African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 8, Issue 3, 2025 (pp. 66-72)
www.abjournals.org

A MULTIVARIATE APPROACH TO UNDERSTANDING TRAIT INTERACTIONS IN SOYBEAN PLANTS

Peter Chimwanda[1] and Edwin Rupi[2]

[1]Department of Mathematics, Chinhoyi University of Technology, Chinhoyi, Zimbabwe.

[2]Department of Mathematics, Masvingo Teachers College, Masvingo, Zimbabwe.

[*]Corresponding Author's Email: pchimwanda@gmail.com

Cite this article:

Chimwanda, P., Rupi, E. (2025), A Multivariate Approach to Understanding Trait Interactions in Soybean Plants. African Journal of Mathematics and Statistics Studies 8(3), 66-72. DOI: 10.52589/AJMSS-MK4VFVIX

Manuscript History

Received: 3 Jun 2025

Accepted: 10 Jul 2025

Published: 25 Jul 2025

ABSTRACT: *Despite widespread use of statistical methods, advanced techniques like multivariate analysis are often underutilized, a trend that can lead to methodological missteps. This article focuses on Multivariate Analysis of Variance (MANOVA) and its necessary follow-up procedures, addressing why MANOVA often remains overlooked. We review its conceptual foundation, analysing multiple dependent variables collectively rather than separately, as in ANOVA, and illustrate its advantages, particularly the reduction of Type I error and the ability to detect multivariate patterns unnoticed in univariate analysis. Drawing on a practical case study involving a 2025 Kaggle soybean dataset (55,450 records across 13 agronomic traits), we apply MANOVA (via Jamovi) across 36 treatment combinations of genotype, salicylic acid, and water stress. Multivariate tests (Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's largest root) were all highly significant ($p < 0.001$), indicating group-level differences across variables. Subsequent univariate ANOVA revealed significance for each trait, and post-hoc pairwise comparisons (630 total) identified numerous significant differences. Finally, mean comparisons highlight the S1C3G3 group as top-performing across multiple key metrics. Our findings demonstrate the value of MANOVA in agricultural research and recommend adopting genotype 3 with salicylic acid at 450 mg under minimal water stress.*

KEYWORDS: Multivariate Analysis of Variance, Soybean, Genotype, Post-hoc Comparisons, Jamovi Statistical Software.

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 8, Issue 3, 2025 (pp. 66-72)

www.abjournals.org

INTRODUCTION TO MULTIVARIATE ANALYSIS

Despite the wide application of statistics in research and everyday situations, many advanced statistical techniques remain underutilized, even in scenarios where they would be most appropriate. Instead, simpler methods, which are more familiar to users, are often employed. This mismatch frequently leads to the misuse of common statistical techniques. This article focuses on Multivariate Analysis of Variance (MANOVA) along with follow-up tests that are carried out when MANOVA is significant.

Multivariate analysis in general refers to a broad and diverse range of techniques used to analyze multiple variables at once (Joseph et al., 2019). At its core, multivariate analysis examines more than one dependent variable simultaneously. This core definition rules out multiple regression, the most frequently used statistical method in statistical modelling, since it typically focuses on identifying the factors influencing a single dependent variable. Multivariate methods include discriminant analysis, factor analysis, hierarchical models, principal components analysis, canonical correlation, structural equation modelling, and multivariate analysis of variance, among others. While MANOVA does not explicitly assume a single underlying construct, users should be cautious when applying it if the dependent variables are unrelated and unlikely to yield similar results (Huberty & Petosky, 2000).

In many agricultural experiments, data is typically collected on multiple traits rather than just one. Soybean farming is a good example, where the variables of interest include grain yield and straw yield (Rajender & Lalmohan, 1997). Additional traits often recorded are plant height, number of green leaves, germination count, pod count, biological weight, protein content, seed yield, and others. However, the analysis is usually focused only on grain yield, and the best treatment is determined based solely on this trait. Straw yield, for example, in spite of its importance for purposes such as cattle feed, mulching, or manuring, is often overlooked in the analysis. Protein content, another important variable, is ignored along with many others.

Considering the agricultural system as a whole, straw yield as well as other important plant features should also be included in the evaluation. Similarly, in varietal trials, data is gathered on a range of plant characteristics and quality parameters. Yet, each trait is typically analyzed separately, and the best treatment or genotype is selected independently for each trait. In such cases, Multivariate Analysis of Variance (MANOVA) offers a more comprehensive approach by enabling simultaneous analysis of multiple traits.

Multivariate analysis of variance (MANOVA) is a generalization of ANOVA, allowing multiple dependent variables. It is interested in the examination of the effect of categorical variables on a combination of several theoretically related dependent variables, (Harris et al., 2012). The technique condenses all dependent variables into one weighted linear combination. It combines the dependent variables and analyses their collective variation with the independent variable. The linear combination of the individual dependent variables becomes the new dependent variable that is of interest. Essentially, MANOVA investigates whether the grouping variable explains significant variations in the combined dependent variables.

Many researchers have a limited understanding of MANOVA and its associated procedures, as evidenced by the common reliance on multiple linear regression, even for datasets that are better suited for MANOVA. This article aims to make MANOVA more accessible by presenting a detailed practical application. The rest of the article is structured into sections covering the rationale for using Multivariate Analysis of Variance and its application in

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

soybean farming, followed by the results and discussion, before concluding with recommendations.

Why Use Multivariate Analysis of Variance?

An alternative to using MANOVA is to perform separate ANOVAs for each dependent variable. However, this method has significant drawbacks: (a) running multiple ANOVA raises the risk of a Type I error, and (b) It does not evaluate whether the independent variable(s) are linked to combinations of the dependent variables. This limitation is particularly important for behavioral scientists, who often examine correlated dependent variables and benefit more from insights into their combined relationships.

MANOVA is applied to scenarios where there are multiple correlated dependent variables and the researcher wants to conduct a single comprehensive statistical test for the entire set, rather than carrying out several separate tests. The second, and often more significant purpose is to examine how independent variables affect the patterns of responses in the dependent variables. The third reason for doing MANOVA is that even when none of the individual ANOVAs produces a significant main effect on the dependent variables, in combination, the factor(s) might produce a significant effect, which suggests that the variables are more meaningful taken together than considered separately.

Even when all dependent variables are completely independent, conducting multiple tests increases the risk of inflated error. This issue becomes even more pronounced in ecological or biological studies, where variables are often interrelated and may exhibit strong actual or potential interactions, further compounding the error. In many such cases, where multiple ANOVAs have been performed, MANOVA would have been the more appropriate analytical approach.

Application on Soybean Dataset

The Kaggle 'Advanced Soybean Agricultural Dataset,' developed to support agricultural research and machine learning applications, was downloaded and analysed. Compiled in 2025 through a collaborative research initiative at the College of Agriculture, University of Tikrit, the dataset contains 55,450 records across 13 variables. It encompasses key agricultural metrics related to soybean plants, including plant height, pod count, biological weight, chlorophyll levels, protein content, seed yield, and leaf relative water content. These features are vital for assessing soybean crop growth, yield potential, and nutritional quality across varying environmental conditions. A key feature of this dataset is the Parameters column, which encodes essential experimental conditions affecting soybean growth. The letters in the Parameters column represent the following:

G: Refers to the genotype of soybean, consisting of six different genotypes; C: Represents salicylic acid, which has two levels (250 mg and 450 mg), along with a third level as a standard control; and S: Indicates water stress, which includes two levels: Water stress at 5% of field capacity and Water stress at 70% of field capacity. Since G has 6 levels, C has 3 levels and S has 2, this amounts to $6x3x2 = 36$ linear combinations altogether. This means that the independent variable, which is our factor, has 36 levels.

The Advanced Soybean Agricultural Dataset is structured to facilitate diverse analytical and predictive modelling applications, particularly in precision agriculture, yield prediction, and

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

crop health assessment. With its comprehensive set of features and randomized samples, this dataset is ideal for researchers, agronomists, and data scientists interested in agricultural optimization and decision-making. Multivariate analysis of variance was run in Jamovi and the results are discussed below.

RESULTS AND DISCUSSION

Table 1: Multivariate Tests

		value	F	df1	df2	p
Parameters	Pillai's Trace	10.00	5426	442	720395	< .001
	Wilks' Lambda	1.41e-13	15838	442	674054	< .001
	Hotelling's Trace	428	53589	442	720215	< .001
	Roy's Largest Root	264	429905	34	55415	< .001

All the four tests in Table 1 show p-values that are less than 0.001. This means that there is a significant difference between the group means across the combined dependent variables. In this case, there are 36 groups, from S1C1G1 to S2C3G6. At least one of the 36 levels of the independent variable had effects that were significantly different.

Table 2: Univariate Tests

	Dependent Variable	Sum of Squares	df	Mean Square	F	p
Parameters	Plant Height (PH)	466255.88	34	13713.40829	15095	< .001
	Number of Pods (NP)	1.05e0+7	34	310229.60720	1473	< .001
	Biological Weight (BW)	1.17e0+8	34	3.45e+6	12441	< .001
	Sugars (Su)	2725.29	34	80.15546	11982	< .001
	Relative Water Content in Leaves (RWCL)	344.08	34	10.11996	10787	< .001
	Chlorophyll A663	443397.87	34	13041.11395	124102	< .001
	Chlorophyll B649	73704.73	34	2167.78621	26912	< .001
	Protein Percentage (PPE)	139109.78	34	4091.46401	1226	< .001
	Weight of 300 Seeds (W3S)	1.28e0+6	34	37783.75037	20703	< .001
	Leaf Area Index (LAI)	22.49	34	0.66147	7286	< .001
	Seed Yield per Unit Area (SYUA)	7.72e+10	34	2.27e+9	13072	< .001

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 8, Issue 3, 2025 (pp. 66-72)

www.abjournals.org

Table 2: Univariate Tests

	Dependent Variable	Sum of Squares	df	Mean Square	F	p
	Number of Seeds per Pod (NSP)	2086.20	34	61.35895	2685	< .001
	Protein Content (PCO)	4254.66	34	125.13692	21009	< .001

Table 2 above is a collection of 13 univariate ANOVAs, one for each dependent variable. From the table, all 13 variables are significant, which means that each of them contributed significantly in making the linear combinations different across the levels of the independent variable.

The significance of the univariate ANOVAs meant that there was a need for further post hoc tests in order to establish which of the 36 levels of the independent variables were significantly different. A total of 630 pairwise comparisons were made. Results showed that most of the pairwise comparisons were significant. Those that were not significantly different included S1C1G1-S1C2G1, S1C1G2-S1C2G1, S1C1G2-S2C1G6, S1C1G5-S1C3G5, S1C1G5-S1C3G5 and S1C1G5-S2C3G6.

Table 3: Means of the Dependent Variables

Par	PH	NP	BW	SU	RWCL	CA	CB	PPE	W3S	LAI	SYUA	NSP	PCO
S1C1G1	50.3	141	67	0.163	0.729	1.25	3.13	35.5	**52.6**	0.08	6320	2.11	0.14
S1C3G1	54.5	160	102	0.594	0.724	9.03	2.2	36.9	29.7	0.0767	5594	**2.47**	0.333
S1C3G3	54.2	164	**235**	0.245	0.458	2.2	3.1	35.2	41.3	**0.0967**	**7641**	2.23	0.573
S2C3G4	41.7	**165**	121	0.613	0.47	1.1	3.37	36.1	34.2	0.09	4297	1.76	0.6
S2C1G5	50.4	148	150	**1.07**	0.669	5.6	1.17	34.6	36.3	0.09	3433	1.78	**1.55**
S2C3G2	**55.3**	146	86.3	0.163	0.611	8.23	2.4	35.6	35	0.09	3221	2.04	0.337
S2C3G3	49.7	128	126	0.318	0.621	**10**	2.3	37	34.4	0.04	5144	2.29	0.38
S2C3G5	51.4	129	133	0.196	**0.821**	1.27	3.33	**39.2**	43.4	0.0433	5491	2.05	0.353
S2C3G6	51.4	124	33.3	0.166	0.719	4.43	**7.6**	35.6	27.3	0.03	2271	2.13	0.32

For purposes of identifying the group that gave the best results, the means of each of the 36 groups and for each of the 13 variables were generated from Jamovi. The group with the highest mean for each variable was identified. Table 3 shows the groups that generated the highest means as well as the corresponding means. The abbreviations in the table are explained here. Parameters (PAR), Plant Height (PH), Number of Pods (NP), Biological Weight (BW), Sugars (Su), Relative Water Content in Leaves (RWCL), Chlorophyll A663 (CA), Chlorophyllb B649 (CB), Protein Percentage (PPE), Weight of 300 Seeds (W3S), Leaf Area Index (LAI), Seed Yield per Unit Area (SYUA), Number of Seeds per Pod (NSP), and Protein Content (PCO). The table shows only 9 out of the 36 rows. It is the 9 rows that carry the highest means of the 13 variables.

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

The table highlights the highest mean values for each variable in bold. As previously noted, only groups that achieved at least one maximum value are included. Specifically, Group S2C3G2 attained the tallest plant height at 55.3 cm, located in row 7, column 2 of Table 3.

Moreover, Table 3 reveals that some groups excelled in multiple variables. Group S1C3G3, for instance, recorded the highest values in Biological Weight (BW), Leaf Area Index (LAI), and Seed Yield per Unit Area (SYUA). Although this group did not achieve the top measurements for Plant Height (PH), Number of Pods (NP), or Number of Seeds per Pod (NSP), its figures for these variables were closely aligned with the highest recorded values.

Based on this comprehensive analysis, Group S1C3G3 emerges as the most outstanding among the 36 groups evaluated.

Table 4: Summary of Means

	PH	NP	BW	SU	RWCL	CA	CB	PPE	W38	LAI	SYUA	NSP	PCO
GROUP	232	134	133	215	235	233	236	235	111	133	133	131	215
VALUE	55.3	165	235	1.07	0.821	10	7.6	39.2	52.6	0.0967	7641	2.27	1.55

Table 4 provides a summary of the groups that achieved the highest values across 13 different variables. In the second row, group identifiers are presented in a condensed format, for instance, "232" corresponds to Group S2C3G2. The third row lists the maximum recorded values for each respective variable.

Taking column 2 as an example, which pertains to plant height, the group with the highest measurement is S2C3G2, recording a plant height of 55.3 cm.

CONCLUSION

Using a one-way MANOVA, the study confirmed that not all 36 soybean treatment groups (combinations of genotype, salicylic acid level, and water stress) were equivalent across 13 key agronomic and physiological traits. This signals a multivariate effect. Subsequently, thirteen individual ANOVAs revealed that, for each trait, at least one group differed significantly, a typical outcome following a significant MANOVA. In the pairwise comparisons (630 in total), the majority showed statistical significance. However, as higher numbers of comparisons increase Type I error risk, the MANOVA-first approach was essential to control false positives. Finally, the analysis of means identified Group S1C3G3, defined by Genotype 3, Salicylic Acid 3, Water Stress Level 1 as outperforming all others.

RECOMMENDATION

Based on these findings, the study strongly recommends adopting the S1C3G3 treatment combination under conditions similar to those tested. Practically, this translates to: growing genotype 3 soybeans, applying 3 salicylic acid, and rearing under mild water stress (level 1) conditions, to achieve significantly enhanced performance across critical traits.

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX

REFERENCES

Hair, J.F., Jr., Black, W. C., Babin, B. J., & Anderson, R.E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

Harris, J.E., Sheean, P.M., Gleason, P.M., Bruemmer, B., & Boushey, C. (2012). Publishing nutrition research: A review of multivariate techniques Part 2 Analysis of variance. Journal of the Academy of Nutrition and Dietetics, 112(1).

Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In Tinsley, Howard EA, and Steven D. Brown, (eds.). Handbook of applied multivariate statistics and mathematical modeling (pp. 183-208). Academic Press.

Joseph F. Hair, Jr., William C. Black, Barry J. Babin and Rolph E. Anderson (2019) Multivariate Data Analysis, EIGHTH EDITION, Cengage Learning.

Masreshaw Yirga, Afework Legesse. Multivariate Analysis Among Soybean (*Glycine max* L.) Genotypes in Southwest Ethiopia. *American Journal of Bioscience and Bioengineering*. Vol. 11, No. 2, 2023, pp. 20-26. doi: 10.11648/j.bio.20231102.12.

Rajender, R. & Lalmohan, L. (1997). Multivariance analysis of variance. Retrieved from https://www.researchgate.net/publication/237227650.

Article DOI: 10.52589/AJMSS-MK4VFVIX
DOI URL: https://doi.org/10.52589/AJMSS-MK4VFVIX