the most critical and responsible variable. They are also strongly associated with other variables. A strong linear relationship is deeply associated with a correlation coefficient of –1 or +1 and a considerably high R^2 value.

Table 3a. Multicollinearity statistics. (dataset I)

Statistic	Elevation	Slope	Gravel	Very coarse sand	Coarse sand	Moderate sand	Fine sand	Very fine sand	Silt and clay
R^2	0.144	0.174	0.510	0.406	0.623	0.679	0.726	0.736	0.374
Tolerance	0.856	0.826	0.490	0.594	0.377	0.321	0.274	0.264	0.626
VIF	1.2	1.2	2.0	1.7	2.7	3.1	3.6	3.8	1.6

Table 3b. Multicollinearity statistics. (dataset II)

Statistics	Elevation	Slope	Sand	Silt	Clay	Hygroscopic moisture (%)	Organic Carbon (%)	Organic matter (%)	Soil pH	Salinity
R^2	0.609	0.268	1.000	1.000	1.000	0.290	1.000	1.000	0.126	0.078
Tolerance	0.391	0.732	0.000	0.000	0.000	0.710	0.000	0.000	0.874	0.922
VIF	2.6	1.4	188519.8	66990.0	86697.2	1.4	14514.1	14620.9	1.1	1.1

When the tolerance value is less than 1, then collinearity exists.

Table 4a. Status of linear relationship (dataset I).

Statistic	R^2	Percentage	State of relation	Remarks
Elevation	0.144	14.4	Weak linear relation	0-25%= Weak linear relation
Slope	0.174	17.4	Weak linear relation	25-50%= Moderate linear relation
Gravel	0.510	51	Strong linear relation	50-75%= Strong linear relation
Very Coarse sand	0.406	40.6	Moderate linear relation	75-100%= Robust linear relation
Coarse sand	0.623	62.3	Strong linear relation	
Moderate Sand	0.679	67.9	Strong linear relation	
Fine sand	0.726	72.6	Strong linear relation	
Very Fine sand	0.736	73.6	Strong linear relation	
Silt and Clay	0.374	37.4	Moderate linear relation	

Table 4a. Status of linear relationship (dataset II).

Statistic	R^2	Percentage	State of relation	Remarks
Elevation	0.609	60.9	Strong linear relation	0-25%= Weak linear relation
Slope	0.268	26.8	Moderate linear relation	25-50%= Moderate linear relation
Sand	1.000	100	Robust linear relation	50-75%= Strong linear relation
Silt	1.000	100	Robust linear relation	75-100%= Robust linear relation
Clay	1.000	100	Robust linear relation	
Hygroscopic moisture	0.290	29	Moderate linear relation	
Organic Carbon	1.000	100	Robust linear relation	
Organic matter	1.000	100	Robust linear relation	
Soil pH	0.126	12.6	Weak linear relation	
Soil salinity	0.078	7.8	Weak linear relation	

Tolerance

Multicollinearity is well understood by tolerance and Variance Inflation Factor (VIF) values of the analysis result. Tolerance is the amount of variability in one independent variable that is not explained by the other independent variables (Daoud, 2017). The tolerance is $1\text{-}R^2$. Tolerance can be considered as an essential standard for filtering of variables. When a particular variable is associated with a tolerance less than a fixed threshold, it is unfair to enter the model, only because its contribution is very insignificant, which might cause some numerical problems. The tolerance value indicates that the variable under consideration is a nearly ideal linear amalgamation of the independent variable. If the variable is associated with a linear relationship, it should have a small tolerance value. Coarse sand, moderate sand, fine sand, and very fine sand (Table 3a) from the dataset I and elevation (Table 3b) from the dataset II are associated with linear relationship. According to tolerance value, elevation, slope, gravel, very coarse sand, silt and clay (Table 3a) from the dataset I and slope, hygroscopic moisture, pH, and salinity (Table 3b) from the dataset II are not associated with a strong linear relationship. However, further investigation is needed when the tolerance value is less than 0.1. A low tolerance value indicates a large standard error toward less significant studies. Tolerance values less than 0.1 indicate collinearity (Daoud, 2017). Sometimes, it is also considered that if the value of tolerance is less than .2 (similarity VIF <20), multicollinearity is present among the data set. As per this concern, collinearity must be present in sand, silt, clay, organic carbon, and organic matter (Table 3b).

Variation Inflation Factor

Variance Inflation Factor (VIF) is equal to the inverse of tolerance. When correlation exists among predictors, the standard error of the predictor's coefficient will increase, consequently inflating the variance of the predictor's coefficients (Daoud,2017). The VIF is a tool to measure and quantify how much the variance is inflated. VIF measures the impact of collinearity among the variables in the regression data model. VIF is used to identify one independent variable's correlation with another variable's batch. Other variables easily explain the degree of variation in a single variable, and VIF can

measure that. VIF value always remains more significant than one because a VIF value of 1 or < 1 indicates no correlation among the variables. A high VIF value stands for high multicollinearity and instability among the variables. VIF value greater than 10 is accompanied by a more significant correlation among the variables. Strong multicollinearity increases the variance of a regression coefficient; the increase in the variance also increases the standard error of the regression coefficient (because the standard error is the square root of the variance), and the increase in the standard error leads to a wide 95% confidence interval of the regression coefficient (Kim, 2019).

Here, several multicollinearities have been detected with the help of tolerance and its reciprocal (VIF). When multicollinearity is present, then it is clear that the coefficient is unstable; it changes to a large extent. When multicollinearity is strongly present in the data set, it is also impossible to test the regression coefficient individually because of inflated standard errors. So, it does not seem easy to convey significant independent variables here. When multicollinearity is present in the data set, it seems complicated to intercept the coefficient of the data set. It can reduce the real power of the regression model in case of detection of statistically significant independent variables. The degree of multicollinearity is a big issue when the question arises about how to solve the problem after its detection in the data set. Correlated specific independent variables must be affected by multicollinearity in the data set. According to the VIF table (Table 5a, 5b), each variable is correlated (Fig. 4). Strong multicollinearity exists in sand, silt, clay, Organic matter, and organic carbon (Table 5b).

Table 5a. Variation Inflation factor (VIF) summary for dataset I.

VIF value	State of correlation	Variables
<1	No correlation	Elevation, Slope, Gravel, Very coarse sand, Coarse sand, Moderate sand, Fine sand, Very fine sand, Silt and Clay
1–5	Minimum correlation	
5–10	Moderate Correlation	
>10	High correlation	

Table 5b. Variation Inflation factor (VIF) summary for dataset II.

VIF value	State of correlation	Variables
<1	No correlation	
1–5	Minimum correlation	Elevation, Slope, Hygroscopic moisture, Soil pH, Conductivity
5–10	Moderate Correlation	
>10	High correlation	Sand, Silt, Clay, Organic carbon, Organic matter

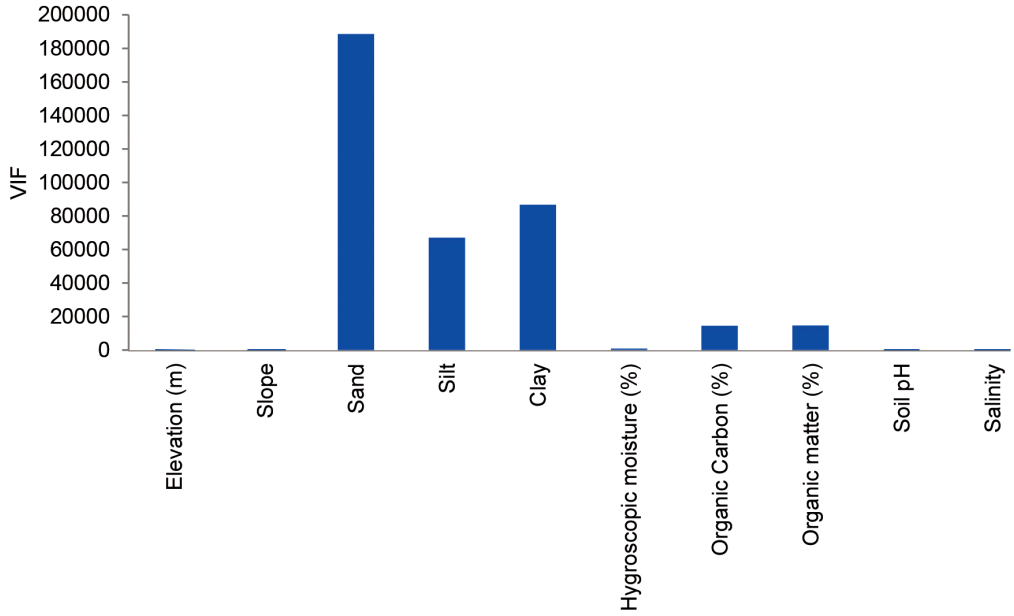

Figure 4. Variance Inflation Factor (VIF) of data set II. High VIF depicts the presence of multicollinearity.

Here, it can be added that the purpose of VIF itself in showing whether the predictors are correlated, $\sqrt{VIF}$ signify how much larger the standard error is, for example if VIF is greater than 9 this means that the standard error for the coefficient of that predictor is 3 times as large as it would be if that predictor is uncorrelated with other predictors

Conclusion

Here, independent variables are highly correlated to each other, which indicates the presence of multicollinearity in the dataset. Multicollinearity can cause issues in regression analysis. The severity of multicollinearity has been quantified through the VIF. According to tolerance and VIF values, multicollinearity strongly exists in sand, silt, clay, organic matter, and organic carbon variables. The strong linear relationship between variables in the data set indicates their influence on each other. When variables are highly correlated, some of them show a strong linear relationship. This high R^2 and VIF values suggest that collinearity may be an issue. Tolerance and VIF are important diagnostic factors for detecting multicollinearity in the data set. If collinearity is detected in the regression output, any explanation of the relationship should be avoided until the problem is resolved.

Further research is needed to address highly correlated variables. It may be necessary to drop one of the variables causing multicollinearity, especially if they are conceptually similar. Furthermore, in future research, Principal Component Analysis can reduce dimensionality and eliminate multicollinearity by creating new uncorrelated variables. Although this may make interpretation more challenging. Multicollinearity can be resolved by combining the highly correlated variables

through principal component analysis or omitting a variable from the analysis that is associated strongly with other variable(s), and it is one of the major problems that should be resolved before starting the process of modelling the data (Daoud, 2017). The impact of multicollinearity may be addressed using regularisation techniques like Ridge regression or Lasso regression. Lasso and Ridge regressions are advanced forms of regression analysis that can handle multicollinearity (Frost, 2021). The problem of multicollinearity may be reduced by changing the approach to sample collection. Further research can be done to increase the number of samples taken from a variety of locations within the study area. A substantial number of samples can solve the multicollinearity problem by varying the data character.

References

Adeboye, N.O., Fagoyinbo, I.S. and Olatayo, T.O. (2014) Estimation of the effect of multicollinearity on the standard error for regression coefficients, *IOSR Journal of Mathematics*, 10(4): 16–20.

Aerts, R. and Chapin, F.S. (1999) The mineral nutrition of wild plants revisited: A re-evaluation of processes and patterns. In Fitter, A.H. and Raffaelli, D.G. (eds) *Advances in Ecological Research*, Academic Press, London, 30: 1–67.

Alesina, A., Dollar, D. (2000) Who gives foreign aid to whom and why? *Journal of Economic Growth*, 5(1): 33–63.

Alvi, E. and Senbeta, A. (2012) Foreign aid: Good for investment, bad for productivity, *Oxford Development Studies*, 40(2): 139–161.

Ando, A. and Modigliani, F. (1963) The "Life cycle" hypothesis of saving: Aggregate implications and tests, *The American Economic Review*, 53(1): 55–84.

Avitabile, V., Baccini, A., Friedl, M.A. and Schmullius, C. (2012) Capabilities and limitations of Landsat and land cover data for aboveground woody biomass estimation of Uganda, *Remote Sensing of Environment*, 117: 366–380.

Ayanu, Y.Z., Conrad, C., Nauss, T., Wegmann, M. and Koellner, T. (2012) Quantifying and mapping ecosystem services supplies and demands: A review of remote sensing applications, *Environmental Science & Technology*, 46: 8529–8541.

Bager, A., Roman, M., Odah, M. and Mohammed, B. (2017) Addressing multicollinearity in regression models: A ridge regression application, *Journal of Social and Economic Statistics*, 16(1): 30–45.

Baharumshah, A.Z., Slesman, L. and Devadason, E.S. (2017) Types of foreign capital inflows and economic growth: New evidence on the role of financial markets, *Journal of International Development*, 29(6): 768–789.

Baidoo, S.T., Boateng, E. and Amponsah, M. (2018) Understanding the determinants of saving in Ghana: Does financial literacy matter? *Journal of International Development*, 30(5): 886–903.

Bandyopadhyay, S., Kar, N.S., Das, S. and Sen, J. (2014) River systems and water resources of West Bengal: A review, *Transactions of the Institute of Indian Geographers*, 36(2): 261–278.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, New York: 292p.

Bhattacharya, B.K., Chakraborty, B.R., Sen, N.N., Mukherji, S., Ray, P., Sengupta, S., Sengupta, K.S. and Maity, T. (1985) *West Bengal District Gazetteers: Puruliya*. Government of West Bengal, Kolkata: 380p.

Box, E.O., Holben, B.N. and Kalb, V. (1989) Accuracy of the AVHRR vegetation index as a predictor of biomass, primary productivity, and net CO2 flux, *Vegetatio*, 80(1): 71–89.

Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A.B., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D. and Rodrigues, A.S.L. (2006) Global biodiversity conservation priorities, *Science*, 313(5783): 58–61.

Mela, C.F. and Kopalle, P.K. (2002) The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations, *Applied Economics*, 34(6): 667–677.

Coenders, G. and Saez, M. (2000) Collinearity, heteroscedasticity and outlier diagnostics in regression: Do they always offer what they claim?

In Ferligoj, A. and Mrvar, A. (eds) *New Approaches in Applied Statistics*, FDV, Ljubljana: 79–98.

Daniel, S., Onyinah, P.O., Baidoo, S.T. and Ayesu, E.K. (2021) Empirical determinants of saving habits among commercial drivers in Ghana, *Journal of African Business*, 22(1): 106–125.

Daoud, J.I. (2017) Multicollinearity and regression analysis, IOP Conference Series: *Journal of Physics*, Conference Series, 949: 012009.

Das, T., Sarkar, D., Chattopadhyay, T., Dutta, D., Singh, D.S., Mukhopadhyay, S., Nayak, D.C. and Banerjee, T. (2010) *Soils of Purulia District, West Bengal for Optimizing Land Use*. NBSS Publication No. 599, NBSS and LUP, Nagpur: 240p.

Dasgupta, P. (2015) Rock characteristics and susceptibility to weathering: A study in the metamorphic terrain of Ajodhya Hill, West Bengal, *Journal of Indian Geomorphology*, 3: 27–47.

De, N.K. (1972) Measuring soil and landform characteristics of parts of Banka Basin, Burdwan. In *Proceedings of Symposium on Geomorphology and Geohydrology*, IIT Kharagpur: 46–58.

De, N.K. (1984a) Pedogeomorphology: A concept in earth science, Burdwan University *Journal of Science*, 1: 18–25.

De, N.K. (1984b) *Measuring Land Potentials in Developing Countries*. University of Burdwan, Burdwan: 210p.

De, N.K. (1987) *Application of Geomorphology in Soil Survey*. In Dixit, K.R. (ed) Exploration in the Tropics: Felicitation Volume, Pune: 123–138.

De, N.K. and Chatterjee, S. (2009) Pedogeomorphological model and land evolution. In Singh, S. (ed) *Geomorphology of India: Felicitation Volume for Prof. Savindra Singh*, Prayag Pustak Bhavan, Allahabad: 547–563.

Frost, J. (2021) *Multicollinearity in regression analysis: Problems, detection, and solutions, Statistics* By Jim. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ (retrieved on 2021-05-05).

Ghosh, A.K. (2013) *Status of Environment in West Bengal*. Society for Environment and Development, Kolkata: 310p.

GSI: Geological Survey of India (1999) *Geology and Mineral Resources of the States of India*, Part 1: West Bengal, Vol. 30. Geological Survey of India, Kolkata: 250p.

Jensen, D.R. and Ramirez, D.E. (2012) Variance inflation in regression, *Advances in Decision Sciences*, 2012: 1–15.

Kim, J.H. (2019) Multicollinearity and misleading statistical results, *Korean Journal of Anaesthesiology*, 72(6): 558–569.

Kroll, C.N. and Song, P. (2013) Impact of multicollinearity on small sample hydrologic regression models, *Water Resources Research*, 49(6): 3756–3769.

Mahapatra, G.B. (2008) *A Text Book of Geology*. CBS Publishers and Distributors, Delhi: 672p.

Mercy, O., Mutanga, O., Odindi, J. and Abdel-Rahman, E.M. (2016) Application of topo-edaphic factors and remotely sensed vegetation indices to enhance biomass estimation in a heterogeneous landscape in the Eastern Arc Mountains of Tanzania, *Geocarto International*, 31(1): 1–21.

Osborne, J.W. and Waters, E. (2002) Four assumptions of multiple regressions that researchers should always test, Practical Assessment, *Research and Evaluation*, 8(2): 1–5.

Richter, D.D., Eppes, M., Austin, J.C., Bacon, A.R., Billings, S.A., Brecheisen, Z., Ferguson, T.A., Markewitz, D., Pachon, J., Schroeder, P.A. and Wade, A.M. (2019) Soil production and the soil geomorphology legacy of Grove Karl Gilbert, *Soil Science Society of America Journal*, 84(1): 1–20.

Sandra, B. (2002) Measuring carbon in forests: Current status and future challenges, *Environmental Pollution*, 116(3): 363–372.

Sarkar, A. (2019a) Soil geomorphology of Garpanchakot Hill area and its influence on land use and land cover, *Journal of Geoscience and Environment Protection*, 7: 108–135.

Sarkar, A. (2019b) Soil geomorphological model and its relation to land use planning: A case study of Biharinath Hill, Bankura District, West Bengal, *Indian Journal of Landscape Systems and Ecological Studies*, 42(1): 5–19.

Sarkar, A. (2019c) Determination of principal soil geomorphological parameters of Garpanchakot Hill area, Purulia District, West Bengal, India using principal component analysis, *Hill Geographer*, 35(2): 49–58.

Sarkar, A. (2021) Analysis of soil geomorphological parameters using data mining techniques in Ajodhya Hills area, *West Bengal, Indian Journal of Spatial Science*, 12(1): 174–182.

Sarkar, A. and Das, P. (2018) Pedogeomorphology of the plateau fringe region of Biharinath Hill, Bankura District of West Bengal and its influence on land use and landcover, *International Journal of Research and Analytical Reviews*, 5(2): 488–501.

Shalabh (2019) *Regression Analysis*, Chapter 9: Multicollinearity, IIT Kanpur. http://home.iitk.ac.in/~shalab/regression/Chapter9-Regression-Multicollinearity.pdf (retrieved on 2021-03-17).

Stephanie, G. (2015) *Multicollinearity: Definition, causes, examples, Statistics* HowTo.com. https://www.statisticshowto.com/multicollinearity/ (retrieved on 2021-05-05).

Zinck, J.A., Metternicht, G., Bocco, G. and Del Valle, H.F. (2016) *Geopedology: An Integration of Geomorphology and Pedology for Soil and Landscape Studies*. Springer, Berlin: 450p.

Date received: 26 March 2024
Date accepted after revision: 11 October 2024