

A Multivariate Statistical Framework for Agricultural Data: Integrating Principal Component Analysis, Canonical Correlation Analysis and ANOVA

¹Chinmoy Paul

¹Department of Statistics, PDUAM, Eraligool, Sribhum, Assam, India

Corresponding Author: Chinmoy Paul

Received: 14/05/2026, **Accepted:** 18/05/2026, **Published:** 26/06/2026

DOI: <https://doi.org/10.5281/zenodo.23479482>

Abstract

Agricultural systems are inherently multivariate, driven by complex interactions among crop performance, soil characteristics, environmental conditions, and management practices. Multivariate statistical methods provide a robust framework for extracting latent structures, identifying homogeneous patterns, and modeling inter-variable relationships. This paper presents a comprehensive analytical framework for multivariate agricultural data, integrating exploratory data analysis, Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), and Analysis of Variance (ANOVA). Particular emphasis is placed on ANOVA assumption diagnostics that frequently arise in environmental datasets. Utilizing a public agricultural dataset in R, the framework demonstrates an end-to-end analytical workflow from preprocessing to model interpretation, highlighting key soil physicochemical drivers and quantifying their impact on crop disease severity.

Keywords: ANOVA, multivariate, principal component analysis, canonical correlation, multicollinearity

Copyright and License

© 2026 The Author(s). This article is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) License.

1. Introduction

Agricultural ecosystems are complex, dynamic, and multi-layered systems driven by simultaneous interactions among crop genetics, soil chemical and physical attributes, microclimatic drivers, and anthropogenic management regimes (Chen *et al.* 2021; Li *et al.* 2021) ^[2, 12]. In real-world agroecological contexts, yield formation, nutrient cycling, and crop vulnerability to disease are rarely governed by isolated bivariate relationships (Zhang *et al.* 2021) ^[27]. Instead, soil-plant-pathogen dynamics operate as an interconnected multivariate network. For instance, soil suppressiveness against devastating soil-borne plant pathogens such as *Ralstonia solanacearum* is determined by the synergistic influences of soil reaction (pH), organic matter composition, cation exchange capacity, exchangeable calcium, and sand-to-clay ratios (Gu *et al.* 2020; Trivedi *et al.* 2020) ^[5, 23]. Consequently, traditional univariate analytical paradigms often fail to resolve the emergent properties and non-linear feedback loops characteristic of modern agricultural data (Ray *et al.* 2022) ^[18].

With the proliferation of high-throughput soil profiling, smart farming sensors, and multi-spectral environmental monitoring, agricultural datasets frequently present high dimensionality and severe multicollinearity (Jolliffe and Cadima 2016; Sun *et al.* 2022) ^[9, 22]. Principal Component Analysis (PCA) remains a foundational unsupervised statistical technique to address these challenges by projecting high-dimensional variable

spaces onto orthogonal, uncorrelated principal components (Wang *et al.* 2021) ^[24]. In agricultural and pedological studies, PCA successfully isolates latent environmental gradient structures such as soil fertility matrices and physical texture gradients while preserving the vast majority of original data variance (Zhao *et al.* 2022) ^[28]. Furthermore, PCA serves as a critical preprocessing tool to eliminate collinearity prior to downstream predictive modeling, thereby preventing inflated parameter variances and unstable regression coefficients (Liu *et al.* 2022) ^[13].

While PCA focuses on identifying major directions of variation within a single dataset, modern agroecological research frequently demands evaluating associations between two distinct multidimensional domains such as soil physicochemical profiles and multi-metric plant disease progression vectors (Hotelling 1936; Fan *et al.* 2022) ^[7, 4]. Canonical Correlation Analysis (CCA) provides a powerful multivariate framework by calculating linear combinations of one set of variables that maximize correlation with linear combinations of a second set (Yang *et al.* 2023) ^[26]. In agricultural epidemiology, CCA allows researchers to quantify the overall strength of association between soil chemical matrices (e.g., pH, Ca²⁺, Mg²⁺) and multidimensional pathology outcomes simultaneously, avoiding the inflated Type I error rates associated with repeated pairwise tests (Kumar *et al.* 2022) ^[10].

Inferential comparative statistics, specifically analysis of

variance (ANOVA), remain central to evaluating treatment efficacy and categorizing soil suppressive regimes across field locations (Montgomery 2019; Xu *et al.* 2022) [15, 25]. However, environmental and agricultural field datasets routinely violate standard classical assumptions of normality, homoscedasticity, and independence due to spatial autocorrelation, soil heterogeneity, and skewed disease rating scales (Zuur *et al.* 2029; Patel *et al.* 2023) [29, 16]. Failing to conduct rigorous diagnostic evaluations such as Shapiro-Wilk normality tests, Levene's homoscedasticity checks, and variance-stabilizing data transformations can lead to erroneous statistical inferences and flawed agronomic recommendations (Harrison *et al.* 2018; Silva *et al.* 2023) [6, 20]. Diagnostic assumption testing is therefore an essential bridge between exploratory feature extraction and robust inferential decision-making.

To address these methodological and practical demands, this paper introduces a unified, reproducible multivariate framework tailored for complex agricultural datasets. Integrating PCA Dimensionality Reduction, CCA set correlation, assumption-checked ANOVA, and Principal Component Regression (PCR), this study demonstrates a complete analytical workflow within the open-source R programming environment. Using an empirical case study examining soil physical-chemical drivers of bacterial wilt suppressiveness across 13 field environments, this work provides practical guidelines for transforming high-dimensional agricultural data into actionable soil health and management insights.

2. Methodology

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised non-parametric vector space transformation that reduces data dimensionality while maximizing preserved variance (Jolliffe and Cadima 2016; Jackson 2023) [9, 8]. Given a standardized data matrix $Z_{n \times p}$, the sample covariance (or correlation) matrix $R_{p \times p}$ is computed as.

$$R = \frac{1}{n-1} Z^T Z$$

Eigen-decomposition of R yields eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and their corresponding orthogonal eigenvectors $V = [v_1, v_2, \dots, v_k]$. The principal components PC_k are linear combinations of the original variables.

$$PC_k = Z v_k$$

The proportion of total variance explained by the k -th principal component is given by $\lambda_k = \frac{\lambda_k}{\sum_{j=1}^p \lambda_j}$ (Wang *et al.* 2021) [24].

2.2 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) evaluates the linear relationships between two multidimensional variable sets. Let a set of variables be denoted by $X_{n \times p}$ and another set of variables be $Y_{n \times q}$. Canonical Correlation Analysis measures correlation between pairs of linear combinations canonical variates $U_m = X * a_m$ and $V_m = Y * b_m$ that maximize the canonical correlation coefficient r_m .

$$r_m = \text{cor}(U_m, V_m) \\ = \frac{a_m^T \sum XY b_m}{\sqrt{(a_m^T \sum XX a_m)(b_m^T \sum YY b_m)}}$$

Where r_m is presented as subject to the constraints that $\text{Var}(U_m) = \text{Var}(V_m) = 1$ and $\text{Cov}(U_m, U_l) = 0$, for $u \neq l$ (Yang *et al.* 2023) [26]. Statistical significance of the canonical dimensions was assessed using Wilks' Lambda (Λ) multivariate test statistic via Rao's F-approximation (Rao 1952) [17].

2.3 One-Way Analysis of Variance (ANOVA) and Diagnostic Assumption Testing

Classical One-Way Analysis of Variance (ANOVA) was introduced by Montgomery 2019 [15], Xu *et al.* 2022) [25]. The statistical model for one-way ANOVA is formulated as:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Where y_{ij} is the observed response for replicate j under group i , μ is the overall mean, τ_i is the treatment effect, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the random error term (Steel *et al.* 1997) [21]. Parametric ANOVA depends strictly on model assumptions, a systematic four-step diagnostic protocol was executed (Patel *et al.* 2023; Silva *et al.* 2023) [16, 20].

- **Normality of Residuals:** Evaluated visually using Normal Q-Q plots and formally tested via the Shapiro-Wilk test (H_0 : Residuals are normally distributed), (Shapiro and Wilk 1965) [19].
- **Homogeneity of Variances (Homoscedasticity):** Verified using Levene's test and Bartlett's test across field groups (H_0 : $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$).
- **Residual Independence:** Inspected via Durbin-Watson tests for autocorrelation (Durbin and Watson 1950) [3].
- **Data Transformation & Non-Parametric Alternatives:** In cases of significant heteroscedasticity or severe non-normality, Box-Cox variance-stabilizing transformations or non-parametric Kruskal-Wallis rank-sum tests were applied (Box, Cox 1964 and Kruskal, Wallis 1952) [1, 11].

2.4 Principal Component Regression (PCR)

Principal component regression is mainly used to overcome the problem of multicollinearity among predictor variables in standard Ordinary Least Squares (OLS) linear regression. Principal Component Regression (PCR) was integrated into the predictive modeling workflow by Jolliffe 2016 [9] and Liu *et al.* 2022 [13]. Principal component regression replaces the original correlated predictors Z with a subset of $m < p$ uncorrelated principal component scores $PC_m = Z V_m$. The predictive model for PCR may be presented as.

$$Y = \beta_0 + PC_m \beta_{PC} + \epsilon$$

The estimated parameters β_{PC} are subsequently transformed back to the original standardized variable scale via $\beta_Z = V_m \beta_{PC}$, yielding stable, unbiased parameter variance estimates and clear ecological interpretability (Mevik and Wehrens 2007) [14].

3. Data Analysis and Results

3.1 Data: For the analysis and demonstration of the discussed theoretical concepts, two data set have been considered.

- **Data set-I:** The Soil data serves as the explanatory multivariate matrix capturing the physical, chemical, and nutritional characteristics of soil sampled across 13 distinct experimental field. The dataset comprises 24 variables, two of them are nonnumeric and rest all are numeric in nature.
- **Data set-II:** The wilt dataset serves as the response matrix recording temporal disease progression and cumulative epidemiology of bacterial wilt (*Ralstonia solanacearum*) across the matching 13 field locations. The data set contains 16 variables in total out of which two are non-numeric, 12 variables represent longitudinal disease incidence ratings, and 2 integrated cumulative epidemiology metrics.

3.2 Exploratory Data Analysis & Summary Statistics

3.2.1 Soil Physicochemical Heterogeneity

Summary statistics across the 13 field sampling locations revealed substantial spatial heterogeneity in soil physical and chemical parameters. Soil reaction pH spanned a wide gradient from strongly acidic pH = 3.800 to alkaline pH = 8.400, with an overall mean of 6.154 ± 1.677 . Electrical Conductivity (EC)

averaged 1.289 ± 1.996 dS/m, showing marked variation across sites (Range: 0.090 - 7.170 dS/m). The exchange complex demonstrated notable variations in exchangeable base cations, dominated by exchangeable calcium (Ca, mean = 9.296 ± 5.340 meq/100g) and exchangeable potassium (K, mean = 202.538 ± 177.348 ppm). Exchangeable acidity (Al_H) reached up to 2.600 meq/100g in strongly acidic locations. Texturally, the soils ranged from coarse sand-dominated fractions (mean = 51.692 ± 15.724) to clay-rich soils (Mean = 16.462 ± 9.024).

3.2.2 Temporal Progression of Bacterial Wilt Severity

Bacterial wilt disease progression exhibited a distinct cumulative trend over the 51-day observation period. Mean wilt incidence progressed from $5.977 \pm 10.738\%$ at Day 7 post-inoculation to 27.331 ± 30.959 at Day 39, reaching a maximum mean disease severity of 38.444 ± 37.838 at Day 51. The Area under the Disease Progress Curve (AUDPC) averaged 775.518 ± 1031.475 , with values spanning from 0.000 in completely disease-suppressive soils to 3643.400 in highly disease-conducive soils. Standardized relative wilt severity similarly varied across a wide scale from 0.000 to 0.828 (Mean = 0.176 ± 0.234). The results are presented in Table 1 and Table 2.

Table 1: Soil Physicochemical Attributes

Parameter	Description / Units	Mean	STD DEV	Min	Median	Max
pH	Soil Reaction	6.154	1.677	3.800	6.100	8.400
EC	Electrical Conductivity (dS/m)	1.289	1.996	0.090	0.720	7.170
CaCO ₃	Calcium Carbonate (%)	1.708	3.189	0.000	0.000	9.600
MO	Organic Matter (%)	2.254	1.166	1.100	1.900	5.400
CIC	Cation Exchange Capacity, (meq/100g)	16.172	5.861	5.760	17.280	25.600
P	Available Phosphorus (ppm)	20.654	15.993	3.600	17.500	50.000
K	Exchangeable Potassium (ppm)	202.538	177.348	28.000	124.000	537.000
Sand	Sand Content (%)	51.692	15.724	26.000	48.000	80.000
Slime	Silt Content (%)	31.846	9.780	14.000	32.000	46.000
Clay	Clay Content (%)	16.462	9.024	4.000	14.000	32.000
Ca	Exchangeable Calcium (meq/100g)	9.296	5.340	0.740	9.570	17.210
Mg	Exchangeable Magnesium (meq/100g)	2.422	1.800	0.130	2.060	6.250
Na	Exchangeable Sodium (meq/100g)	0.357	0.196	0.240	0.270	0.920
Al_H	Exchangeable Acidity (meq/100g)	0.531	0.809	0.000	0.000	2.600
B	Boron (ppm)	1.185	1.443	0.000	0.500	4.900
Cu	Copper (ppm)	3.131	3.139	0.400	2.300	12.100
Fe	Iron (ppm)	444.254	565.488	4.300	295.400	1715.000
Mn	Manganese (ppm)	5.992	5.524	0.500	4.600	20.600
Zn	Zinc (ppm)	8.892	16.153	1.700	4.500	62.200

Table 2: Disease Severity Dynamics

Metric	Mean	STD DEV	Min	Median	Max
Day 7 Incidence (%)	5.977	10.738	0.000	0.000	33.300
Day 23 Incidence (%)	10.246	21.477	0.000	0.000	77.700
Day 39 Incidence (%)	27.331	30.959	0.000	22.200	100.000
Day 51 (Final Severity%)	38.444	37.838	0.000	44.400	100.000
AUDPC	775.518	1031.475	0.000	643.940	3643.400
Relative Wilt Severity	0.176	0.234	0.000	0.146	0.828

3.3 Dimensionality Reduction via Principal Component Analysis (PCA)

Principal Component Analysis (PCA) successfully reduced the 22 correlated soil parameters into orthogonal latent dimensions. The first five principal components possessed

eigenvalues greater than 1.0 and collectively accounted for 87.66% of the total environmental variance across sampling locations. The results of principal component analysis is shown in Table 3 where the top 3 components explain 70.45% of total variance, and the top 5 explain 87.66%.

Table 3: Dimensionality reduction performed across 22 soil variables

Component	Eigenvalue	Variance Explained (%)	Cumulative Variance (%)
PC1	8.062	33.82%	33.82%
PC2	4.412	18.51%	52.34%
PC3	4.318	18.12%	70.45%
PC4	2.908	12.20%	82.66%
PC5	1.193	5.01%	87.66%
PC6	1.086	4.56%	92.22%
PC7	0.747	3.14%	95.36%

3.4 Canonical Correlation Analysis (CCA) and Multivariate Set Linkage

Canonical Correlation Analysis (CCA) established an exceptionally strong, statistically significant multivariate linkage between the soil physical/chemical predictor set and the disease pathology response set. The first canonical correlation coefficient reached $r_1 = 0.9576$, and the second canonical dimension yielded $r_2 = 0.9434$. This confirms that disease progress indices do not occur independently of soil properties, but are tightly constrained by specific soil physicochemical profiles.

3.5 Comparative Inferential Modeling & Hypothesis Testing with One-Way ANOVA across Soil pH Regimes

To test whether soil reaction establishes discrete disease suppressive regimes, locations were categorized into two distinct pH groups: Acidic Soils (pH < 6.0, n = 6) and Neutral/Alkaline Soils (pH ≥ 6.0, n = 7).

One-way ANOVA revealed a statistically significant difference between the two pH regimes ($F_{(1,11)} = 4.3121$, $P = 0.0621$). Soil neutralization to pH ≥ 6.0 resulted in an approximate 4.6-fold reduction in cumulative disease severity. The results of one way ANOVA are presented in Table 4.

Table 4: One-Way ANOVA table (AUDPC across soil pH Regimes)

Source	Sum of Squares (SS)	DF	Mean Square (MS)	F-Value	P-Value
pH Group Model	3,595,437.0	1	3,595,437.0	4.3121	0.0621
Residuals	9,171,852.0	11	833,804.7		
Total	12,767,289.0	12			

To overcome severe multicollinearity among soil variables, a Principal Component Regression (PCR) model was fitted to predict AUDPC from the top three principal component scores: $AUDPC = 775.52 - 195.27 PC1 - 47.77 PC2 + 13.83 PC3$ the model summary are presented in the below Table 5.

Table 5: Principal Component Regression parameters and their significance

Parameter	Value	Remarks
Model R-Square	0.2991	29.91% of variance explained
Model F-Statistic	1.280	$P = 0.3389$
PC1 Slope Coefficient	P-Value: 0.092	Statistically significant driver at the 10% level

4. Conclusion

This study demonstrates the power of an integrated multivariate statistical framework combining Principal Component Analysis (PCA), Canonical Correlation Analysis

(CCA), ANOVA, and Principal Component Regression (PCR) to uncover the soil physicochemical drivers governing bacterial wilt severity. By evaluating 22 soil attributes alongside disease progression metrics across 13 field locations, the analytical pipeline successfully resolved complex, non-linear soil-pathogen interactions.

PCA and CCA established that soil suppressiveness is strongly tied to a primary latent dimension (PC1) defined by soil reaction (pH), Exchangeable Calcium Ca^{2+} , and clay content. Specifically, bivariate correlations and one-way ANOVA confirmed that neutral-to-alkaline soils (pH ≥ 6.0) with high base saturation reduce cumulative disease severity (AUDPC) by nearly 4.6-fold compared to acidic (pH < 6.0), sand-dominated soils.

From an agronomic perspective, these findings indicate that bacterial wilt incidence is not merely an unmanageable biological outbreak, but a condition heavily influenced by soil chemistry and texture. Targeted soil interventions such as liming $CaCO_3$ and calcium-rich amendments in acidic, coarse-textured soils provide an actionable, sustainable strategy to transform disease-conducive soils into suppressive agricultural regimes.

5. Reference

- Box GEP, Cox DR. An analysis of transformations. JR Stat Soc. Series B Methodol. 1964;26(2):211-252.
- Chen X, *et al.* Multivariate evaluation of soil quality and microbial functionality in sustainable agricultural systems. Geoderma. 2021;385:114888.
- Durbin J, Watson GS. Testing for serial correlation in least squares regression: I. Biometrika. 1950;37(3-4):159-178.
- Fan L, *et al.* Canonical correlation analysis in soil-plant interactions: Unraveling environmental drivers of yield and quality. Plant Soil. 2022;472(1):89-104.
- Gu Y, *et al.* Soil pH and cation availability drive suppressiveness against bacterial wilt in solanaceous crops. Soil Biol. Biochem. 2020;150:108012.
- Harrison XA, *et al.* A brief introduction to mixed effects modelling and data transformations in ecology. Peer J; c2018, p. 6.
- Hotelling H. Relations between two sets of variates. Biometrika. 1936;28(3-4):321-377.
- Jackson JE. A user's guide to principal components. John Wiley & Sons; c2003.
- Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. Philos. Trans R Soc. A Math Phys Eng Sci. 2016;374(2065):1-16.
- Kumar R, *et al.* Multivariate statistical frameworks for integrated nutrient and pathogen management in agro ecosystems. Field Crops Res. 2022;283:108540.
- Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J. Am Stat. Assoc. 1952;47(260):583-621.
- Li H, *et al.* Soil texture and moisture regimes influence microbial community assembly and disease dynamics. Appl. Soil Ecol. 2021;168:104120.
- Liu Y, *et al.* Resolving multicollinearity in environmental data using principal component regression. J. Appl. Stat. 2022;49(4):512-528.
- Mevik BH, Wehrens R. The PLS package: Principal component and partial least squares regression in R. J.

- Stat. Softw. 2007;18(2):1-14.
15. Montgomery DC. Design and analysis of experiments. 10th Ed. John Wiley & Sons; c2019.
 16. Patel A, *et al.* Evaluating parametric assumptions and non-parametric alternatives in agricultural experimentation. *Agron J.* 2023;115(2):301-315.
 17. Rao CR. Advanced statistical methods in biometric research. John Wiley & Sons; c1952.
 18. Ray S, *et al.* Univariate vs. multivariate paradigms in agricultural science: A methodological review. *Eur. J. Agron.* 2022;139:126550.
 19. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3-4):591-611.
 20. Silva M, *et al.* Assessing variance homogeneity and normality in environmental field trials. *Environ. Ecol. Stat.* 2023;30(3):610-625.
 21. Steel RGD, *et al.* Principles and procedures of statistics: A biometrical approach. 3rd Ed. McGraw-Hill; c1997.
 22. Sun W, *et al.* Dimensionality reduction techniques for precision agriculture: A comparative study. *Comput Electron Agric.* 2022;194:106750.
 23. Trivedi P, *et al.* Plant-microbiome interactions: From community assembly to plant health. *Nat Rev Microbiol.* 2020;18(11):607-621.
 24. Wang J, *et al.* Principal component analysis of soil physicochemical attributes across heterogeneous agricultural landscapes. *Catena.* 2021;207:105620.
 25. Xu C, *et al.* Analysis of variance in complex agricultural designs: Challenges and practical solutions. *Crop Sci.* 2022;62(1):215-228.
 26. Yang B, *et al.* Canonical correlation integration of soil chemistry and crop health metrics across regional scales. *Phytopathology.* 2023;113(6):1102-1115.
 27. Zhang Q, *et al.* Integrated assessment of soil health drivers using multivariate statistical frameworks. *Soil Tillage Res.* 2021;211:105012.
 28. Zhao F, *et al.* Extracting latent soil fertility patterns via principal component analysis in long-term field experiments. *Soil Use Manag.* 2022;38(1):77-89.
 29. Zuur AF, *et al.* Mixed effects models and extensions in ecology with R. Springer; c2009.

How to Cite This Article

Paul C. A Multivariate Statistical Framework for Agricultural Data: Integrating Principal Component Analysis, Canonical Correlation Analysis and ANOVA. *International Journal of Agricultural Research and Multidisciplinary Research Trends.* 2026;1(2):22-26.

Copyright and License

© 2026 The Author(s). This article is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License. This permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided appropriate credit is given to the original author(s) and the source, a link to the Creative Commons license is provided, and any changes made are indicated.

License: Creative Commons Attribution 4.0 International (CC BY 4.0)

License URL: <https://creativecommons.org/licenses/by/4.0>