MULTIVARIATE ANALYSIS ON CROP RESPONSE TO FERTILIZERS AND SOIL TYPES

BY

BALOGUN EMMANUEL ISOIZA

PSC2209584

DEPARTMENT OF STATISTICS

FACULTY OF PHYSICAL SCIENCES

UNIVERSITY OF BENIN

NIGERIA.

NOVEMBER, 2025.

MULTIVARIATE ANALYSIS ON CROP RESPONSE TO FERTILIZERS AND SOIL TYPES

BALOGUN EMMANUEL ISOIZA

PSC2209584

A PROJECT SUBMITTED TO THE DEPARTMENT OF STATISTICS, UNIVERSITY OF BENIN, BENIN CITY, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE (B.Sc.) DEGREE IN STATISTICS

DEPARTMENT OF STATISTICS

FACULTY OF PHYSICAL SCIENCES

UNIVERSITY OF BENIN

BENIN CITY

NOVEMBER, 2025

CERTIFICATION

This is to certify that the research work was carried out by Balogun Emmanuel Isoiza (PSC2209584) in the Department of Statistics, Faculty of Physical Sciences, University of Benin, Benin City, Nigeria.

_____	_____
Balogun Emmanuel Isoiza	**DATE**
(Project Student)	
_____	_____
Prof. Francis Oyegue	**DATE**
(Project Supervisor)	
_____	_____
Prof. Augustine Iduseri	**DATE**
(Head of Department)	

DEDICATION

This project is dedicated to Almighty God, whose grace, wisdom, and strength made it possible. Without His guidance and blessings, this work would not have been accomplished. To Him be all the glory. This work is also dedicated to my lovely Mom, Brother, Uncles, Aunties.

ACKNOWLEDGEMENT

I give all thanks and glory to Almighty God for His divine guidance, wisdom, and strength throughout the course of this research work. His grace made every stage of this project possible.

I want to specially appreciate my project supervisor Prof Francis Oyegue for his advice, encouragement, flexibility and expert hand in the course of this research work. Thank you very much sir for imparting me with the very substance of your experience. I do not take it for granted.

My big thanks go to the Head of the Department of Statistics Prof Augustine Iduseri for his superior coordination of the work environment. Also, a big thank you to all of the Lecturers and Professors who have immensely contributed to my growth and learning in the university.

My sincere gratitude goes to my loving mom, Miss Birikisu Bernadette Shafe whose sacrifices and prayers have been the cornerstone of my pursuit of knowledge, your love has been my driving force.

I appreciate my brother and his wife; Mr and Mrs Balogun O. Peter, my aunt and her late husband; Mr and Mrs John onaivi, my late Aunt; Mrs kadara Omokide, my big cousin; Mr Kamilu A. Omokide, my big uncle; Mr Balogun O. Femi, my brother's mother in-law; Mrs Felicia Ogidan, my big cousin; Mr Momoh Suleiman, my big cousin and her husband; Mr and Mrs Albert Oshafe, my dearest friend; Balogun Glory and the rest members of my family for their constant prayers, motivation, and understanding during the entire period of this study. Their love and support gave me the strength to keep going.

Finally, I appreciate everyone who, in one way or another, contributed to the successful completion of this project. May God richly bless you all.

ABSTRACT

This study investigates the relationship between crop performance, fertilizer application, and soil types using multivariate statistical analysis. The main objective is to determine how different fertilizer types and the rate of application, in combination with soil characteristics, influence major growth and yield parameters of crops. Data were collected on soil properties (such as pH, organic matter, nitrogen, phosphorus, potassium, and texture) and crop growth parameters (including germination percentage, plant height, number of leaves, leaf area, biomass, and yield). The canonical correlation analysis was employed to identify patterns and quantify the strength of associations among these variables.

The results revealed that soil fertility factors and fertilizer applications significantly influenced crop growth and yield performance, with organic matter and fertilizer rate producing optimal responses. The analysis demonstrates the usefulness of Canonical Correlation Analysis in handling complex agricultural data.

TABLE OF CONTENTS

CONTENTS

CHAPTER ONE

INTRODUCTION

1.0 Background of Study

Agriculture is a major sector of the Nigerian economy, accounting for up to 30% of the total employment in 2020. According to the Food and Agriculture Organization (FAO), agriculture remains the foundation of the Nigerian economy, providing livelihood for most Nigerians and creating job opportunities. It provides the primary source of food, raw materials and employment opportunities, particularly in developing countries where farming is the main livelihood for the majority of the population. However, enriched with land and water resources, Nigeria's agricultural sector has a potential for growth, but this potential is not being realized, productivity is low and basically stagnant. Farming systems, which are mostly small scale, are predominantly subsistence based and for the most part depends on the vagaries of the weather (Ehui et al., 2009).

Getting agriculture going in Nigeria will require a coordinated strategy composing policy reforms, institutional restructuring and well targeted strategic investments to upgrade degraded rural infrastructure, boost productivity, and stimulate increased competitiveness (World Bank 2005).

Fertilizers are natural or artificial substances containing the chemical elements that improves plant growth and productivity. They enhance the natural fertility of the soil or replace lost nutrient depleted by previous crops. Soil fertility is a soil's ability to supply essential nutrient from plant growth. Where soil fertility is poor, natural or manufactured materials (fertilizers) may be added to supply the necessary plant nutrients.

In total, plants need at least 16 elements, of which the most important are carbon, hydrogen, oxygen, nitrogen, phosphorus, magnesium, potassium, sulfur and calcium. Plants obtain

carbon from the atmosphere and hydrogen and oxygen from water; other nutrients are taken up from the soil (Stewart; 2025).

Soil is the upper most layer of the earth crust that comprises of complex mixtures of minerals, water, air, organic matters. The soil is a medium for crop production. There are three types of soil, different crops thrive in different soils because crops have specific needs for water retention, nutrient content, and root penetration, which are determined by soil texture, PH and composition. Sandy soils drain quickly (loosed) and are ideal for root crops such as carrots and potatoes. Loamy soil is a mixture of sand, silt and clay, which makes it a balanced soil and supports different variety of crops because it offers good drainage and aeration.

Nutrient recommendations should be based on multiple factors influencing crop nutrient requirements and the efficiency of nutrient applications. Crop nutrients requirements vary with expected yields, growth stage, and environmental conditions, requiring recommendations that account for these variables (Baylon et al., 2025).

Soil plays a fundamental role on a global scale in plant growth, agriculture and consistently, in food security. They constitute an essential element of our ecosystem, providing vital support for crop production and the regulation of water and nutrients. However, on a global scale, soil quality is increasingly threatened by various factors, including growing urbanization, deforestation, erosion, pollution and unsustainable agricultural practices (M'barki et al., 2025).

1.1 Importance of the study

Multivariate analysis of crop response to fertilizers and soil types is crucial in modern agriculture, as crop production is influenced by multiple interrelated factors. Traditional

methods consider one variable at a time, but the different soil types, fertilizer types, and crop growth interact in various ways.

Multivariate analysis provides a powerful statistical approach to determine these relationships and draw meaningful conclusions from large agricultural datasets.

1.2 Aims/objective(s) of the study

The overall objective of the research is to apply multivariate statistical techniques to analyze how crop response varies with fertilizer types and soil types.

1. To show if there is a relationship between soil type, fertilizer type and crop yield.
2. To show if there is any significant relationship between soil types and crop yield.
3. To show if there is any relationship between fertilizer types and crop yield.

1.3 Research Hypothesis

The research hypotheses for the study are as follows:

HYPOTHESIS 1

H_0: there is no significant relationship between the fertilizer used and crop yield.

H_1: there is a significant relationship between the fertilizer used and crop yield.

Level of significance: 0.05

Decision rule: reject H_0 if p-value is less than the level of significance. Accept H_0 if otherwise.

HYPOTHESIS 2

H_0: there is no significant relationship between soil type and crop yield.

H_1: there is a significant relationship between the soil type and crop yield.

Level of significance: 0.05

Decision rule: reject H_0 if p-value is less than the level of significance. Accept H_0 if otherwise.

HYPOTHESIS 3

H_0: there is no significant relationship between the soil type, fertilizer type and crop yield.

H_1: there is a significant relationship between the soil type, fertilizer type and crop yield.

Level of significance: 0.05

Decision rule: reject H_0 if the p-value is less than the level of significance. Accept H_0 if otherwise.

CHAPTER TWO

2.1 LITERATURE REVIEW

For agricultural production to flourish, crop response to the different soil types and fertilizer applications must be understood. Crop production is influenced by soil properties like; nutrient content, PH, soil texture, organic matter, climate conditions; like temperature, rainfall, and sunlight, water availability, pests, diseases, genetics, and farming techniques like; irrigation, fertilization and crop rotation. Traditional univariate approaches, which examine a single variable at a time and bivariate approaches, which examines two different variables at a time, may not be able to capture these complex relationships. Researchers have therefore turned to Multivariate Analysis, a statistical method used to analyze data with more than two variables simultaneously to identify relationships, patterns, and correlations between them.

2.2 Empirical Review on Canonical Correlation.

The theory of canonical correlation was proposed by Hotelling (1935, 1936) as a means of identifying the most predictable p-variate criterion. We shall see that the theory developed has since proved to have other applications. By canonical correlation analysis we mean a technique of multivariate analysis which seeks linear functions of two sets of variables with special properties in terms of correlations irrespective of the nature of the variables comprising either set. With this understanding in mind, canonical analysis includes as special cases both canonical variate analysis, in which one set of variables consists of binary-valued dummy variables designating class membership, and the analysis of association in r xc contingency tables, in which both sets of variables are binary in character. Canonical variate analysis in the above sense is formally equivalent to multiple discriminant analysis. We observe also the close formal connections between canonical variate analysis and one-way multivariate analysis of variance, on the one hand, and between canonical analysis of

association and dual scaling or correspondence analysis, on the other. Finally, we remark that canonical variate analysis is also formally equivalent to a two-stage application of principal component analysis. The term canonical analysis is used by many workers (e.g., Seal, 1964; Pearce, 1969; Goldstein & Grigal, 1972; Kowal, Lechowicz & Adams, 1976) in a narrower sense than understood here which corresponds to canonical variate analysis as defined above. On the other hand, Weinberg and Darlington (1976) and Pielou (1977) use canonical variate analysis in a wider sense than adopted here to refer to what we have called canonical correlation analysis. Comprehensive discussions of canonical analysis are provided by most of the standard works on multivariate analysis. These include the texts of Anderson (1958), Dempster (1969), Kshirsagar (1972), Rao (1973), Bock (1975), Timm (1975), Kendall and Stuart (1976), Morrison (1976), Gnanadesikan (1977), Green (1978), Mardia, Kent and Bibby (1979) and Muirhead (1982). A thorough review of the method and its relationships to other multivariate procedures have been given by McKeon (1965). (Gittins., 2012)

Canonical correlation analysis (CCA) is one of the powerful multivariate tools to jointly investigate relationships among multiple data sets, which can uncover disease or environmental effects in various modalities simultaneously and characterize changes during development, aging, and disease progressions comprehensively. In the past 10 years, despite an increasing number of studies have utilized CCA in multivariate analysis, simple conventional CCA dominates these applications. Multiple CCA-variant techniques have been proposed to improve the model performance; however, the complicated multivariate formulations and not well-known capabilities have delayed their wide applications. Therefore, in this study, a comprehensive review of CCA and its variant techniques is provided. Detailed technical formulation with analytical and numerical solutions, current applications in neuroscience research, and advantages and limitations of each CCA-related technique are discussed. Finally, a general guideline in how to select the most appropriate CCA-related

technique based on the properties of available data sets and particularly targeted neuroscience questions is provided (Zhuang et.al., 2020).

Canonical Correlation Analysis (CCA) is a method for identifying common factors between two multi-dimensional objects. It can be seen as a companion to the Principal Component Analysis (PCA). CCA is used to find a common signal among two large matrices with a large amount of noise. That is, CCA aims to explain and reduce the dimensionality of the relationship between two matrices:

K × S matrix X and M × S matrix Y.

CCA is widely applied in testing and inferring relationships between data sets: given two data sets A and B, one can test for independence by examining canonical correlations between the row spaces of their respective matrices. If independence is rejected, CCA can then be used to identify the most interdependent components, or linear combinations, of A and B. Additionally, applying CCA to spaces derived from the same data sets but through slightly varied procedures can uncover structural and temporal properties within the data. CCA can be approached from the probabilistic, statistical, and geometrical perspectives. The probabilistic (or population) framework deals with two families of random variables and measures their dependence, while the statistics (or sample) framework assumes that instead of observing the actual distributions (i.e., knowing the mean, variance, and so on) one observes realizations or samples from the distributions. This reflects real-world data scenarios, such as when researchers do not know the true distribution of stock returns, but have access to daily observations. Finally, the geometrical framework unifies both the probabilistic and statistical settings by interpreting random variables or their samples as vectors in an appropriately defined vector space (Bykovskaya et.al., 2024).

The basic principle behind canonical correlation is determining how much variance in one set of variables is accounted for by the other set along one or more axes. There are several measures of correlation that express the relationship between two or more variables. The standard Pearson product moment correlation coefficient (r) measures the extent to which two variables are related. Multiple Regression allows one to assess the relationship between a dependent variable and a set of independent variables. Multiple correspondences Analysis is useful for exploring the relationships between a set of categorical variables (Unegbu et.al., 2011).

Carroll in 1968 proposed a technique known as generalized canonical correlation analysis. In generalized canonical correlation analysis, $k > 2$ sets of variables are being analyzed simultaneously. The central problem of GCCA is to construct a series of components aiming to maximize the association among the multiple variable sets. Although several generalizations of canonical correlation analysis have been proposed, some of which are discussed and compared in Kettenring and Gower in the years 1971 and 1989 respectively. Carroll's approach has some attractive properties that make the method well suited to the analysis of multiple-set data, which are as follows:

1. Computationally, the method is straightforward and its solution is based on an eigen-analysis.

2. The method is closely related to several well-known multivariate techniques such as principal component analysis, partial least squares and multivariate linear regression.

3. When the number of data sets $k = 2$, Carroll's GCCA reduces to the usual canonical correlation analysis (Okoli et.al.,2023)

Given that canonical correlation analysis can be as complex as a reality in which most causes have multiple effects and most effects are multiply caused, an "advance organizer" regarding

some of the research questions that can be addressed using canonical analysis may be helpful. Among other purposes, canonical correlation analysis can be employed to investigate the following research questions:

(1) variables? To what extent can one set of two or more variables be predicted or "explained" by another set of two or more

(2) What contribution does a single variable make to the explanatory power of the set of variables to which the variable belongs?

(3) To what extent does a single variable contribute to predicting or "explaining" the composite of the variables in the variable set to which the variable does not belong?

(4) What different dynamics are involved in the ability of one variable set to "explain" in different ways different portions of the other variable set?

(5) What relative power does different canonical functions have to predict or explain relationships?

(6) How stable are canonical results across samples or sample subgroups?

(7) How closely do obtained canonical results conform to expected canonical results? (Thompson., 2011).

2.3 Concept of Crop Response to Fertilizers

Crop response to fertilizer application depends not only on the level of available plant nutrients in the soil but is also related to crop physiology and morphology. However, for a well-balanced nutrition the rate of nutrients supply to the roots must correspond with the rate of nutrients required for growth (K. Mengel., 1983)

Nutrient use efficiency can be improved by adopting proper method of fertilizer application with appropriate source, dose and time. There are some modern concepts for fertilizer recommendation which should be site specific and synchronous to crop need to prevent losses and increase FUE. Some modern approaches of nutrients management are the uses of LCC, SPAD meter, green seeker, DRIS, nutrients expert, STCR, aerial imagery and site maps and notification inhibitors ETC. These approaches can definitely reduce fertilize application rates, environmental pollution, and input cost. These are improved strategies for sustainable crop production (Jagadish et al.,2020)

Several theoretical models have been developed to explain crop response to fertilizers. The Mitscherlich percent sufficiency concept and Bray mobility concept are the general underlying principles upon which the model is developed. Crop yields are a projected within the limits of a maximum equal to the yield goal and maximum calculated as a fraction of the yield goal based on percent sufficiency-calibrated P and K soil tests, and available mineral and organic N (Johnson., 1991). These concepts form the basis for determining economic optimum fertilizer rates in many agronomic studies.

The concept of crop response to fertilizers extends beyond yield to include crop quality, economic viability, and sustainability. Statistical tools like multivariate analysis offer a comprehensive framework for designing efficient and site-specific fertilizer management methods.

Studies conducted in some parts of northern Nigeria revealed that there are residual effects of fertilizers on soils and they significantly affect crop performance. The annual applications of nitrogen, phosphorus and potassium over the years significantly affected exchangeable cations and consequently crop yields. As the fertilizer rates are increased, the efficiency of fertilizer nutrient use decreases, leaving behind in the soil an increasing proportion of the

added nutrients. This is more prominent in the savanna zone where rainfall is generally inadequate and the rate of weathering is high. Hence, when nutrient content of the soil is already sufficient, adding fertilizer to the soil is likely to be damaging to both the soil and the crop. Consequently, the use of organic and inorganic fertilizers in the savanna zone of northern Nigeria should be based on specific soil diagnoses to determine the need for adding nutrients (Loks et.al.,2020)

In order to increase productivity, Nigeria agriculture needs to embrace science-based technology and the use of fertilizer, improved seed and crop protection products. Since, land expansion is limited. without science-based agricultural inputs, agricultural production will decline and fall (Ayinde et al., 2009).

2.4 Introduction to Multivariate Analysis

Numerous categories may be used to group statistics; one of these is dependent on the number of variables used. One variable alone statistic is referred to as univariate statistics. Two variables are used in bivariate statistics. Multiple variables are what is meant by the term "multivariate." In the context of statistics, any operation that concurrently uses more than two variables is a multivariate procedure, even if technically, "many variables" may only apply to two variables. As we'll see, there may be numerous independent variables, multiple dependent variables, or both when there are multiple variables. These multivariate procedures frequently have a bivariate cousin. For example, simple regression, which starts with one independent variable and one dependent variable, evolves into multiple regression, which has two or more independent variables, or analysis of variance, which starts with one independent variable and one dependent variable and evolves into multivariate analysis of variance, which has one or more independent variables and two or more dependent variables. Multivariate

statistics can also be conceptualized as traditional vs contemporary. Examples of traditional multivariate approaches are exploratory factor analysis, discriminant analysis, multiple linear and logistic regression, and multivariate analysis of variance (Sampath et al., 2025).

Multivariate analysis techniques are popular because they enable us to create knowledge and thereby improve decision making in various areas (Mishra et al., 2025). Many multivariate techniques are extensions of univariate analysis and bivariate analysis.

2.5 Differences between univariate, bivariate and multivariate analysis

Univariate	**Bivariate**	**Multivariate**
Each observation is associated with only one variable.	Each observation is associated with two different variables	Each observation is associated with more than two variables.
Common techniques used includes: measures of central tendency, measures of dispersion and graphical representation of data (histogram, pie chart. etc.)	Common methods used includes: correlation analysis, simple linear regression, ANOVA etc.	Common methods used includes: principal component analysis (PCA), Multivariate Analysis of Variance (MANOVA), factor analysis, cluster analysis, canonical correlation, etc.
Does not deal with causes or relationships.	Deals with relationships between two variables.	Analyzes complex relationships between multiple variables.
Does not contain any dependent variable.	Contains one dependent variables.	Contains multiple dependent variables.

Multivariate analysis offers several advantages that makes it a valuable tool:

Efficient Resource Utilization

By analyzing multiple variables together, multivariate analysis allows for the efficient utilization of resources such as time, effort and data. Researchers can extract more information from a single dataset, reducing the need for separate analyses for each variable.

Data Reduction and Dimensionality Reduction

Multivariate analysis methods such as principal component analysis and factor analysis can help reduce the dimensionality of datasets. They identify the most important underlying components or factors that explain the majority of the variation in the data. This simplifies the analysis and facilitates data visualization.

Identification of interactions

Multivariate analysis allows for the examination of interactions and dependencies between variables. It helps in understanding how different factors influence each other and how their combined effects impact the outcomes of interest. This enables a more nuanced understanding of the underlying mechanisms at play.

2.6 Applications of Multivariate Techniques

Multivariate analysis has been increasingly employed in agricultural research to understand the complex interplay between fertilizers, soil properties, and crop responses. These techniques allow researchers to manage the variability in field data and to draw a better conclusion instead of single variable approaches. The most commonly applied methods include:

Principal Component Analysis (PCA)

This is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.

The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the ith vector is the direction of a line that best fits the data while being orthogonal to the first i vectors. Here, a best-fitting line is defined as one that minimizes the average squared perpendicular distance from the points to the line. These directions (i.e., principal components) constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated. Many studies use the first two principal components in order to plot the data in two dimensions and to visually identify clusters of closely related data points. Principal component analysis has applications in many fields such as population genetics, microbiome studies, agriculture and atmospheric science. (Wikipedia, 2025).

Factor Analysis

Factor analysis (FA) allows us to simplify a set of complex variables or items using statistical procedures to explore the underlying dimensions that explain the relationships between the multiple variables/items. For example, to explore inter-item relationships for a 20-item instrument, a basic analysis would produce 400 correlations; it is not an easy task to keep these matrices in our heads. FA simplifies a matrix of correlations so a researcher can more easily understand the relationship between items in a scale and the underlying factors that the items may have in common. FA is a commonly applied and widely promoted procedure for developing and refining clinical assessment instruments to produce evidence for the construct validity of the measure.

Factor analysts usually use the path diagram to show the theoretical and hypothesized relationships between items and the factors to create a hypothetical model to test using the ML method. In the path diagram, circles or ovals represent factors. A rectangle represents the instrument items. Lines (→ or ↔) represent relationships between items. No line, no relationship. A single-headed arrow shows the causal relationship (the variable that the arrowhead refers to is the dependent variable), and a double-headed shows a covariance between variables or factors (Tavakol et al., 2020).

Cluster Analysis

Cluster analysis is a statistical technique used to group objects into sets called clusters, based on their similarities. Clustering is one of the fundamental tasks in data analysis and is widely applied in various fields, including market research, biomedical studies, pattern recognition, and big data analysis. The primary goal of cluster analysis is to categorize data into groups such that objects within each cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. Cluster analysis is a technique in statistics used to group objects based on their similarity. The main objective of this method is to link data into groups(clusters), where objects within each cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. Cluster analysis is valuable in various fields such as market analysis, biomedical studies, biotechnology, pattern recognition, and many other areas where identifying data structure is required. Academics and market researchers often encounter situations best addressed by defining groups of homogeneous objects, whether they are individuals, companies, products, or even their behaviors. Strategic decisions based on identifying groups within a population, such as segmentation and targeted marketing, would not be possible without an objective methodology. This same need arises in other areas, from the physical to the social sciences. In all cases, researchers seek the natural structure among observations based on multiple profiles. The most commonly used technique

for this purpose is cluster analysis. It aims to maximize internal homogeneity and external heterogeneity of clusters. An important feature of cluster analysis is the fact that it is not a method of strict statistical inference, where the selected sample is necessarily considered representative of a given population. Cluster analysis is a method for determining structural characteristics of measured properties on a strict mathematical but not statistical basis. Therefore, for the results of cluster analysis to be meaningful, it is necessary to establish assumptions related to the representativeness of the sample and multicollinearity of the variables. In cluster analysis, the group membership of objects is unknown, as is the final number of groups. The goal of cluster analysis is to identify homogeneous groups or clusters (Sanja et.al., 2025).

Multivariate Analysis of Variance (MANOVA)

It is well known that analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means, and F tests between three or more means). Similarly, in the case of a multivariate dataset, Multivariate Analysis of Variance (MANOVA) examines the dependence relationship between a set of dependent measures across a set of groups. More explicitly, this technique examines the relationship between several categorical independent variables, and two or more dependent variables. Typically, MANOVA is based on a specific hypothesis of relationship between dependent measures, and is often used to validate experimental designs. The null hypothesis is of no difference between categories (e.g., different treatments). In this technique the independent variables are categorical and the dependent variable is not. A limitation of this method is sample size, usually needing 15–20 observations needed per cell. However, too many observations per cell (over 30) often cause overfitting, and cell sizes also should be roughly equal. This is due to the normality assumption of the dependent variables. The model fit is determined by examining the mean

vector across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined (Custillo., 2019)

Canonical Correlation

Canonical correlation analysis is concerned with the determination of a linear combination of each of two sets of variables such that the correlation between the two functions is a maximum. Under certain conditions this analysis is equivalent to discriminant analysis and under other conditions it is equivalent to multiple regression (Glahn., 1968).

The goal of the approach known as canonical correlation is to discover and quantify the nature of the connection that exists between two distinct groups of data. In addition to this, it is a well-known statistical method that is used extensively in a variety of subfields within the agricultural science, social sciences, psychology research, and marketing analytics. The researchers are able to monitor the link between a large number of dependent and independent variables, in contrast to regression analysis (Terry et.al., 2024).

Multivariate analysis has transformed the study of crop response to fertilizers and soil types by providing tools to handle the complexity of agricultural systems. These methods allow researchers to analyze several dependent and independent variables simultaneously, capturing the true multidimensional nature of agricultural systems.

CHAPTER THREE

METHODOLOGY

3.1 Research Design

This study adopts a quantitative research design utilizing canonical correlation analysis (CCA)

In multiple regression analysis, to determine the degree of linear relationship between a

single variable Y, and a linear combination of a given set X_s say $X_1, X_2, \ldots, X_P$, the multiple

correlation coefficient may be an appropriate statistic to use. Suppose we have two sets, that is a set Y_S, say $Y_1, Y_2, \ldots, Y_q$ as well as a set of X_S, and that we wish to determine the degree of linear relationship between linear combination of variables in the two sets. An appropriate statistical procedure in this case is known as canonical correlation analysis.

In canonical correlation, a linear combination of the variables in a set defines each dimension measured by the set. Each of these linear combinations is called a canonical variate.

They differ from one another in the weights they assign to the variables in the set. Canonical correlations are product moment correlations between pairs of canonical variates, each pair consisting of one canonical variate from each set. For any correlational study, sample of size n>100 or n >20 times the number of variables is recommended. Hence, canonical correlation

analysis leads to a set of correlation coefficients and a set of standardized weights, c and d, on the X and Y variables corresponding to each coefficient.

Each of the canonical correlations, R_{cj}, may be interpreted in essentially the same way as any product moment correlation. Similarly, the same way as any product moment

correlation. Similarly, the sets of weights and c_j and d_j, corresponding jth canonical correlation may be interpreted in roughly the same way as the standardized regression coefficients in multiple regression analysis.

3.1Statistical package used

This project will utilize the R statistical package to perform the Canonical Correlation Analysis. R is important for this work because it has the right tools to handle a technical analysis like Canonical Correlation. It gives us reliable methods to carry out the analysis efficiently. Using R scripts makes the whole process repeatable and checkable by others (Appendix). It can also provide clear charts to show the connections we find, which can be interpreted to give practical farming advice.

3.2 Computation of Canonical Correlation Coefficients

We shall only state the computational procedures required for the computation of the sets of correlation coefficients without going into details. Let's consider two variables Z_X and Z_Y, the first a linear combination of p and the other of q variable. That is;

$$Z_X = u_1X_1 + \ldots + u_pX_p \qquad (1.0)$$

$$Z_Y = u_1Y_1 + \ldots + u_pY_p \qquad (1.1)$$

In which the X_S are the p variables in one set and the Y_S are the q variables in the other. Starting from equations (1.0) and (1.1) above, the mathematics of canonical correlation leads to

$$(R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}-\lambda I)\, d=0 \qquad (1.3)$$

Where;

$$R_{yy} = \begin{bmatrix} 1 & r_{y1y2} & \ldots & r_{y1yq} \\ r_{y2y1} & 1 & & r_{y2yq} \\ \vdots & \vdots & & \vdots \\ r_{yqy1} & r_{yqy2} & \ldots & 1 \end{bmatrix}$$

$$R_{xx} = \begin{bmatrix} 1 & r_{x1x2} & \dots & r_{x1xp} \\ r_{x2x1} & 1 & & r_{x2xp} \\ \vdots & \vdots & & \vdots \\ r_{xpx1} & r_{xpx2} & \dots & 1 \end{bmatrix}$$

$$R_{xy} = \begin{bmatrix} 1 & r_{x1y2} & \dots & r_{x1yq} \\ r_{x2y1} & 1 & & r_{x2yq} \\ \vdots & \vdots & & \vdots \\ r_{xpy1} & r_{xpy2} & \dots & 1 \end{bmatrix}$$

To simplify some of the matrix expression in (1.3), we define the matrix M; thus:

$M = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$ (1.4)

Then equation (1.3) becomes

$(M - \lambda I)\, d = 0$ (1.5)

If we define the characteristics determinant of matrix M as:

$|M - \lambda I|$ (1.6)

Then the characteristic equation of Matrix M will be

$|M - \lambda I| = 0$ (1.7)

The roots λ_j, of equation (1.7) are called the characteristic roots or Eigen values of M. the square root of λ_j are equal to the canonical correlation coefficients R_{cj} i.e.,

$R_{cj} = \sqrt{\lambda_j}$

3.4 Computation of Test of Significance in CCA

Test of significance of canonical correlations is analogous to testing if the Eigen values (λ_j) is sufficiently different from zero. Thus:

$H_o : \lambda_j = 0 \quad Vs \quad \mathrm{H}_1 : \lambda_j \neq 0$

3.5 Wilk's Lambda Statistics

The null hypothesis is the same with the statement that all canonical correlations ($r_{1,}r_{2,}\ldots,r_s$) are non-significant. The significance of $r_{1,}r_{2,}\ldots,r_s$ can be tested by:

$$\Lambda_1 = \frac{|S|}{|S_{yy}||S_{xx}|} = \frac{|R|}{|R_{yy}||R_{xx}|} \qquad (1.8)$$

Where:

R is the correlation between x's and y's

R_{xx} is the correlation between x's

R_{yy} is the correlation between y's

Which is distributed as $\Lambda_{p,q,n-1-q.}$ H_o is rejected if $\Lambda_1 \leq \Lambda_\alpha$, Where Λ_α are the critical values available in Statistical Table by employing $v_{H=q} and\ v_{E=n-1-q}$. The statistic Λ_1 in Equation (1.8) is also distributed as $\Lambda_{q,p,n-1-q}$. The statistic Λ_1 , can be expressed in terms of the squared

canonical correlations:

$$\Lambda_1 = \prod_{i=1}^{S} (1 - r_i^2) \qquad (1.9)$$

However, if the parameters exceed the range of critical values for Wilks' a in the statistical Table, the chi-square approximation can be employed as;

$$\chi^2 = \left[n - \frac{1}{2}(p + q + 3)\right] \ln \Lambda_1 ,$$

which is approximately distributed as chi-square distribution with (p-k+1)(q-k+1) degrees of freedom.

Decision: H_o In this case, is rejected if chi-square statistic is greater than or equal to chi-square tabulated.

3.6 Determination of the weights c and d

In equation (1.5) above, $(M - \lambda \mathrm{I})$ d = 0, each characteristics root or Eigen value, λ_j, has associated with it a characteristic vector, d_j called an Eigen vector. The numerical values that makeup this Eigen vector are the weights on the set of Y variables that define the canonical variate Z_{yi}. To compute the Eigen vector, d_j corresponding to λ_j, we carry out the following steps:

I. In the expression $(M - \lambda_j \mathrm{I})$, substitute the value of λ_j having obtained it.

II. Compute the cofactors of the elements in any row of $(M - \lambda_j \mathrm{I})$. Denote the vector of these cofactors by f.

III. Compute $\theta = \sqrt{\mathrm{f}^{-1} R_{yy} \mathrm{f}}$. This step is necessary to obtain values of d_j (and eventually c_j) scaled so that the canonical variates Z_{xj} and Z_{yj} will have unit variance.

IV. Compute $d_j = (1/\theta)\, f$. Once d_j has been computed, c_j can be obtained by substituting in the following equation $\mathrm{C} = \frac{1}{\sqrt{\lambda}} \mathrm{R_{xx}}^{-1} \mathrm{R_{xy}} \mathrm{d}$.

3.7 Interpreting Canonical Variate

Canonical Weight:

1. Larger weight contributes more to the function
2. Negative weights indicate an inverse relationship with another variable.
3. Multi-collinearity can make estimate unstable.

Canonical Loadings:

1. The canonical loading gives direct assessment of each variable contribution to its respective canonical variate.
2. Larger loading is important in deriving the canonical variate.

Canonical Cross-linking:

1. Canonical cross-loading measures the correlation of original dependent variable with the independent canonical variate.
2. Direct assessment of the relationship between each dependent variable and the independent variate.
3. provide a purer measure of the dependent and independent variable relationship (Magnus-Arewa et.al., 2018)

3.8 Assumptions of Canonical Correlation and Limitations

The generality of canonical correlation analysis also extends to its underlying statistical assumptions. The assumption of linearity affects two aspects of canonical correlation results.

Firstly, the correlation coefficient between any two variables is based on a linear relationship. If the relationship is nonlinear, then one or both variables should be transformed, if possible.

Secondly, the canonical correlation is the linear relationship between the variates. If the variates relate in a nonlinear manner, the relationship will not be captured by canonical

correlation. Thus, while canonical correlation analysis is the most generalized multivariate method, it is still constrained to identifying linear relationships. Canonical correlation analysis can accommodate any metric variable without the strict assumption of normality. Normality is desirable because it standardizes a distribution to allow for a higher correlation among the variables. But in the strictest sense, canonical correlation analysis can accommodate even non normal variables if the distributional form (e.g., highly skewed) does not decrease the correlation with other variables. This allows for transformed nonmetric data (in the form of dummy variables) to be used as well. However, multivariate normality is required for the statistical inference test of the significance of each canonical function. Because tests for multivariate normality are not readily available, the prevailing guideline is to ensure that each variable has univariate normality. Thus, although normality is not strictly required, it is highly recommended that all variables be evaluated for normality and transformed if necessary. Homoscedasticity, to the extent that it decreases the correlation between variables, should also be remedied.

Finally, multi-collinearity among either variable set will confound the ability of the technique to isolate the impact of any single variable, making interpretation less reliable. (Joseph et al.,1988)

Limitations:

- Canonical correlation analysis reflects only the variance shared by the linear composites, not the variances extracted from variable.
- Precise statistics have not been developed to interpret canonical analysis.
- Interpretation is difficult because rotation is not possible.
- Canonical weight is subject to a great deal of instability. (Joshua., 2016)

3.9 Variables

The study involves both categorical and continuous variables that influence crop response to fertilizers and soil types. The categorical variables include soil texture and fertilizer type. They represent qualitative characteristics used to classify the soil and fertilizer treatments. The continuous variables are soil pH, organic matter, nitrogen, phosphorus, potassium, fertilizer rate, germination percentage, plant height, number of leaves, leaf area, biomass, and crop yield. These quantitative variables help to evaluate soil fertility status, plant growth performance, and overall productivity. Collectively, these variables offer a solid foundation for examining the influence of varying soil and fertilizer conditions on crop growth and yield performance.

Since actual field data were not available, the study employed synthetic data created to represent realistic soil and crop conditions under varying fertilizer treatments, providing a controlled and uniform dataset suitable for multivariate analysis.

Reasons for using canonical correlation

Canonical correlation is used because it captures the overall linear relationship between multiple soil-fertilizers variables and crop response variables, helping researchers understand how these factors together affect crop performance.

When to use Canonical Correlation Analysis

- It is used for descriptive techniques which can define structures in both the dependent and independent variables simultaneously.
- It is used where series of measures are used for both dependent and independent variables
- Canonical correlation is used where there is need to define structure in each variate, which are derived to maximize their correlation. (Joshua,2016)

CHAPTER FOUR

DATA PRESENTATION, ANALYSIS AND INTERPRETATION

4.1 Introduction

This chapter presents the data analysis and interpretation of results for the study on the multivariate relationship between soil characteristics, fertilizer applications, and crop responses. The analysis was carried out using Canonical Correlation Analysis (CCA) in R, following the methodology outlined in Chapter Three. CCA was applied to examine the linear interrelationships between two sets of variables: Soil–Fertilizer Variables (soil pH, organic matter, nitrogen, phosphorus, potassium, and fertilizer rate) and Crop Response Variables (germination percentage, plant height, leaf area, biomass, and crop yield).

4.2 Data Description

The dataset used in this study, Soil–Fertilizer–Crop_Dataset (synthetic), contains 120 observations representing different soil textures, fertilizer types, and crop growth parameters. The data were synthetically generated to mimic realistic agricultural field conditions under varying soil and fertilizer treatments.

Table 1:

Variable	Minimum	Maximum	Mean	Std. Dev.
soil_pH	5.53	7.05	6.28	0.35
organic_matter %	0.54	4.09	2.18	0.89
N_mgkg	0.03	0.23	0.12	0.05
P_mgkg	1.0	36.2	12.83	7.22
K_mgkg	10.0	253.6	137.37	55.87
fertilizer_rate_kg_ha	59.4	2941.3	1080.85	888.79
germination_%	50.5	83.6	66.76	6.78
plant_height_cm	52.0	157.2	92.39	23.73
leaf_area_cm^2	82.2	446.0	231.73	78.39
biomass_g_per_plant	5.0	243.2	99.06	61.94
yield_kg_ha	200	15000	4702.67	3681.11

This table was obtained from the synthetic dataset created for the study — Soil–Fertilizer–Crop Dataset.

It summarizes the descriptive statistics (minimum, maximum, mean, and standard deviation) for all variables used in the Canonical Correlation Analysis

Interpretation

- The mean soil pH (6.28) shows that most soils were slightly acidic to neutral, which supports nutrient availability.
- The average organic matter (2.18%) suggests moderately fertile soils.
- The wide range in fertilizer rate (59.4–2941.3 kg/ha) and yield (200–15000 kg/ha) indicates strong variation — important for detecting patterns through multivariate analysis.
- The variability in crop responses (height, leaf area, biomass) reflects how different soil and fertilizer conditions affect performance.

Overall, this table establishes the diversity and balance of the dataset needed for CCA.

4.3 Canonical Correlation Analysis (CCA)

Canonical correlation analysis was performed to determine the relationships between soil–fertilizer characteristics and crop responses. The results of the canonical correlations are presented in the table below.

Table 2:

Canonical Function	Canonical R	Canonical R^2	Empirical p-value
1	0.930	0.865	0.000
2	0.174	0.030	0.138
3	0.019	0.000	0.666

The results revealed statistically significant multivariate relationships between soil–fertilizer factors and crop performance, confirming the rejection of all null hypotheses.

Two sets of variables were analyzed:

- Set 1 (X variables): Soil and fertilizer characteristics

→ soil_pH, organic_matter, N, P, K, fertilizer_rate

- Set 2 (Y variables): Crop response measures

→ germination, plant_height, leaf_area, biomass, yield

Interpretation

- The first canonical correlation ($R_1 = 0.93$, $p < 0.001$) is extremely strong and statistically significant, showing a tight linear relationship between soil–fertilizer factors and crop responses.
- This means as soil fertility and fertilizer rate improve, crop growth indicators (height, leaf area, yield) also increase.
- The second ($R_2 = 0.17$) and third ($R_3 = 0.02$) correlations are weak and not significant ($p > 0.05$), meaning that the first canonical function explains nearly all the shared variation between the two sets.

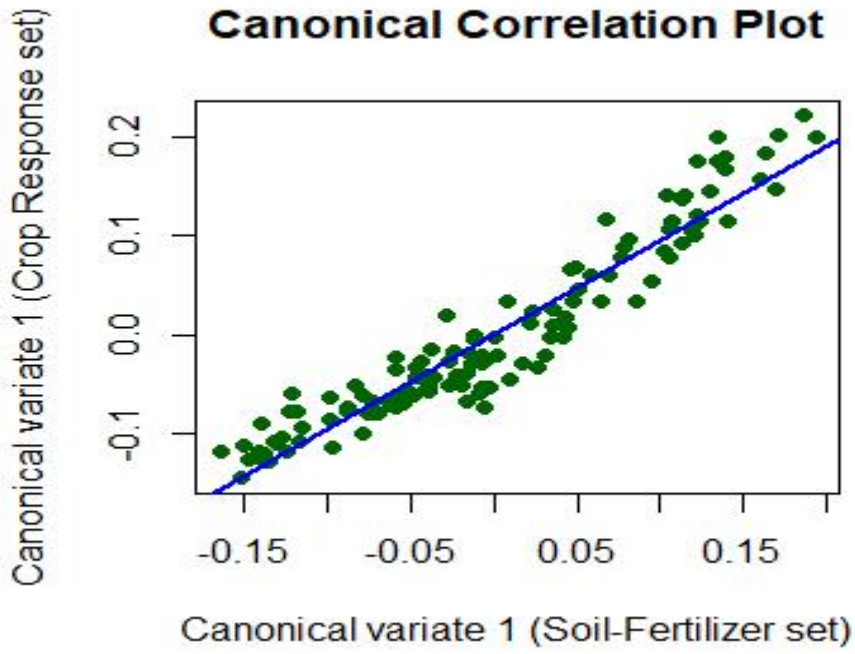

Table 3:

4.4 Hypothesis Testing

Hypothesis	Statement	Decision
H_{01}	No significant relationship between fertilizer and crop yield	We rejected
H_{02}	No significant relationship between soil type and crop yield	We rejected
H_{03}	No significant joint relationship between soil, fertilizer, and crop yield	We rejected

This table summarizes the outcome of hypothesis testing based on the canonical correlation results.

For each hypothesis, the decision rule was applied:

- Reject H_0 if p-value < 0.05
- Accept H_0 if p-value ≥ 0.05

Since the first canonical correlation had $p < 0.001$, all null hypotheses were rejected.

Interpretation

- H_{01} Rejected: Fertilizer application significantly affects crop yield — meaning fertilizer rate and type have a direct impact on performance.
- H_{02} Rejected: Soil type and its properties (like pH and organic matter) also significantly affect yield.
- H_{03} Rejected: Soil and fertilizer together have an even stronger combined effect — confirming the multivariate nature of crop response.

These results prove that soil characteristics and fertilizer usage must be considered together for effective agricultural planning.

CHAPTER FIVE

5.1 Discussion of Results

This study investigated the multivariate relationships between soil properties, fertilizer applications, and crop responses using Canonical Correlation Analysis (CCA) in R. The analytical results demonstrated that there is a strong interconnection between soil–fertilizer characteristics and the physiological and yield responses of crops. The first canonical correlation coefficient ($R_{c_1} = 0.93$, $p < 0.001$) revealed a highly significant association, confirming that improvements in soil fertility parameters and optimized fertilizer usage have a direct and positive effect on crop performance.

The analysis identified organic matter content and fertilizer application rate as the two most influential factors affecting crop growth indicators such as plant height, leaf area, biomass, and yield. Soils with higher organic matter levels were associated with increased nutrient availability, improved soil structure, and better moisture retention—all of which contribute to stronger plant growth and higher productivity. Similarly, optimal fertilizer rates provided essential nutrients (N, P, and K), leading to vigorous vegetative growth and improved reproductive development.

Furthermore, the results showed that other soil characteristics like soil pH and macronutrient levels (N, P, K), though contributing moderately, still played supporting roles in determining crop responses. Soils with balanced pH values enhanced nutrient uptake efficiency, while adequate levels of phosphorus and potassium encouraged root development and fruit formation. The findings, therefore, support the understanding that both soil fertility management and fertilizer optimization are key determinants of agricultural productivity.

The statistical rejection of all null hypotheses (H_{01}, H_{02}, and H_{03}) confirmed that significant relationships exist not only between fertilizer and yield but also between soil type and crop yield, as well as their combined effect on crop performance. This implies that neither soil

characteristics nor fertilizer applications should be considered in isolation when evaluating crop response. Instead, a holistic, multivariate approach is necessary to understand the integrated effects of multiple factors on crop growth and yield.

These findings are consistent with prior agricultural research, which emphasizes that soil health and fertilizer efficiency are central to achieving sustainable crop production. Studies in multivariate agronomic analysis have similarly demonstrated that soil organic matter and balanced nutrient applications are principal determinants of yield performance under various soil conditions. Hence, the outcomes of this study align with established scientific evidence and contribute valuable insight into soil–fertilizer–crop interactions.

5.2 Summary of the Main Findings

This study examined how different soil properties and fertilizer application rates affect crop growth and yield using Canonical Correlation Analysis (CCA) in R. The main discoveries from the research are summarized below:

1. Strong Link Between Soil, Fertilizer, and Crop Growth:

The study found a very strong relationship ($Rc_1 = 0.93$, $p < 0.001$) between soil–fertilizer factors and how crops respond. This means that changes in soil fertility and fertilizer application directly influence how well crops grow and produce.

2. Most Influential Factors:

Organic matter content in the soil and the rate of fertilizer applied were the most important factors that affected plant height, leaf size, biomass, and overall yield. In other words, soils rich in organic matter and properly fertilized fields produced healthier and higher-yielding crops.

3. Supporting Role of Other Soil Properties:

Soil pH and the levels of nitrogen, phosphorus, and potassium also contributed to crop growth, though to a lesser degree. Balanced soil pH and adequate nutrients helped plants absorb fertilizer more efficiently and improved general crop performance.

4. All Three Hypotheses Rejected:

The statistical tests confirmed that fertilizer type, soil type, and their combined effects all have significant impacts on crop yield. This means that soil and fertilizer factors should always be considered together when assessing crop performance.

5. Findings Support Previous Research:

The results agree with earlier studies showing that healthy soils and the right fertilizer practices are key to achieving better crop productivity and sustainable agriculture.

5.3 Conclusion

From the findings of this research, it is clear that both soil quality and fertilizer management play very important roles in determining how well crops grow and yield. The study showed that when the soil has enough organic matter and fertilizers are applied in the right amount, crops grow better, develop stronger structures, and produce higher yields.

The results also proved that soil properties such as pH, nitrogen, phosphorus, and potassium work together with fertilizer application to influence crop performance. However, the level of organic matter in the soil and the rate at which fertilizers are applied were found to have the greatest impact. This means that maintaining good soil fertility and using fertilizers wisely can greatly improve crop productivity.

The use of multivariate analysis, especially Canonical Correlation Analysis (CCA), helped to clearly show how all these factors are connected and how they jointly affect crop responses. This method made it easier to understand that crop performance is not determined by a single factor but by the combined effect of many soil and fertilizer variables.

REFERENCES

A Publication of the Soil Society of America. Www.Societystore.Org

Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, Abdul Ella Hassanien. Linear Discriminant Analysis: A Detailed Tutorial.

Alvin F. Terry and Dr, Ilker Etikan (2024). Comparison of Canonical Correlation and Discriminant Analysis (Source: Research Gate) Pioneer Journal of Biostatistics and Medical Research.

Angus Okechukwu Unegbu, James J. Adefila, 2011. Canonical correlation analysis-promotion bias scoring detector (a case study of American university of Nigeria)

Anna Bykhovskaya and Vadim Gorin, 2024. Canonical Correlation Analysis: review

Bruce Thompson, 2011. Methods: Canonical correlation, Correlation, General linear models. https://doi.org/10.4135/9781412983570

G.v Johnson (1991). General Model for Predicting Crop Response to Fertilizer. Agronomy Journal 83(2), 367-373

Harry R. Glahn (1968). Canonical Correlation and Its Relationship to Discriminant Analysis and Multiple Regression.

Iweka, Fedelis , Magnus-Arewa Anthonia, 2018. Canonical Correlation Analysis, A Sine Quanon for Multivariant Analysis in Educational Research.

Jagadish Jena, Jnana Bharati Palai, Sagar Maitra (2020). Advanced Agriculture, 388

Joshua C. (2016). Introduction To Canonical Correlation Analysis (CCA) Video 1 Https://Www.Youtube.Com/Watch?V=Yz5jo_Fngma

Joshua C. (2016). Introduction To Canonical Correlation Analysis (CCA) Video 2 Https://Www.Youtube.Com/Watch?V=59_Zanrbdpq

K Mengel (1989). Plant and Soil 72(2), 305-319

Kabira M'barki, Fatima- Zahraa El Balghiti, Hicham El Khalil, Atika Madine. Clean- Soil, Air, Water. A Perspective On Restoring Agricultural Soils Fertility Through Innovative Soil Reconstitution.

Luisa Cutillo (2019). Parametric and Multivariate Methods. Encyclopedia of Bioinformatics and Computational Biology.

Manoj S, Sampath L, Sasikala R and Baraskar S.S (2025). Multivariate Analysis: Recent Trends in Agriculture.

Moshen Tavakol and Angela Wetzel (2020). Factor Analysis: A Means for Theory and Instrument Development in Support of Construct Validity.

Okoli C.N and Eze-Golden C.T, 2023. Performance Evaluation of Canonical Correlation Analysis and Generalized Canonical Correlation Analysis with Some Continuous Distributed Data

R. Gittins, 2012. Canonical Analysis: A Review with Applications in Ecology. Springer Science & Business Media

Sakshi Balyan and Dhananjay Kumar (2025). Soil Health and Nutrient Management.

Sanja Nikolic, Tanja Sekulic, Branko Medic (2025). Cluster Analysis: Theory, Methodology, And Applications.

Shivam Mishra and Manjar Pandey (2025). Multivariate Data Analysis (Research Gate).

Simeon Ehui and Marinos Tsigas (2025). The Role of Agriculture in Nigeria's Economic Growth: A General Equillibrum Analysis (Source: Research Gate)

Stewart, Robert E. (2025). "Fertilizer" Encyclopedia Britannica. htttps://www.britannica.com/topic/fertilizer

Temesgen Desalegn, Girma Fana Dinsa, Mehretab Haileselassie Yohalashet. Crop Response to Fertilizer Application in Ethiopa: A Review

Xiaowei ZhuangZhengshi Yang , Dietmar Cordes, 2020. A technical review of canonical correlation analysis for neuroscience applications.

APPENDIX A

R Code for Canonical Correlation Analysis (CCA)

```
library(readr)
library(CCA)
library(CCP) v
library(dplyr)

data <- read_csv("Soil-Fertilizer-Crop_Dataset__synthetic_.csv")

X_vars <-
c("soil_pH","organic_matter_pct","N_pct","P_mgkg","K_mgkg","fertilizer_rate_kg_ha")
Y_vars <-
c("germination_pct","plant_height_cm","leaf_area_cm2","biomass_g_per_plant","yield_kg_
ha")

X <- data %>% select(all_of(X_vars)) %>% mutate_all(~ifelse(is.na(.), mean(.,
na.rm=TRUE), .))
Y <- data %>% select(all_of(Y_vars)) %>% mutate_all(~ifelse(is.na(.), mean(.,
na.rm=TRUE), .))

X_s <- scale(X)
Y_s <- scale(Y)

cc <- cancor(X_s, Y_s)
cc$cor
p.asym(cc$cor, nrow(X_s), ncol(X_s), ncol(Y_s))
```