Journal of Agricultural Science and Technology (JAST), 29(6)
In Press, Pre-Proof Version

Soil Quality Assessment and Yield Prediction for Sustainable Banana Production

B. R. Jagadeesh[1]*, T. Bhagyalakshmi[1], R. Suma[1], P. S. Fathima[2], and P. Praveen[3]

Abstract

Soil quality strongly influences banana yields, but few field studies have linked soil properties to actual farmer-managed production, particularly in India's diverse banana regions. This work assessed the physical, chemical and micronutrient status of soils under banana in five villages of Nanjangud taluk, Mysore district, Karnataka, and related these traits to recorded yields. Samples and yield data were gathered from 150 geo-referenced farmers' fields. The mean observed banana yield was 46.53 ± 10.14 t ha^{-1}, ranging from 30.00 to 77.60 t ha^{-1}. Principal Component Analysis singled out iron, porosity, pH, available nitrogen, available phosphorus and zinc as the main soil-quality indicators, these were combined into a Soil Quality Index (SQI). Village-mean SQI ranged from 0.401 in Halepura to 0.523 in Kanenuru, with Huskuru and Karya at 0.490 and Hosaveedu at 0.479. Across individual fields, 22 % of soils ranked low, 72 % moderate and 6 % high in quality. The Random Forest model showed promising predictive performance, with a squared correlation of 0.765 between observed and predicted yields, a test-set R^2 of 0.700, an RMSE of 5.32 t ha^{-1} and an MAE of 4.30 t ha^{-1}.Banana variety proved the strongest driver, followed by zinc, iron, latitude, organic carbon, available potassium and nitrogen. By linking soil-quality indices with machine-learning yield models, we can identify which soil and crop factors most strongly influence banana yield. This, in turn, provides a basis for developing site-specific nutrient management strategies in the region.

Keywords: Principal Component Analysis, Random Forest Model, Soil Quality Index.

1. INTRODUCTION

Banana (*Musa* spp.) is a major fruit crop in tropical and subtropical regions, supporting food security, nutrition and smallholder livelihoods. Global production reached approximately 139 million tonnes in 2023, with India among the leading producers (FAO, 2023; FAO, 2025). Yield variability across regions reflects differences in soil fertility, nutrient availability, cultivar, water supply and crop management (Wairegi et al., 2010; Wairegi and van Asten, 2010).

[1] Department of Soil Science and Agricultural Chemistry, College of Agriculture, V.C. Farm, Mandya, University of Agricultural Sciences, Bangalore, Karnataka, India.
[2] College of Agriculture, V.C. Farm, Mandya, University of Agricultural Sciences, Mandya, Karnataka, India.
[3] Department of Agricultural Engineering, College of Agriculture, V.C. Farm, Mandya, University of Agricultural Sciences, Bangalore, Karnataka, India.
*Corresponding author; e-mail: jagushreeram@gmail.com

Conventional soil fertility assessments focus mainly on individual nutrients, whereas crop productivity depends on interactions among soil physical, chemical and biological properties. Soil quality reflects the capacity of soil to sustain plant growth and ecosystem functions (Bünemann et al., 2018). Biological properties contribute to nutrient cycling and other soil functions, although they are dynamic and sensitive to environmental conditions and management practices (Maurya et al., 2020). A Soil Quality Index (SQI) integrates selected soil indicators into a composite measure of soil condition. Principal Component Analysis (PCA) is commonly used to identify a representative Minimum Data Set (MDS) from a larger set of indicators while reducing redundancy among correlated variables (Abass et al., 2023). The selection of indicators depends on the objectives of the assessment, local soil characteristics and environmental conditions. Consequently, an MDS developed for a particular study area should be interpreted within its geographical and agricultural context, and its applicability to other regions requires further evaluation. Soil quality is an important component of banana productivity, but yield is also influenced by cultivar, planting material, irrigation and nutrient management. Recent studies have highlighted the roles of soil nutrients, environmental conditions and farm practices in crop-yield variation (Huang et al., 2025; Zhang et al., 2025; Ndayambaje and Ngango, 2025). Random Forest (RF) modelling can capture nonlinear relationships and estimate the relative predictive importance of different variables. Recent research has applied machine-learning approaches to banana-yield prediction using soil properties, crop characteristics, environmental variables and remote-sensing data (Yánez-Cajo et al., 2025). However, field-level assessments integrating soil quality with crop and management factors to predict banana yield under farmer-managed conditions remain limited.Integrating PCA-based SQI development with RF modelling provides an approach for assessing soil-quality variation and identifying variables associated with banana yield. Accordingly, the present study was conducted in five banana-growing villages of Nanjangud taluk, Karnataka, to assess soil quality and examine its relationship with banana productivity. The specific objectives were to:

1. Assess soil quality in banana fields using a PCA-based SQI.
2. Identify key soil, crop and management factors associated with banana yield.
3. Develop and evaluate an RF model for banana-yield prediction under field conditions.

2. MATERIALS AND METHODS

2.1 Study Area

The study was conducted in five banana-growing villages Huskuru, Hosaveedu, Kanenuru, Karya and Halepura in Nanjangud taluk, Mysuru district, Karnataka, India (Figure 1a). Variations in soil properties, nutrient status and crop management provided a basis for assessing soil quality and its association with banana yield. Village-wise banana cultivation area is presented in Table 1. Sampling-site coordinates were recorded using GPS, while information on banana variety, planting material, irrigation and yield was collected through farmer interviews and field surveys.

Table 1. Selected banana-growing villages and cultivated area in the study region.

Hobli	Village	Area (Hectares)
Hullahalli	Huskuru	97.14
	Hosaveedu	55.22
	Kanenuru	59.61
	Karya	51.01
Doddakowlande	Halepura	83.72

Source: Crop survey data 2023, Department of Horticulture.

2.2. Sampling design and banana yield measurement

A total of 150 banana-growing fields were selected from five villages of Nanjangud taluk, Mysuru district, Karnataka, namely Halepura, Hosaveedu, Huskuru, Kanenuru and Karya, with 30 fields per village selected through purposive sampling based on field suitability and availability. Banana yield data were collected through farmer interviews and expressed in tonnes per hectare (t ha^{-1}). Soil quality and yield data were used to examine the relationship between soil quality and banana productivity across the study area.

2.3 Soil Sampling and Laboratory Analysis

A total of 150 geo-referenced surface soil samples were collected from banana fields across five villages at a depth of 0–30 cm. The samples were shade-dried, gently crushed and sieved before laboratory analysis. Soil physical properties, including bulk density, particle density, porosity and maximum water-holding capacity (MWHC), were determined using standard methods. Soil pH, electrical conductivity (EC), organic carbon (OC), available nitrogen (N), phosphorus (P_2O_5) and potassium (K_2O) were analysed following Jackson (1973). Available iron (Fe), manganese (Mn), zinc (Zn) and copper (Cu) were extracted using the DTPA method and quantified according to Lindsay and Norvell (1978). Information on banana variety, planting material, irrigation practices and yield was collected for each field through farmer

interviews and direct observation. These data were integrated with soil properties to examine factors associated with banana yield.

2.4 Development of the Soil Quality Index

A Soil Quality Index (SQI) was developed to integrate multiple soil properties into a single measure of soil condition, reflecting the soil's capacity to support plant growth and ecosystem functions (Bünemann et al., 2018). Principal Component Analysis (PCA) was used to identify the main sources of variation in the data and to select a representative Minimum Data Set (MDS), thereby reducing redundancy among correlated variables (Karlen et al., 2008). Suitability of the dataset for multivariate analysis was first confirmed with Bartlett's test of sphericity. Principal components with eigenvalues greater than 1 were retained, and variables with high factor loadings on these components were chosen for the MDS. When multiple indicators were associated with the same principal component, their factor loadings and inter-indicator correlations were considered in selecting the final indicators to minimize redundancy. The soil quality index (SQI) was calculated using a weighted-additive approach based on six minimum data set indicators: Fe, soil porosity, pH, available N, available P_2O_5 and Zn. Indicator values were transformed into unitless scores ranging from 0 to 1. A linear more-is-better scoring function was applied to Fe, porosity, available N, available P_2O_5 and Zn:

$S_i = (x_i - x_{i,min}) / (x_{i,max} - x_{i,min})$ where S_i is the score of indicator i, x_i is its observed value, and $x_{i,min}$ and $x_{i,max}$ are the minimum and maximum observed values, respectively. Soil pH was scored using a piecewise function, assigning a score of 1.0 within the pH range of 6.5–7.5 and progressively lower scores outside this range:

$$S_{ph} = \begin{cases} pH/6.5, & pH < 6.5 \\ 1, & 6.5 \le pH \le 7.5 \\ 1/[1 + (2/15)(pH - 7.5)], & pH > 7.5 \end{cases}$$

The scored indicators were aggregated using the weighted-additive equation:

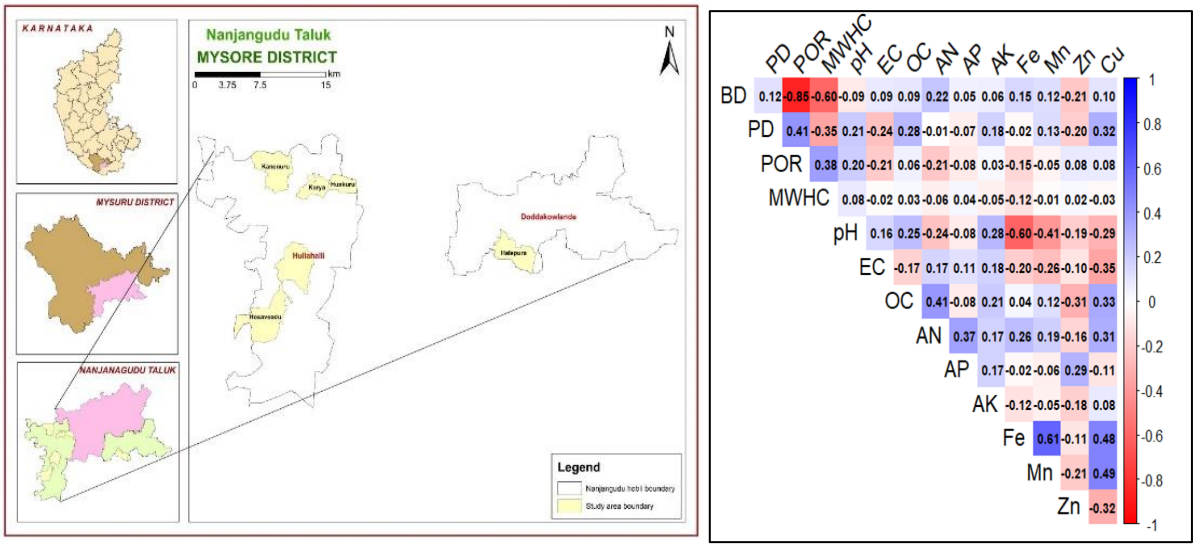
$$SQI = \sum_{i=1}^{6} w_i S_i$$

where w_i is the weight assigned to each indicator and S_i is its corresponding score. The weights used were 0.294 for Fe, 0.225 for porosity, 0.207 for pH, 0.076 for available N, 0.076 for available P_2O_5 and 0.121 for Zn. The weighted contributions were summed to obtain the SQI for each sampled field. Soil quality was classified into five categories following Abass et al., 2023: very low (SQI ≤ 0.20), low (0.20 < SQI ≤ 0.40), moderate (0.40 < SQI ≤ 0.60), high (0.60 < SQI ≤ 0.80) and very high (0.80 < SQI ≤ 1.00). Each sampling location was assigned

to a category based on its calculated SQI, and the number and percentage of locations in each category were determined to describe soil-quality distribution across the study area. The same classification criteria were applied to village-level mean SQI values to compare soil-quality status among the five villages.

2.5 Random Forest-Based Yield Prediction

Banana yield was predicted using a Random Forest (RF) regression model incorporating soil, nutrient, crop and management variables. RF is an ensemble machine-learning technique capable of capturing complex, nonlinear relationships among predictors (Breiman, 2001). The predictors included soil physical and chemical properties, micronutrient status, banana variety, planting material and irrigation method.The dataset was randomly divided into training (122 observations) and testing (28 observations) subsets, using a fixed random seed of 123 to ensure reproducibility, and the model was constructed with 500 decision trees. Model performance was evaluated using the squared correlation (r^2) between observed and predicted yields, the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) based on the test dataset. The relative predictive importance of individual predictors was assessed using the percentage increase in mean squared error (%IncMSE). Similar RF approaches have been applied previously to banana-yield prediction using crop, environmental, soil and remote-sensing variables (Yánez-Cajo et al., 2025).



(a) (b)

Figure 1. (a) Location of the study area in Karnataka, India; (b) correlation heatmap of soil properties in banana-growing fields.

The testing subset was used to evaluate predictions for observations excluded from model training. However, independent external validation using a separate dataset was not performed. Thus, overfitting cannot be ruled out, and further validation across locations, seasons and cultivars is needed to assess the model's generalizability.

2.6 Statistical Analysis

All statistical analyses were performed using R version 4.5.1 in RStudio. Descriptive statistics (mean, standard deviation, minimum, maximum and range) were calculated for soil properties. Bartlett's test and principal component analysis (PCA) were used for MDS selection and SQI construction.The normality of ANOVA residuals was assessed using the Shapiro–Wilk test, and homogeneity of variance among villages was assessed using Levene's test. When homogeneity of variance was violated, Welch's ANOVA was used to compare mean soil quality index (SQI) values among villages. Where significant differences were detected, the Games–Howell post-hoc test was used for pairwise comparisons. Statistical significance was assessed at $p < 0.05$. RF model performance was evaluated using r^2, R^2, RMSE and MAE. Predictor importance was assessed using %IncMSE.

3. RESULTS

3.1 Soil Properties of Banana-Growing Fields

The 150 soil samples showed substantial variation in physical, chemical and micronutrient properties across the five villages (Table 2). Bulk density ranged from 0.90 to 1.73 Mg m^{-3}, porosity from 31.79 to 65.33%, and maximum water-holding capacity (MWHC) from 20.96 to 80.19%. Soil pH ranged from 5.20 to 8.80 (mean 7.36), although most samples were neutral to alkaline. Electrical conductivity (EC) ranged from 0.05 to 1.67 dS m^{-1}, while organic carbon (OC) ranged from 0.12 to 0.87% (mean 0.52%). Available nitrogen, phosphorus and potassium ranged from 112.90 to 399.08, 4.10 to 112.34 and 122.17 to 464.62 kg ha^{-1}, respectively. Available zinc ranged from 0.03 to 2.25 mg kg^{-1} (mean 0.64 mg kg^{-1}). Overall, the results indicate considerable variability in soil fertility and micronutrient status among the sampled banana fields. However, the geological origins and parent materials of the sampled soils were not specifically investigated therefore, the observed variation cannot be attributed to differences in geological origin based on the present data.

3.2 Principal Component Analysis and Minimum Data Set

Five principal components with eigenvalues greater than one together explained 73.22 % of the total variance in the soil dataset (Table 3). Factor loadings on these components identified iron, porosity, pH, available nitrogen, available phosphorus and zinc as the most influential indicators. These six variables captured the main sources of physical, chemical and micronutrient variation and were retained as the Minimum Data Set (MDS) for constructing the Soil Quality Index (SQI).

Table 2. Descriptive Statistics of Soil Surface Physical Properties (n = 150).

Soil parameter	Mean ± SD	Median	Min	Max	Range	Skewness	Kurtosis	SE
Bulk Density (Mg m^{-3})	1.32 ± 0.12	1.32	0.9	1.73	0.83	0.04	0.81	0.01
Particle Density (Mg m^{-3})	2.57 ± 0.13	2.59	2.08	2.79	0.71	-1.23	2.13	0.01
Porosity (%)	48.56 ± 5.14	49.03	31.79	65.33	33.53	-0.19	0.44	0.42
MWHC (%)	45.55 ± 9.36	45.36	20.96	80.19	59.24	0.24	1.01	0.76
pH	7.36 ± 0.84	7.4	5.2	8.8	3.6	-0.31	-0.68	0.07
EC (dS m^{-1})	0.20 ± 0.24	0.14	0.04	1.67	1.63	3.82	16.63	0.02
Organic Carbon (%)	0.52 ± 0.20	0.52	0.12	0.87	0.75	-0.14	-0.83	0.02
Available N (kg ha^{-1})	204.28 ± 52.42	197.57	112.9	399.08	286.19	0.61	0.42	4.28
Available P_2O_5 (kg ha^{-1})	23.43 ± 20.07	16.16	4.1	112.34	108.23	2.73	7.8	1.64
Available K_2O (kg ha^{-1})	290.07 ± 90.14	297.23	122.17	464.62	342.45	-0.09	-0.97	7.36
Fe (mg kg^{-1})	4.77 ± 2.93	5.03	0.66	13.39	12.73	0.65	0.07	0.24
Mn (mg kg^{-1})	3.53 ± 2.75	2.73	0.25	13.05	12.8	1.2	1.41	0.22
Zn (mg kg^{-1})	0.64 ± 0.36	0.57	0.03	2.25	2.22	1.6	3.66	0.03
Cu (mg kg^{-1})	1.38 ± 0.92	1.18	0.09	2.94	2.85	0.15	-1.54	0.07

3.3 Soil Quality Index and Spatial Variation

The mean Soil Quality Index (SQI) differed significantly among the five villages (Table 4). Kanenuru recorded the highest mean SQI (0.523), followed by Huskuru and Karya (0.490 each), Hosaveedu (0.479) and Halepura (0.401).

Table 3. Principal Component Analysis: Explained Variance.

Principal Component	Eigenvalue	Variance Explained (%)	Cumulative Variance (%)
PC1	3.016	21.54	21.54
PC2	2.315	16.53	38.07
PC3	2.122	15.16	53.23
PC4	1.559	11.13	64.37
PC5	1.239	8.85	73.22
PC6	0.887	6.34	79.55
PC7	0.698	4.98	84.54
PC8	0.556	3.97	88.51
PC9	0.481	3.44	91.95
PC10	0.354	2.53	94.48
PC11	0.32	2.28	96.76
PC12	0.249	1.78	98.54
PC13	0.203	1.45	99.99
PC14	0.002	0.01	100

The Shapiro–Wilk test indicated a significant departure from normality of the ANOVA residuals ($W = 0.97825$, $p = 0.01765$), while Levene's test indicated significant heterogeneity of variance among villages ($F = 3.5186$, $p = 0.00898$). Accordingly, Welch's ANOVA revealed significant differences in mean SQI among villages ($F = 19.653$, $df_1 = 4$, $df_2 = 71.85$, $p < 0.001$; Table 5). Games–Howell post-hoc comparisons indicated that Halepura had a significantly lower mean SQI than each of the other four villages, whereas differences among Hosaveedu, Huskuru, Karya and Kanenuru were not statistically significant ($p > 0.05$; Table 6). Although all five villages belonged to the moderate soil-quality class, the significant differences in SQI indicated spatial variation in soil quality within the study area. The radar chart (Figure 2) further illustrated differences in indicator-level scores among the villages.

Table 4. Mean Soil Quality Index and statistical grouping of the five villages.

Village	Mean SQI	Soil quality class	Statistical group
Halepura	0.401	Moderate	A
Hosaveedu	0.479	Moderate	B
Huskuru	0.490	Moderate	B
Karya	0.490	Moderate	B
Kanenuru	0.523	Moderate	B

Note: Means sharing the same letter are not significantly different according to the Games–Howell post-hoc test at the 5% significance level.

Table 5. Tests of statistical assumptions and Welch's ANOVA for soil quality index (SQI) among five villages.

Test	Test statistic	df_1	df_2	P-value
Shapiro–Wilk test	$W = 0.97825$	—	—	0.01765
Levene's test	$F = 3.5186$	4	145	0.00898
Welch's ANOVA	$F = 19.653$	4	71.85	5.811×10^{-11}

Note: The Shapiro–Wilk test was applied to ANOVA residuals. Levene's test assessed homogeneity of variance. As unequal variances were detected, Welch's ANOVA was used to compare mean SQI among the five villages.

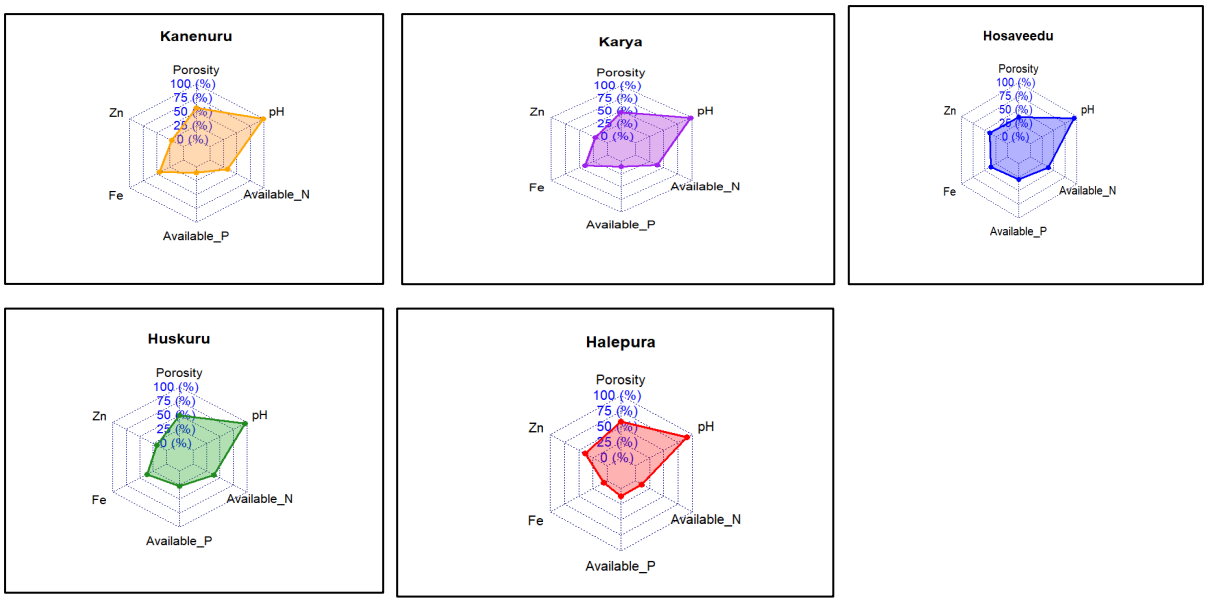


Figure 2. Radar charts showing soil-quality indicators for each village.

3.4 Random Forest Model Yield Prediction

Banana yield varied considerably across the 150 sampled fields, with a mean of 46.53 ± 10.14 t ha^{-1}. Yields ranged from 30.00 to 77.60 t ha^{-1}, indicating substantial field-level variability. This variation provided a basis for examining the relationships of soil properties, crop characteristics and management factors with banana yield (Table 8). The Random Forest model was evaluated using a testing subset comprising 28 observations (Table 7). Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). Based on the 28 observed and predicted yield values in the test dataset, the test-set R^2 was 0.700, with an RMSE of 5.32 t ha^{-1} and an MAE of 4.30 t ha^{-1}. The squared correlation between observed and predicted yields was 0.765. Observed and predicted yields showed a positive relationship in the scatter plot (Figure 3). Banana variety was by far the strongest predictor (%IncMSE = 34.44), well ahead of zinc, iron, latitude, organic carbon, available potassium and available nitrogen (Table 9). The variable-importance value for banana variety was 5.19 times that of zinc, the next-ranked predictor. Planting material, manganese, village and irrigation method contributed comparatively little. Electrical conductivity, bulk density, porosity and available phosphorus returned negative or near-zero %IncMSE values, suggesting that these variables added little predictive power once the other factors were taken into account. Variable-importance values reflect the relative contribution of predictors to model performance but do not establish causal relationships with banana yield. The high importance of banana variety does not demonstrate that changing varieties would increase yield by a

specific amount, nor do the rankings quantify individual nutrient effects. The importance of zinc, iron and soil organic carbon highlights their potential relevance to soil fertility assessment. However, nutrient-management recommendations should be guided by soil-test results and validated through targeted field experiments rather than inferred from variable importance alone.

4. DISCUSSION

4.1 Soil Variability and Fertility Status

The wide variation in soil properties indicates substantial spatial heterogeneity across the sampled banana fields (Table 2), potentially reflecting differences in soil-forming processes, organic-matter management, fertiliser history, irrigation and cultivation duration. Alkaline soil pH can reduce zinc and iron availability through adsorption, precipitation and the formation of sparingly soluble compounds (Saleem et al., 2023; Chandrakar et al., 2024; Rengel, 2015). The correlation analysis (Figure 1b) revealed negative correlations between soil pH and iron ($r = -0.60$) and manganese ($r = -0.41$). Iron was positively correlated with manganese ($r = 0.61$) and copper ($r = 0.48$). These relationships highlight the interdependence of soil properties and the relevance of PCA for selecting representative indicators for the Soil Quality Index (SQI). (Abass et al., 2023). Similar studies have identified organic carbon, pH, nutrient status, bulk density and micronutrients as important indicators of soil quality (Abass et al., 2023; AbdelRahman and Metwaly, 2023). The low-to-moderate organic-carbon levels observed suggest scope for improvement through organic amendments. In banana systems, Sathya et al. (2024) reported improved soil fertility and productivity under banana–cowpea intercropping with appropriate nutrient management. Integrating organic amendments and suitable intercrops may therefore support soil-fertility management in the study area. Furthermore, the observed variability in pH, organic carbon and nutrient availability highlights the limitations of uniform fertiliser recommendations. Field-specific soil testing could help target limiting nutrients and avoid unnecessary applications where nutrient supply is adequate.

4.2 Importance of Zinc and Iron

Zinc and iron were notable because both were selected for the Minimum Data Set (MDS; Table 3) and ranked among the stronger predictors in the Random Forest model (Table 9). Both micronutrients play important roles in banana physiology: zinc contributes to enzyme activation, hormone synthesis and growth regulation, while iron is essential for chlorophyll formation, respiration and electron transport (Arthi and Shakila, 2024). Chandrakar et al.

(2024) reported improved banana growth following zinc application and increased chlorophyll content following iron application. Zhang et al. (2025) further demonstrated that synchronising fertiliser timing and nutrient release with crop uptake can improve nutrient-use efficiency in banana. Nevertheless, the importance of Zn and Fe in the present study indicates predictive association rather than causation. Their %IncMSE values reflect their contribution to model performance, not the yield response expected from increasing their availability. Similarly, their inclusion in the MDS reflects their contribution to soil-quality assessment rather than independent control over yield. Therefore, micronutrient management should be guided by soil-test results and, where feasible, plant-tissue analysis, considering pH, organic matter, nutrient interactions and fertiliser application methods. Site-specific assessment of Zn and Fe status may help improve nutrient-management efficiency in banana fields.

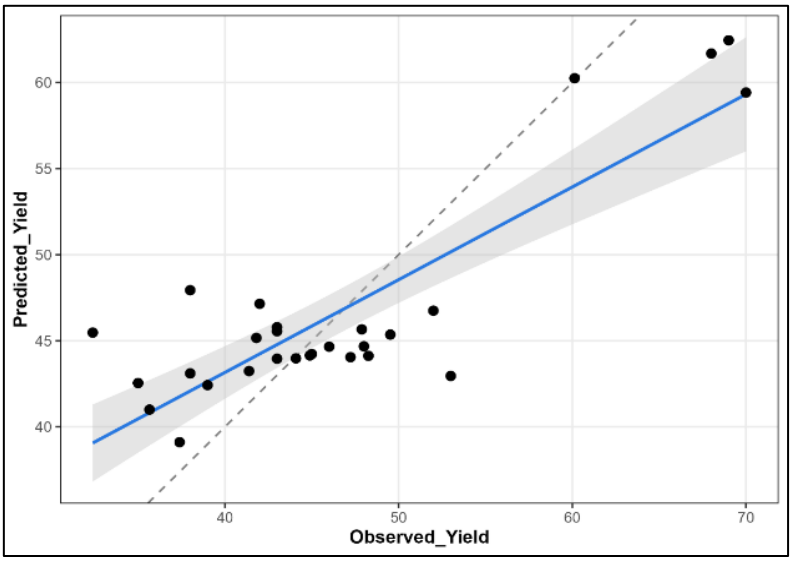


Figure 3. Observed versus predicted banana yield from the Random Forest model.

Table 6. Pairwise comparison of soil quality index (SQI) among five villages using the Games–Howell test.

Comparison	Mean difference	95% CI	Adjusted p-value	Significance
Halepura – Hosaveedu	0.07776	0.01971 to 0.13581	0.003719	**
Halepura – Huskuru	0.08887	0.04162 to 0.13612	0.00002156	****
Halepura – Karya	0.08936	0.04532 to 0.13340	0.00000436	****
Halepura – Kanenuru	0.12179	0.08146 to 0.16213	1.0612×10^{-10}	****
Hosaveedu – Huskuru	0.01112	−0.05252 to 0.07475	0.987699	ns
Hosaveedu – Karya	0.01160	−0.04987 to 0.07307	0.983419	ns
Hosaveedu – Kanenuru	0.04404	−0.01506 to 0.10313	0.231387	ns
Huskuru – Karya	0.00049	−0.05108 to 0.05206	1.000000	ns
Huskuru – Kanenuru	0.03292	−0.01566 to 0.08150	0.323542	ns
Karya – Kanenuru	0.03243	−0.01305 to 0.07792	0.275072	ns

Note: Mean differences are group 1 minus group 2; CI = 95% confidence interval; ** $p < 0.01$; **** $p < 0.0001$; ns = not significant ($p \geq 0.05$).

4.3 Spatial Differences in Soil Quality

Significant differences in SQI among the five villages (Table 4) indicated spatial variation, with Kanenuru recording the highest mean SQI and Halepura the lowest. Although all villages were classified as having moderate soil quality, the radar chart (Figure 2) revealed differences in individual indicators, suggesting varying soil-management priorities. Previous studies have highlighted the importance of spatial soil variability for site-specific management (AbdelRahman and Metwaly, 2023; Ibno Namr and Bel-Lahbib, 2023). SQI is a useful diagnostic tool but is not a direct proxy for yield, which also depends on crop and management factors. The Random Forest model identified variety as the strongest yield predictor, followed by zinc, iron, latitude, organic carbon, potassium and nitrogen (Table 9), demonstrating the value of integrating soil quality with crop and management information.The soil quality assessment was based on a Minimum Data Set (MDS) comprising iron, porosity, soil pH, available nitrogen, available phosphorus and zinc. However, biological properties were not comprehensively assessed; therefore, the SQI represents an indicator-based assessment of soil condition rather than a complete measure of soil health. Future studies incorporating biological indicators, such as microbial biomass and soil enzyme activities, alongside physical and chemical properties could provide a more comprehensive assessment of soil quality in banana-growing systems.

4.4 Random Forest Performance and Yield Determinants

The Random Forest model showed promising predictive performance, with a squared correlation of 0.765, a conventional test-set R^2 of 0.700, an RMSE of 5.32 t ha^{-1} and an MAE of 4.30 t ha^{-1} (Table 7). Observed and predicted yields showed a positive relationship (Figure 3). The model identifies predictive associations rather than causal relationships.

Table 7. Random Forest model performance.

Model	Trees	Test-set squared correlation (r^2)	Test-set (R^2)	RMSE (t ha^{-1})	MAE (t ha^{-1})
Random Forest regression	500	0.765	0.700	5.32	4.30

Banana variety ranked highest in variable importance (%IncMSE = 34.44), followed by zinc, iron, latitude, organic carbon, potassium and nitrogen (Table 9). This highlights the contribution of variety to yield prediction and the potential relevance of soil factors to site-specific management. However, variable importance does not establish the direction or magnitude of effects, and nutrient application should not be assumed to increase yield without field validation. Soil-test-based nutrient management and locally appropriate practices are

recommended. Similar findings were reported by Huang et al. (2025), Ndayambaje and Ngango (2025) and Pathak et al. (2023). Random Forest performance may decline beyond the conditions represented in training data (Silva et al., 2023). As the model was developed in a limited geographical area, its wider applicability remains uncertain without independent validation. Further testing across locations, seasons and varieties is needed; therefore, the model should be considered exploratory rather than universally validated.

Table 8. Descriptive statistics of banana yield across the sampled farms (n = 150).

Statistic	Banana yield (t ha^{-1})
Number of farms	150
Mean	46.53
Standard deviation	10.14
Minimum	30.00
Maximum	77.60
Range	47.60

Table 9. Variable importance ranking of predictors for banana yield.

Rank	Predictor Variable	%IncMSE	IncNodePurity
1	Variety	34.44	4668.31
2	Zn	6.64	406.58
3	Fe	4.3	669.23
4	Latitude	3.24	418.68
5	Organic Carbon	2.69	595.31
6	Available K_2O	2.64	677.02
7	Available N	2.53	226.4
8	Particle Density	2.48	267.36
9	Planting material	2.4	253.82
10	Mn	2.11	375.24
11	Village	1.47	167.25
12	Irrigation method	0.62	134.69
13	MWHC	0.46	287.3
14	Longitude	0.05	583.61
15	pH	-0.03	577.42
16	Cu	-0.05	376.45
17	Available P_2O_5	-0.67	363.48
18	Porosity	-1.03	236.19
19	Bulk Density	-1.1	256.88
20	EC	-1.43	513.33

5. CONCLUSIONS

This study integrated a PCA-based Soil Quality Index (SQI) with Random Forest modelling to assess soil quality and predict banana yield in five villages of Nanjangud taluk, Karnataka. Six indicators-iron, porosity, soil pH, available nitrogen, available phosphorus and zinc-formed the Minimum Data Set (MDS). Kanenuru recorded the highest mean SQI, while Halepura recorded the lowest, although all villages had moderate soil quality. The Random Forest model showed promising predictive performance, with a squared correlation of 0.765, an RMSE of

5.32 t ha^{-1} and an MAE of 4.30 t ha^{-1}. Banana variety was the strongest predictor, followed by zinc and iron. Integrating soil-quality assessment with crop and management information may support site-specific decision-making. However, the selected indicators reflect local soil conditions, and independent validation across locations, seasons and varieties is required before wider application of the model.

6. ACKNOWLEDGMENTS

ChatGPT (OpenAI) was used during manuscript preparation for language editing and improving the clarity of the text. The authors take full responsibility for the accuracy, originality, interpretation, and integrity of the manuscript.

7. REFERENCES

1. Abass A, Fanuel L, Gifole G, Lakew G. Explaining the soil quality using different assessment techniques. Appl Environ Soil Sci. 2023; 2023:6699154. https://doi.org/10.1155/2023/6699154

2. AbdelRahman MAE, Metwaly MM. Digital soil characteristics mapping for aiding site-specific management practices in the West Nile Delta, Egypt. Discov Sustain. 2023; 4:47. https://doi.org/10.1007/s43621-023-00162-6

3. Arthi LR, Shakila A. Effect of micronutrients on enhancing the quality of banana (Musa spp.) cv. Red Banana. Int J Curr Microbiol Appl Sci. 2024;13(3):117-120. https://doi.org/10.20546/ijcmas.2024.1303.010

4. Breiman L. Random forests. Mach Learn. 2001; 45:5-32. https://doi.org/10.1023/A:1010933404324

5. Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T. W., Mäder, P., Pulleman, M., Sukkel, W., van Groenigen, J. W. and Brussaard, L. 2018. Soil quality – A critical review. Soil Biology and Biochemistry, 120: 105–125.

6. Chandrakar M, Tiwari SP, Verma LS, Guhey A, Rajput GS. Physiological functions of micronutrients in enhancing banana (Musa paradisiaca L.) yield. Int J Adv Biochem Res. 2024;8(10):336-345. https://doi.org/10.33545/26174693.2024.v8.i10e.2484

7. Food and Agriculture Organization of the United Nations. FAOSTAT: Crops and livestock products-Bananas: Production, harvested area and yield data. Rome: FAO; 2023. https://www.fao.org/faostat/en/#data/QC

8. Food and Agriculture Organization of the United Nations. FAOSTAT analytical brief 96: Global fruit production. Rome: FAO; 2025.

9. Food and Agriculture Organization of the United Nations. Bananas: Markets and trade. Rome: FAO; 2026. https://www.fao.org/markets-and-trade/commodities-overview/bananas-tropical-fruits/bananas/en

10. Huang S, Wang L, Liu Y, Jiang C, Song K, Zhao Y, et al. Banana yield prediction based on machine learning algorithm. J Trop Biol. 2025;16(1):21-30. https://doi.org/10.15886/j.cnki.rdswxb.20240031

11. Ibno Namr K, Bel-Lahbib S. Use of spatial variability of soil quality index models and soil properties for soil quality evaluation in the irrigated perimeter, semi-arid region of Morocco. Earth Syst Environ. 2023; 7:857-879. https://doi.org/10.1007/s41748-023-00352-2

12. Jackson ML. Soil chemical analysis. New Delhi: Prentice-Hall of India Private Limited; 1973.

13. Karlen DL, Andrews SS, Wienhold BJ, Zobeck TM. Soil quality assessment: past, present and future. J Integr Biosci. 2008;6(1):3-14.

14. Lindsay WL, Norvell WA. Development of a DTPA soil test for zinc, iron, manganese, and copper. Soil Sci Soc Am J. 1978; 42:421-428. https://doi.org/10.2136/sssaj1978.03615995004200030009x

15. Maurya S, Abraham JS, Somasundaram S, Toteja R, Gupta R, Makhija S. Indicators for assessment of soil quality: a mini-review. Environ Monit Assess. 2020; 192(9):604. https://doi.org/10.1007/s10661-020-08556-z

16. Ndayambaje A, Ngango J. Predicting banana crop yield in Eastern Rwanda using machine learning models: Insights from agricultural and meteorological data. Cogent Food Agric. 2025;11(1):2578320. https://doi.org/10.1080/23311932.2025.2578320

17. Paul GC, Saha S, Ghosh KG. Assessing the soil quality of Bansloi river basin, eastern India using soil-quality indices (SQIs) and Random Forest machine learning technique. Ecol Indic. 2020. https://doi.org/10.1016/j.ecolind.2020.106804

18. Pathak D, Miranda M, Mena F, Sanchez C, Helber P, Bischke B, et al. Predicting crop yield with machine learning: An extensive analysis of input modalities and models on a field and sub-field level. arXiv. 2023. https://doi.org/10.48550/arXiv.2308.08948

19. Rengel Z. Availability of Mn, Zn and Fe in the rhizosphere. J Soil Sci Plant Nutr. 2015;15(2):397-409.

20. Robinson JC, Galán Saúco V. Site selection, soil requirements and soil preparation. In: Bananas and plantains. 2nd ed. Wallingford: CABI; 2010. p.115-123. https://doi.org/10.1079/9781845936587.0115

21. Saleem A, Zulfiqar A, Saleem MZ, Ali B, Saleem MH, Ali S, et al. Alkaline and acidic soil constraints on iron accumulation by rice cultivars in relation to several physio-biochemical parameters. BMC Plant Biol. 2023; 23:397. https://doi.org/10.1186/s12870-023-04400-x

22. Sathya S, Akila N, Velmurugan K, Paulpandi S, Muthusamy N. Soil fertility changes and crop productivity of banana by continuous adoption of banana-cowpea intercropping with nutrient management practices. Asian J Soil Sci Plant Nutr. 2024;10(4):114-120. https://doi.org/10.9734/ajsspn/2024/v10i4387

23. Silva JV, van Heerwaarden J, Reidsma P, Laborte AG, Tesfaye K, van Ittersum MK. Big data, small explanatory and predictive power: Lessons from random forest modeling of on-farm yield variability and implications for data-driven agronomy. Field Crops Res. 2023; 302:109063 https://doi.org/10.1016/j.fcr.2023.109063

24. Wairegi LWI, van Asten PJA. The agronomic and economic benefits of fertilizer and mulch use in highland banana systems in Uganda. Agric Syst. 2010;103(8):543-550. https://doi.org/10.1016/j.agsy.2010.06.002

25. Wairegi LWI, van Asten PJA, Tenywa MM, Bekunda MA. Abiotic constraints override biotic constraints in East African highland banana systems. Field Crops Res. 2010;117(1):146-153. https://doi.org/10.1016/j.fcr.2010.02.010

26. Yánez-Cajo D, Vásconez-Montúfar G, Villamar-Torres RO, Godoy-Montiel L, Jazayeri SM, Pérez-Porras F, Mesas-Carrascosa F. Banana yield prediction using Random Forest, integrating phenology data, soil properties, spectral technology, and UAV imagery in the Ecuadorian littoral region. Sustainability. 2025;17(22):10098. https://doi.org/10.3390/su172210098

27. Zhang X, Zhao L, Huang Q, Pang B, Zhou Z, Lu Y. Synchronized fertilization based on crop nutrient uptake and fertilizer nutrient release characteristics increases nutrient use efficiency in banana. Sci Rep. 2025; 15:34449. https://doi.org/10.1038/s41598-025-17602-0