Application of Linear Algebra in Crop Yield Analysis

Gurpreet Kaur
Assistant Professor
Department of Mathematics
Khalsa College for Women, Civil Lines, Ludhiana, India
gkbrar97@gmail.com

Abstract: Agricultural productivity is governed by multiple interrelated factors, making quantitative analysis essential for understanding yield variations across different crops. This study presents an application of linear algebra to examine agricultural yield patterns using matrix representation, eigenvalues, eigenvectors, and the least squares method. Three representative crops are considered, and four key influencing variables—rainfall, fertilizer usage, temperature, and yield—are organized into a structured mathematical framework to capture their interdependence.

Eigenvalue analysis is employed to identify dominant factors influencing overall crop performance, where the principal eigenvector provides insight into the relative contribution of each input variable. To estimate yield values corresponding to the selected agricultural factors, the least squares method is used to obtain approximate solutions to an overdetermined system of linear equations.The Moore-Penrose pseudo-inverse is used to estimate model coefficients, ensuring the numerical stability even when normal inverse does not exist. A numerical illustration is included to demonstrate the applicability of the proposed approach and to facilitate interpretation of the results.

The findings indicate that even a straightforward linear algebraic framework can provide meaningful insights into agricultural systems. The proposed model highlights the potential of basic mathematical tools in agricultural data analysis and offers a foundation for future studies incorporating real-world datasets and extended modeling techniques.

Keywords: Linear Algebra, Eigenvalues, Eigen Vectors, , Least Squares Method, Pseudoinverse, Agricultural Yield Analysis, Crop Modeling.

I. Introduction

Agriculture has always been the backbone of human civilization. It not only provides food and raw materials but also contributes greatly to the economy. However, the success of agriculture depends on several natural and man-made factors such as rainfall, soil fertility, temperature, and fertilizer use[4]. In today's world, farmers face several challenges—unpredictable weather, changing soil conditions, variations in irrigation, and increasing demand for food. All these factors make it harder to estimate crop yield accurately. Traditional farming practices rely mostly on experience, observation, and intuition. While these methods have value, they are not always reliable, especially when many factors influence crop health simultaneously. Predicting crop yield has become an important area of study because it helps farmers and planners to make better decisions and manage resources effectively.

This is where mathematics, especially linear algebra becomes important. It provides useful tools to handle and analyze such data and helps us study systems involving many variables at the same time. Agriculture is full of such variables—soil nutrients, rainfall, seed quality, fertilizer levels, temperature, pest attack intensity, irrigation hours, and many others. These factors affect crop yield collectively, not individually. Therefore, we need a mathematical tool that can understand relationships among many variables at once.

This kind of study allows researchers to develop simple prediction models that can guide farmers to make better decisions. In this way, mathematical models become a bridge between theoretical concepts and practical applications in the field of agriculture.

This study aims to show how the simple ideas of linear algebra — such as matrices, eigenvalues, and eigenvectors — can be applied to predict crop yield. It demonstrates how agricultural data can be arranged in a mathematical form and analyzed to identify which factors have the greatest impact on productivity.

II. Agricultural Challenges Requiring Quantitative Analysis

Agriculture is affected by many factors that farmers cannot fully control. Predicting crop yield becomes difficult because numerous environmental, biological, and management variables interact. Understanding these factors is necessary before developing a mathematical model.

Climate Variability

Rainfall, temperature, humidity, and sunlight vary from season to season. Too much or too little rainfall affects growth, and extreme temperatures can damage crops. Mathematical modelling helps farmers quantify these risks.

Soil Variation

Different fields contain different levels of nutrients such as nitrogen, phosphorus, and potassium. Soil texture, pH level, and organic matter content also vary. These differences make agricultural production inconsistent from one plot to another.

Fertilizer Application

Farmers apply fertilizer based on experience, but the exact amount needed depends on crop type, soil, and climate. Underuse reduces yield, while overuse damages soil and increases cost.

Irrigation Differences

Some fields depend on rainfall, while others use tube wells or canal irrigation. The amount of water supplied affects plant growth directly.

Human Limitations

Farmers may recognize patterns but cannot scientifically measure how strongly each factor affects yield. Human intuition alone cannot handle complex relationships between many variables.

Hence, there is a need for a systematic mathematical method to analyze these factors together. Linear algebra provides such a method by organizing data in matrices and studying relationships using eigenvalues.

III. Why Linear Algebra in Agricultural Modeling ?

Linear algebra is the branch of mathematics concerned with vectors, matrices, and transformations. It is used in engineering, physics, economics, machine learning, and even social sciences. Agriculture, too, is a field where interactions between multiple variables occur, making linear algebra a natural fit.

Matrices as Data Organizers

A matrix is a rectangular arrangement of numbers. In agriculture, each row of the matrix can represent a crop, and each column can represent a factor such as rainfall, soil nitrogen, fertilizer amount, and irrigation hours.

Relations Between Variables

Agricultural variables influence each other. For example, fertilizer usage may depend on soil nutrient levels. Linear algebra allows us to see how strongly factors are related.

Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors help identify the most important patterns in data. For instance:

- The largest eigenvalue may represent the dominant factor affecting crop yield.
- The corresponding eigenvector shows how strongly each agricultural factor contributes.

This approach is simple but powerful as

- It helps planners focus on the most sensitive input, and
- Avoids wasting resources on less influential factors. For example: If fertilizer dominates, training and subsidies can be targeted there.

Advantages Over Simple Observation

- Handles large data easily
- Reduces error compared to human judgment
- Helps identify hidden patterns
- Provides an objective (not emotional) way of predicting yield

Linear algebra is not only a mathematical subject—it is a tool that converts agricultural data into useful information.

IV. Objectives of the Study

The main objectives of this study are:

1. To represent agricultural data in the form of matrices and vectors.
2. To apply linear algebraic techniques to estimate the effect of various factors such as rainfall, temperature, and fertilizer use on crop yield.
3. To use eigenvalue and eigenvector analysis to determine the most influential factors affecting crop productivity.
4. To show that linear algebra provides a simple yet effective method for predicting crop yield and understanding the relationship between multiple agricultural variables.

V. Review of Literature

Mathematical modelling has long been used as a tool for understanding complex systems in agriculture. Linear models, in particular, have been widely applied due to their simplicity and interpretability.

Several researchers have applied linear regression and least squares techniques to study crop yield responses to environmental and management factors. These studies show that even simple linear relationships can capture important trends when the number of variables is limited.

Matrix-based methods are commonly used in multivariate agricultural analysis, especially when multiple crops and factors are involved. The matrix representation allows simultaneous analysis of all variables and facilitates computational efficiency.

The use of eigenvalues and eigenvectors in applied sciences is well established, particularly in data reduction, sensitivity analysis, and principal component analysis[5][6]. In agriculture, eigenvalue-based methods have been employed to identify dominant factors affecting yield variability.

This study differs from previous work by focusing exclusively on basic linear algebra concepts—matrices, eigenvalues, eigenvectors, and least squares—and presenting them in a pedagogical and application-oriented manner.

VI. Mathematical Framework

The methodology of this study is structured in two complementary stages. In the first stage, a linear regression model is formulated using a data matrix representing multiple agricultural factors across different crops. In the second stage, eigenvalue and eigenvector analysis of the matrix $A^T A$ is carried out to examine the internal structure of the data. This analysis supports the regression results by identifying dominant direction of variability and revealing the relative influence of individual agricultural factors. Together, these two stages provide both predictive capability and structural insight into the agricultural system.

The Data Matrix

Suppose we have *n* crop fields and *m* agricultural factors. We create a matrix A of size n × m, where:

- each row represents a crop
- each column represents a factor

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

Eigenvalue

Let A be a square matrix. A number λ is called an eigenvalue of A if there exists a non-zero vector v such that: $Av = \lambda v$ [1]
Here, the matrix A stretches or compresses the vector v but does not change its direction.

Eigenvector

A non-zero vector v satisfying: $Av = \lambda v$ is called an eigenvector corresponding to eigenvalue λ.

Linear Regression Model

The relationship between yield and influencing factors is modeled using a linear equation:

$$AX = B$$

where X is the vector of unknown coefficients representing the contribution of each factor to crop yield. Since the matrix A is not square, the system does not admit a unique solution in the usual sense. Therefore, an approximate solution is obtained using the least squares method.

Least Squares Solution Using the Pseudo-inverse

While estimating the coefficient vector X, if the number of observations exceeds the number of unknown coefficients, the resulting system is over-determined and an exact solution generally does not exist. In this case, the least square approach is employed using Moore-Penrose pseudo-inverse of the data matrix[7] to minimize the residual error between the observed and predicted yields, leading to the solution $X = A^+ B$. Conversely, if the number of observations is smaller than the number of unknowns, the system becomes under-determined and admits infinitely many solutions. Here also, the Moore-Penrose pseudo-inverse is used to obtain the minimum-norm solution, which provides a unique and stable estimate of the coefficient vector. Thus, the pseudo-inverse offers a unified framework for solving both over-determined and under-determined linear systems

The least squares solution is given by:

$X = A^{+} B$

where A^{+} denotes the Moore–Penrose pseudoinverse of matrix A.

For the case of over-determined system, the pseudoinverse is computed using the relation:

$A^{+} = (A^{T}A)^{-1} A^{T}$ (if $A^{T}A$ is invertible)

For the case of under-determined system, the pseudoinverse is computed using the relation:

$A^{+} = A^{T} (AA^{T})^{-1}$ (if AA^{T} is invertible)

Role of Eigenvalues and Eigenvectors

Since the matrix A is rectangular, eigenvalue analysis is performed on the symmetric matrix $A^{T}A$. The matrix $A^{T}A$ captures the combined effect of all influencing factors and plays a central role in the least squares formulation.

Eigenvalues of $A^{T}A$ indicate the relative importance of different combinations of agricultural factors. A larger eigenvalue corresponds to a direction in which the variation of yield is high, while smaller eigenvalues indicate weaker or redundant influences.

The corresponding eigenvectors represent dominant patterns among the agricultural factors. For example, an eigenvector with large components associated with rainfall and irrigation suggests that these factors jointly have a strong impact on yield.

Eigenvalue decomposition is therefore essential for understanding the structure of the data and for constructing a stable pseudo-inverse.

Interpretation of the Model

The estimated coefficient vector X provides quantitative insight into the influence of each factor on crop yield. A larger coefficient value indicates a stronger contribution of the corresponding factor.

The resulting linear model can be written as:

Predicted Yield = x_1 (Rainfall) + x_2 (Fertilizer) + x_3 (Temperature) + x_4 (Irrigation)

This model allows yield estimation even when future experimental data are unavailable. By substituting known values of rainfall, fertilizer usage, temperature, and irrigation, the yield can be predicted for subsequent seasons.

Practical Significance of the Methodology

The proposed methodology offers practical advantages foe the agricultural modeling by eigenvalue based analysis. The use of the Moore- Penrose pseudo-inverse ensures numerical stability and avoids computational difficulties arising from singular or ill-conditioned data matrices, which are common in real world agricultural datasets. The estimated regression coefficients provide quantitative measures of the contribution of individual factors such as rainfall, fertilizer, temperature, and irrigation to crop yield. In addition, eigenvalue and eigenvector analysis enhances the interpretability of the model by identifying dominant factor combination and revealing underlying relationships among variables. This integrated approach is particularly suitable for agricultural systems where data are limited, correlated, or imprecise, making the methodology a reliable and practical tool for regional crop yield estimation and decision support.

VII.Case Study: Application of Proposed Model

In the illustrative case study, yield data from three crops are considered with four influencing agricultural factors, resulting in an under-determined liner system. The Moore-Penrose pseudo-inverse is therefore employed to obtain a unique minimum-norm solution for the model coefficients. Although the real-world agricultural datasets typically lead to over-determined systems due to large number of observations, the present example is used for conceptual clarity and methodological demonstration.

Selection of Crops

In this study, three widely cultivated crops are selected in order to keep the model simple and interpretable:

1. Wheat
2. Rice
3. Maize

These crops are chosen because they differ in their response to environmental and input factors, making them suitable for comparative analysis.

Selection of Influencing Factors

Four major agricultural factors that significantly affect crop yield are considered:

- Rainfall (mm)
- Fertilizer usage (kg/ha)
- Average temperature (°C)
- Irrigation intensity (hours/week)

These factors are commonly used in crop modelling studies and are supported by agricultural literature.

Construction of the Data Matrix

Let the matrix A represent the agricultural input data for the three crops.
Each row corresponds to a crop, and each column corresponds to a factor.

$$A = \begin{bmatrix} 120 & 80 & 28 & 40 \\ 140 & 75 & 30 & 45 \\ 110 & 90 & 27 & 38 \end{bmatrix}$$

Explanation of entries:

Crop	Rainfall (mm)	Fertilizer (kg/ha)	Temperature(°C)	Irrigation (hrs/week)
Wheat	120	80	28	40
Rice	140	75	30	45
Maize	110	90	27	38

These values are realistic, positive, and representative of typical agricultural conditions, and they are used for methodological demonstration rather than exact prediction.

Yield Vector

Let the vector b represent the average yield (kg/ha) of the three crops:

$$B = \begin{bmatrix} 2600 \\ 3000 \\ 2400 \end{bmatrix}$$

This yield data reflects commonly reported yield ranges for wheat, rice, and maize.

Least Squares Estimation

The relationship between yield and influencing factors is modeled as:

$A X = B$

where:

- A is the 3×4 input matrix,
- X is the 4×1 vector of unknown coefficients,
- B is the 3×1 yield vector.

Since the system AX = B is under-determined, the Moore-Penrose pseudo-inverse will be employed to obtain the minimum-norm least-square solution $X = A^T(AA^T)^{-1}B$

Computation of AA^T

First, compute the transpose of A:

$$A^T = \begin{bmatrix} 120 & 140 & 110 \\ 80 & 75 & 90 \\ 28 & 30 & 27 \\ 40 & 45 & 38 \end{bmatrix}$$

Now multiply AA^T:

$$AA^T = \begin{bmatrix} 23184 & 25440 & 22676 \\ 25440 & 28150 & 24670 \\ 22676 & 24670 & 22373 \end{bmatrix}$$

This matrix is symmetric and positive definite, which ensures stability of the solution.

Least Squares Solution

The least squares estimate is given by:

$X = A^T(A^TA)^{-1}B$

After computing the inverse and multiplication, we obtain:

$$X = \begin{bmatrix} 17.204 \\ -0.412 \\ 11.469 \\ 6.185 \end{bmatrix}$$

Interpretation of the Coefficients

Using the least squares method based on the Moore–Penrose pseudoinverse, the coefficient vector X was estimated from the observed data. The solution provides numerical values for the contribution of each agricultural factor—rainfall, fertilizer usage, temperature, and irrigation—to crop yield.The estimated coefficient vector reflects the relative influence of these factors on yield formation.

- Rainfall coefficient (17.204) indicates a strong positive influence on crop yield.
- Fertilizer (-0.412) shows negative impact (possible overuse or correlation effect).
- Temperature (11.469) has the significant sensitivity.
- Irrigation (6.185) has moderate positive impact.

The negative coefficient does not imply an error, but it reflects interdependence among factors, which is common in agricultural data.

Final Estimated Model
Predicted Yield = 17.204 $R - 0.412\ F + 11.469\ T + 6.185\ I$
where R, F, T, and I stand for rainfall, fertilizer, temperature, and irrigation respectively.

VIII. Discussion of Results

Eigenvalue and Eigenvector Analysis

Eigenvalues of A^TA

We obtained the symmetric matrix:

$$A^TA = \begin{bmatrix} 46100 & 30000 & 10530 & 15280 \\ 30000 & 20125 & 6920 & 9995 \\ 10530 & 6920 & 2413 & 3496 \\ 15280 & 9995 & 3496 & 5069 \end{bmatrix}$$

To find the eigenvalues, we solve the characteristic equation:

$\det(A^TA - \lambda I) = 0$

Since A^TA is real and symmetric, all eigenvalues are real and non-negative, which is a fundamental result from linear algebra.

After computation (using standard numerical eigenvalue techniques), the eigenvalues are:

$\lambda_1 \approx 73282.3$, $\lambda_2 \approx 424.5$, $\lambda_3 \approx 0$, $\lambda_4 \approx 0.20$

Interpretation of Eigenvalues

- One very large eigenvalue indicates one dominant combined factor strongly affects yield.
- One eigenvalue ≈ zero which shows the strong correlation/redundancy among factors.
- Confirms why normal inverse fails.
- Justifies the use of the pseudo-inverse.

Eigenvectors of A^TA

The eigenvector corresponding to the dominant eigenvalue λ_1 is computed from:

$(A^TA - \lambda_1 I)\, v = 0$

After normalization, the dominant eigenvector is approximately:

$$v_1 \approx \begin{bmatrix} 3.012 \\ 1.978 \\ 0.690 \\ 1 \end{bmatrix}$$

Each component corresponds to:

Component	Factor
3.012	Rainfall
1.978	Fertilizer
0.690	Temperature
1	Irrigation

Interpretation of the Dominant Eigenvector

The dominant eigenvector corresponds to the largest eigenvalue of the matrix A^TA indicates the principal direction of variation in the agricultural data[5]. each component of this eigenvector quantifies the relative contribution of the corresponding agricultural factor to the overall crop yield variability.

- From the values, rainfall component has the highest magnitude (3.012), indicating that rainfall is the most influential factor affecting crop yield in the constructed model.
- Fertilizer application follows with a substantial contribution(1.978), showing its significant role in enhancing productivity.
- Temperature exhibits a moderate influence(0.690), suggesting that while important, its impact is comparativelyless dominant.
- Irrigation, normalized to unity(1.000), serves as a reference factor and contributes across crops.

Thus, the dominant eigenvector reveals the combined effect and relative importance of agricultural inputs rather than isolated influences. This analysis helps identify which factors primarily drive yield variation and provides a quantitative basis for prioritizing agricultural planning and resource allocation.

IX. Limitations and Scope for Future Work

The present model is based on a limited dataset and assumes a linear relationship between yield and influencing factors. While this assumption simplifies analysis, real agricultural systems may exhibit nonlinear behavior.

Future studies may incorporate larger datasets, nonlinear models, or time-dependent variables to enhance predictive accuracy. Nevertheless, the current approach demonstrates that even simple linear algebra tools can provide meaningful insights into agricultural systems.

Thus, the least squares–based linear model serves both as an analytical tool for understanding factor influence and as a practical framework for yield prediction in data-limited scenarios.

The study is subject to the following limitations:

1. The dataset is small and illustrative.
2. Values are representative rather than experimentally measured.
3. The model assumes linear relationships.
4. External factors such as soil type and pests are not included.
5. Despite these limitations, the study successfully demonstrates the methodological usefulness of linear algebra in agricultural analysis.

X. Conclusion

This research demonstrates that basic linear algebra tools—matrices, eigenvalues, eigenvectors, and least squares—can be effectively applied to agricultural yield analysis.

The study shows that:

- Linear algebra provides numerical insights.
- Eigenvalues identify dominant agricultural patterns.
- Least squares estimation quantifies factor influence.
- The approach is simple, interpretable, and educationally valuable.
- The model can be extended using real field data and additional variables, making it a foundation for future interdisciplinary research.

XI. References

[1] Strang, G. *Linear Algebra and Its Applications*, 4th ed., Brooks/Cole, 2006.

[2] Lay, D. C., Lay, S. R., and McDonald, J. J. *Linear Algebra and Its Applications*, Pearson, 2016.

[3] Montgomery, D. C., Peck, E. A., and Vining, G. G. *Introduction to Linear Regression Analysis*, Wiley, 2012.

[4] Food and Agriculture Organization of the United Nations (FAO). *FAOSTAT Statistical Database*. Rome: FAO.

[5] Jolliffe, I. T. *Principal Component Analysis*, Springer, 2002.

[6] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

[7] Ben-Israel, A.,& Greville, T.N.E.(2003). *Generalized Inverses: Theory and Applications(2nd ed.). Springer*