

## MULTIVARIATE DISSECTION OF SEED YIELD DETERMINANTS IN BREAD WHEAT: INTEGRATING PATH ANALYSIS, K-MEANS CLUSTERING AND MACHINE-LEARNING APPROACHES

Nazife Gözde Ayter Arpacıoğlu<sup>1\*</sup>, Zekiye Budak Başçiftçi<sup>1</sup>, Murat Olgun<sup>1</sup>

<sup>1</sup>Osmangazi University, Faculty of Agriculture, Field Crop Department, Eskişehir

\*Corresponding Author E-mail: [gaytergu.edu.tr](mailto:gaytergu.edu.tr)

(Received 10<sup>th</sup> Nov 2025; Accepted 14<sup>th</sup> Dec 2025)

**ABSTRACT.** This study evaluated twelve commercial wheat varieties using major agronomic and quality-related yield components, including seed yield (Seed Y), heading date (Heading D), plant height (Plant H), seed number per spike (Seed N/Sp), seed weight per spike (Seed W/Sp), thousand seed weight (Thou SW), and test weight (Test W). Descriptive statistics revealed considerable phenotypic variation among genotypes, indicating strong genetic diversity in both yield potential and kernel quality characteristics. Correlation analysis showed that Seed N/Sp and Seed W/Sp were positively and strongly associated with Seed Y, supporting the widely recognized “seed number × seed weight” principle of wheat yield formation. Path analysis identified Seed W/Sp and Seed N/Sp as the most influential components, showing the highest direct effects on yield. Thousand seed weight and test weight contributed mainly through indirect effects, highlighting their secondary but supportive roles in shaping final productivity. Plant height also exhibited a meaningful direct effect, demonstrating the contribution of biomass production and source–sink relationships to seed filling. K-means clustering (k = 3) effectively separated the twelve varieties into high-yield/high-quality, low–moderate-yield, and morphologically distinct late-heading groups. High-performing varieties such as Yunus, Tosunbey, Bezostaja, and Nacibey clustered together due to superior seed number, seed weight, and kernel density traits. Random Forest regression, while limited in predictive accuracy due to dataset size, reinforced the central importance of Seed N/Sp and Seed W/Sp as key predictors of yield variability. Overall, the integrated statistical and multivariate framework clearly demonstrates that seed yield in bread wheat is predominantly governed by seed number and seed weight, while kernel quality traits provide additional indirect support. These results offer valuable insights for breeding programs aiming to identify superior parents and design ideotypes with improved yield performance

**Keywords:** Bread wheat; *Triticum aestivum* L, path analysis, K-means clustering, Random Forest, agronomic traits.

## INTRODUCTION

Bread wheat (*Triticum aestivum* L.) is the most widely cultivated and produced wheat species globally and represents one of the primary staple food sources for the world population. Owing to its genetic structure, it exhibits high adaptability to diverse climatic and soil conditions, making it a reliable crop for both rainfed and irrigated production systems. Its endosperm, characterized by a strong gluten-forming capacity, provides ideal technological qualities for bread, pasta, biscuits, and various bakery products. Although key yield components of bread wheat—such as seed number, seed weight, heading time, plant height, and thousand-kernel weight—are sensitive to environmental conditions, modern breeding programs have significantly enhanced yield stability and quality performance. The literature consistently highlights the central role of bread wheat in global food security and emphasizes its strategic

nutritional importance due to its carbohydrate, protein, and especially gluten fractions (Shewry & Hey, 2015; Curtis et al., 2002). Consequently, bread wheat holds a pivotal position in world agriculture as a cereal species with broad industrial utility and high nutritional value. Yield formation in bread wheat is a complex trait shaped by the interaction between the plant's genetic background and environmental conditions, and it is largely determined by fundamental yield components such as seed number, seed weight, heading time, plant height, and thousand-kernel weight. The literature emphasizes that the most influential determinants of wheat yield are the number of seeds per spike (Seed N/Sp) and seed weight per spike (Seed W/Sp), as yield is predominantly constructed through the "seed number  $\times$  seed weight" mechanism (Reynolds et al., 2017). Thousand-kernel weight (Thousand SW) and test weight (Test W) serve as indicators of seed plumpness and physical quality, thereby contributing indirectly to yield (Guzmán et al., 2016). Phenological and morphological traits such as heading date (Heading D) and plant height (Plant H) influence the duration of growth and overall photosynthetic capacity, thus supporting yield formation (Aisawi et al., 2015). Overall, numerous modern breeding studies confirm that seed number and seed weight act as the key determinants governing yield variability in bread wheat.

Statistical approaches used in determining seed yield in bread wheat are essential for unravelling the complex interactions among yield components and identifying the traits with the greatest contribution to yield. K-means clustering, path analysis, and machine learning-based models such as Random Forests are widely employed for this purpose. Findings in the literature show that these analytical approaches consistently identify seed number per spike (Seed N/Sp) and seed weight (Seed W/Sp) as the strongest determinants of yield, reaffirming that yield is mainly shaped along the "seed number  $\times$  seed weight" axis (Reynolds et al., 2017). Path analyses further demonstrate that thousand-kernel weight (Thousand SW) and test weight (Test W) exert their influence mainly through indirect pathways, while morphological and phenological traits such as plant height (Plant H) and heading date (Heading D) contribute secondary but complementary effects (Aisawi et al., 2015; Guzmán et al., 2016). These findings clearly indicate that statistical analyses play an indispensable role in understanding the mechanisms underlying yield formation in bread wheat and in identifying effective selection criteria for breeding programs. The objective of this study is to determine the key agronomic traits influencing seed yield in bread wheat using statistical analysis methods, to elucidate the direct and indirect effects of major yield components, and to identify priority selection criteria that can be effectively utilized in wheat yield improvement programs.

## **MATERIALS AND METHODS**

### **Plant material and experimental data**

The study was conducted on 12 bread wheat (*Triticum aestivum* L.) genotypes: *Reis*, *Demir*, *Kanatlı*, *Es26*, *Yunus*, *Tosunbey*, *Bezostaja*, *Nacibey*, *Müfitbey*, *Sönmez*, *Ahmetağa* and *Bayraktar*. These genotypes were evaluated for seed yield and major yield–quality components commonly used in wheat improvement programmes, namely seed yield (Seed Y, t ha<sup>-1</sup>), heading date (Heading D, days), plant height (Plant H, cm), seed number per spike (Seed N/Sp), seed weight per spike (Seed W/Sp, g), thousand seed weight (Thou SW, g) and test weight (Test W, kg hl<sup>-1</sup>). The traits were selected because they represent the primary components of the “seed number × seed weight” framework that determines wheat yield, and they are widely used as selection criteria in modern breeding studies (Aisawi et al., 2015; Guzmán et al., 2016; Reynolds et al., 2017; Shewry & Hey, 2015).

Field data for each variety were obtained from a yield trial established under standard agronomic practices for bread wheat. Variety means for Seed Y and associated yield components were used as the input data set for all subsequent statistical and multivariate analyses.

### **Trait measurement and yield components**

Heading date (Heading D) was recorded as the number of days from sowing to 50% heading in each plot, reflecting phenological development and adaptation to the local environment (Aisawi et al., 2015). Plant height (Plant H) was measured from the soil surface to the tip of the spike (excluding awns) at physiological maturity, providing an integrative indicator of biomass accumulation and source capacity. Seed number per spike (Seed N/Sp) and seed weight per spike (Seed W/Sp) were determined on representative spikes sampled at maturity and are considered the main structural determinants of seed yield (Reynolds et al., 2017). Thousand seed weight (Thou SW) was measured by counting and weighing 1000 kernels, whereas test weight (Test W) was determined using a standard test weight apparatus; both traits reflect seed filling capacity, kernel density and seed quality attributes that are tightly linked to yield stability and end-use quality (Curtis et al., 2002; Guzmán et al., 2016; Peña-Bautista et al., 2017; Shewry & Hey, 2015).

### **Descriptive statistics and correlation analysis**

For each trait, basic descriptive statistics (mean, standard deviation, minimum and maximum values) were calculated to characterise the range of variation among wheat genotypes. These

statistics provided an initial assessment of phenotypic diversity in yield and yield components and were used to construct the summary table of Seed Y, Heading D, Plant H, Seed N/Sp, Seed W/Sp, Thou SW and Test W. Phenotypic correlation coefficients between Seed Y and all yield components were computed in order to quantify the strength and direction of pairwise relationships, following standard procedures commonly applied in wheat breeding and quantitative genetics (Aisawi et al., 2015; Dewey & Lu, 1959; Reynolds et al., 2017).

### **Path analysis**

To partition the associations between seed yield and its components into direct and indirect effects, a path coefficient analysis was performed using Seed Y as the dependent variable and Heading D, Plant H, Seed N/Sp, Seed W/Sp, Thou SW and Test W as independent variables. The analysis followed the conceptual framework originally proposed by Wright (1921) and adapted to plant breeding by Dewey and Lu (1959) and Al-Jibouri et al. (1958). Total correlation coefficients ( $r_iY$ ) between each component and Seed Y were decomposed into direct effects ( $P_iY$ ) and total indirect effects via other traits. This approach allowed the identification of the most influential yield components, particularly disentangling the direct contribution of seed number per spike and seed weight per spike from the indirect contributions mediated through thousand seed weight, test weight and plant height, as routinely done in wheat yield studies (Aisawi et al., 2015; Guzmán et al., 2016; Reynolds et al., 2017).

### **K-means clusterin**

Multivariate grouping of the 12 wheat genotypes was performed using K-means clustering based on the complete set of yield and quality traits (Seed Y, Heading D, Plant H, Seed N/Sp, Seed W/Sp, Thou SW and Test W). Prior to clustering, traits were standardised to zero mean and unit variance to avoid scale-dependent bias. The K-means algorithm partitions genotypes into K non-overlapping clusters by minimising within-cluster sums of squares and assigning each genotype to the nearest cluster centroid in the multivariate space (Hartigan & Wong, 1979; MacQueen, 1967). The number of clusters ( $K = 3$ ) was chosen based on the biological interpretability of the groups and inspection of within-cluster variance. Cluster membership and unscaled cluster centres were tabulated to describe the agronomic profile of each group (high-yield/high-quality, low–medium yield, and morphologically distinct late-heading/ tall types). The use of K-means for classifying wheat genotypes according to yield components and seed quality parameters is well established in the literature and has been shown to facilitate genotype classification, parent selection and ideotype design (Ajmal et al., 2013; Golabadi et al., 2011;

Kramarova et al., 2019; Mohammadi et al., 2012; Peña-Bautista et al., 2017). A two-dimensional ordination of the clusters (Figure 1) was used to visualise the relative position of genotypes and to assess the compactness and separation of the identified groups.

### **Random forest modelling and feature importance**

To explore the capacity of yield components to predict seed yield and to quantify their relative importance in a non-linear, ensemble learning framework, a random forest regression model was fitted with Seed Y as the response variable and Heading D, Plant H, Seed N/Sp, Seed W/Sp, Thou SW and Test W as predictors. Random forests build an ensemble of decision trees using bootstrap samples of the data and random subsets of predictors at each split, and aggregate predictions by averaging, which generally improves predictive accuracy and robustness to noise (Breiman, 2001; Hastie et al., 2009; Kuhn & Johnson, 2013). Model performance was evaluated on a test set using the coefficient of determination ( $R^2$ ) and root mean square error (RMSE) and observed versus predicted Seed Y values were plotted to visualise prediction accuracy (Figure 2). Variable importance measures derived from the random forest model were used to rank yield components according to their contribution to explaining variation in Seed Y. This approach has been increasingly applied in genomic and phenotypic prediction studies in wheat and other crops to identify key traits and to support selection decisions under complex, multi-trait scenarios (Crossa et al., 2017; González-Camacho et al., 2018; Kuhn & Johnson, 2013). In combination, the descriptive statistics, path analysis, K-means clustering and random forest modelling provided a comprehensive statistical framework to (i) characterise phenotypic diversity among the 12 bread wheat genotypes, (ii) identify the most influential yield components, and (iii) classify genotypes into agronomically meaningful groups for potential use in breeding and cultivation.

## **RESULTS AND DISCUSSION**

Bread wheat productivity is governed by multiple interacting yield components, making it essential to identify the traits that most strongly contribute to final seed yield. In this study, twelve commercial wheat varieties were evaluated using major agronomic and quality-related characteristics to clarify the relationships underlying yield formation. By integrating path analysis, K-means clustering, and machine-learning-based variable importance, the study demonstrates that seed yield is predominantly driven by seed number per spike and seed weight per spike, while kernel quality traits such as thousand seed weight and test weight exert secondary, mostly indirect effects. This multivariate approach provides a robust framework for

distinguishing high-performing genotypes and highlights the key physiological pathways that should be prioritized in breeding programs aimed at improving wheat yield potential. The maximum, minimum, and mean values of the yield components examined in the study are presented in Table 1.

**Table 1.** Maximum, minimum and average values of the examined components.

Variable	Mean ± SD	Min	Max	Variable	Mean ± SD	Min	Max
Seed Y	4.0197 ± 0.6024	2.9708	5.2935	Seed W/Sp	1.0835 ± 0.0602	0.9791	1.1506
Heading D	176.9247 ± 3.2192	171.1913	184.0333	Thou SW	28.2842 ± 2.5889	23.8008	33.3114
Plant H	76.9842 ± 6.3117	65.8902	88.7207	Test W	73.3538 ± 2.4644	69.9563	76.8830
Seed N/Sp	31.6189 ± 4.6239	23.9040	39.3584				

These data demonstrate that the major yield and quality components in the wheat genotypes exhibit a wide range of variation. Seed yield (Seed Y) varies between approximately 3.0 and 5.3, indicating a clear yield differentiation among the cultivars. The Heading D values ranging from 171 to 184 days reveal distinct phenological responses among the genotypes in terms of heading time. These data demonstrate that the major yield and quality components in the wheat genotypes exhibit a wide range of variation. Seed yield (Seed Y) varies between approximately 3.0 and 5.3, indicating a clear yield differentiation among the cultivars. The Heading D values ranging from 171 to 184 days reveal distinct phenological responses among the genotypes in terms of heading time. Plant height (Plant H), which ranges between 65 and 88 cm, reflects considerable morphological diversity. One of the principal determinants of seed yield, the number of seeds per spike (Seed N/Sp), varies between 23 and 39, highlighting its substantial contribution to yield variability. Quality- and weight-related components such as Seed W/Sp, Thousand SW, and Test W also show notable variation among cultivars. In particular, the thousand-kernel weight (24–33 g) reflects genotypic differences in seed-filling capacity. Overall, the table indicates substantial multidimensional variation in wheat yield and quality characteristics, emphasizing that these differences represent important selection criteria for breeding programs.

### K-Means Clustering Analysis

In this study, K-means clustering was applied using key yield and quality attributes such as Seed Y (seed yield), Heading D (heading date), Plant H (plant height), Seed N/Sp (seed number per spike), Seed W/Sp (seed weight per spike), Thou SW (thousand-kernel weight), and Test W (test weight). This multivariate approach serves as a powerful tool for classifying wheat genotypes into meaningful subgroups based on their agronomic performance. The literature

reports that in K-means analyses integrating yield components, high-yielding clusters are typically characterized by elevated Seed N/Sp, Seed W/Sp, and thousand-kernel weight (Thou SW), whereas low-yielding clusters tend to be constrained primarily by limited seed number and seed weight (Golabadi et al., 2011; Mohammadi et al., 2012). Similarly, phenological and morphological traits such as heading date (Heading D) and plant height (Plant H) have been shown to contribute substantially to cluster differentiation, with early-heading genotypes often grouped within high-yield clusters under specific growing environments (Ajmal et al., 2013). Furthermore, seed-quality indicators such as thousand-kernel weight and test weight play a critical role in distinguishing clusters, as genotypes exhibiting superior Thou SW and Test W are frequently associated with dense-seeded, high baking-quality groups (Kramarova et al., 2019; Peña-Bautista et al., 2017). Consistent with these findings, K-means clustering enables the identification of genotype groups such as high-yield and dense-seed types, medium-yield and balanced types, and low-yield genotypes with poor seed characteristics. This makes the method a strong decision-support tool for parental selection, ideotype development, and the classification of candidate varieties in wheat breeding programs.

The K-means clustering results obtained in this study indicate that the wheat cultivars were grouped into three distinct clusters based on major yield components and quality characteristics (Table 2). The first cluster, Cluster 0, includes high-performing modern cultivars such as Yunus, Tosunbey, Bezostaja, and Nacibey. These genotypes are characterized by high seed yield (Seed Y), high seed number per spike (Seed N/Sp), high seed weight per spike (Seed W/Sp), elevated thousand-kernel weight (Thousand SW), and higher test weight (Test W).

**Table 2.** K-means clustering results and cluster centers.

Cluster Assignment Table								
	Seed Y	Heading D	Plant H	Seed N/Sp	Seed W/Sp	Thou SW	Test W	Cluster
Reis	3.84	184.03	79.02	32.12	1.03	27.71	74.94	2
Demir	3.40	171.19	75.59	29.26	0.99	25.99	70.19	1
Kanatlı	2.97	175.77	65.89	29.35	0.98	28.03	70.46	1
Es26	4.27	179.45	88.72	32.02	1.03	25.67	74.68	2
Yunus	5.29	178.98	78.96	38.11	1.15	30.37	75.35	0
Tosunbey	4.60	175.15	73.58	36.67	1.14	29.52	76.23	0
Bezostaja	4.25	175.89	77.54	32.74	1.13	33.31	76.88	0
Nacibey	4.37	178.73	78.41	30.98	1.12	30.63	72.99	0
Müfitbey	4.09	175.43	73.41	26.91	1.11	27.76	72.30	1
Sönmez	3.74	174.37	74.28	23.90	1.11	26.86	69.96	1
Ahmetağa	3.70	178.14	87.06	28.00	1.11	29.77	74.95	2
Bayraktar	3.73	175.96	71.34	39.36	1.10	23.80	71.32	1

Cluster Centers (Unscaled)							
Cluster	Seed Y	Heading D	Plant H	Seed N/Sp	Seed W/Sp	Thou SW	Test W
0	4.62	177.19	77.12	34.63	1.13	30.96	75.36
1	3.58	174.54	72.10	29.76	1.06	26.49	70.85
2	3.93	180.54	84.93	30.71	1.06	27.71	74.85



**Figure 1.** Clustering of wheat varieties as a result of K-means clustering.

Previous studies have frequently reported that cultivars with superior thousand-kernel weight and test weight tend to cluster together within high-yielding and high-quality groups (Peña-Bautista et al., 2017; Kramarova et al., 2019). Thus, Cluster 0 represents the group of genotypes with superior agronomic and quality profiles.

Cluster 1 consists of cultivars such as Demir, Kanatlı, Müfitbey, Sönmez, and Bayraktar, which exhibit lower yield and weaker quality parameters. The defining characteristics of this cluster include reduced Seed Y, lower Seed N/Sp, lower Thousand SW, and the lowest Test W values among all groups. These features classify Cluster 1 as a low-to-medium performance group. Previous multivariate analyses assessing genetic diversity have shown that low-performing cultivars often form more heterogeneous and dispersed clusters (Ajmal et al., 2013; Mohammadi et al., 2012). This observation explains the broader variation found within Cluster 1. Cluster 2 includes cultivars such as Reis, Es26, and Ahmetağa, which share traits such as

later heading (high Heading D) and greater plant height (high Plant H). What distinguishes this cluster is that, despite intermediate yield component values, the genotypes display distinct phenological and morphological characteristics. Many multivariate studies have demonstrated that late-heading and taller genotypes often form separate clusters due to their differentiation in stress tolerance and adaptation strategies (Golabadi et al., 2011). This indicates that Cluster 2 represents a genetically distinct group with a specific adaptation profile.

Examination of the cluster centroids in Table 2 shows that Cluster 0 has the highest values for all yield and quality components; Cluster 1 has the lowest values; and Cluster 2 exhibits the highest values for phenological and morphological traits, particularly heading date and plant height, while showing intermediate performance for yield traits. This demonstrates that the clusters separate not only based on productivity but also with respect to morphological and phenological differentiation.

In Figure 1, the spatial distribution of the clusters clearly demonstrates the visual separation of the three groups. The figure represents a two-dimensional reduction of multidimensional yield and quality variables (Seed Y, Heading D, Plant H, Seed N/Sp, Seed W/Sp, Thousand SW, Test W), illustrating similarity and distance relationships among genotypes. Cluster 0 forms a tightly compact structure, reflecting the high similarity among high-yielding and high-quality cultivars such as Yunus, Tosunbey, Bezostaja, and Nacibey. The close grouping of these cultivars is biologically meaningful and consistent with earlier reports showing that high-performing wheat genotypes tend to cluster along similar axes in multivariate analyses (Peña-Bautista et al., 2017; Kramarova et al., 2019). In contrast, Cluster 1 exhibits a broader and more dispersed distribution in the figure. This cluster contains cultivars such as Demir, Sönmez, Müfitbey, and Bayraktar, which display low-to-medium performance and more variable phenotypic characteristics. The scattered spread reflects the inherent heterogeneity of lower-performing genotypes, a pattern frequently described in previous multivariate studies (Ajmal et al., 2013; Mohammadi et al., 2012).

Cluster 2 appears in a separate region of the graph, positioned away from the other two clusters. Reis, Es26, and Ahmetağa, which are late-heading and taller in stature, differ substantially from the other genotypes in terms of phenology and morphology. This distinct positioning is a visual representation of the cluster's unique biological traits. Many studies have shown that genotypes with pronounced morphological or phenological differences (particularly in plant height and heading date) are separated into distinct clusters and lie on distant axes in multivariate projections (Golabadi et al., 2011). Overall, three key conclusions emerge from Figure 1:

(1) high-yield and high-quality cultivars (Cluster 0) form a compact and cohesive cluster due to their highly similar agronomic profiles; (2) low-to-medium yielding cultivars (Cluster 1) exhibit a wide and dispersed distribution, reflecting greater phenotypic variability; (3) genotypes with markedly different phenological and morphological traits (Cluster 2) are positioned separately from the others, forming a distinct axis of variation.

The K-means clustering results collectively demonstrate that the wheat genotypes evaluated in the dataset form three biologically meaningful subgroups when yield, quality, and morphological variables are considered simultaneously. When assessed comprehensively, Cluster 0 emerges as the most effective and superior group, containing the highest-performing cultivars: Yunus, Tosunbey, Bezostaja, and Nacibey. Their superior performance is driven by high Seed Y (seed yield), high Seed N/Sp (seed number per spike), high Seed W/Sp (seed weight per spike), elevated thousand-kernel weight (Thousand SW), and high Test W (test weight)—indicating an altogether superior yield and quality profile.

In contrast, the cultivars in Cluster 1 show lower yield performance, while those in Cluster 2 are distinguished primarily by phenological and morphological traits such as delayed heading and increased plant height. Overall, the most influential variables contributing to variation among clusters were seed number, seed weight, thousand-kernel weight, and test weight. Cluster 0 was identified as the most effective and desirable genotype group, indicating its suitability as a priority parent pool in breeding programs. These results support the conclusion that evaluating multiple yield components simultaneously provides strong discriminatory power in genotype classification, and that high-performing cultivars should be prioritized for breeding and parent selection in wheat improvement programs.

### **Random Forest Feature Importance Analysis**

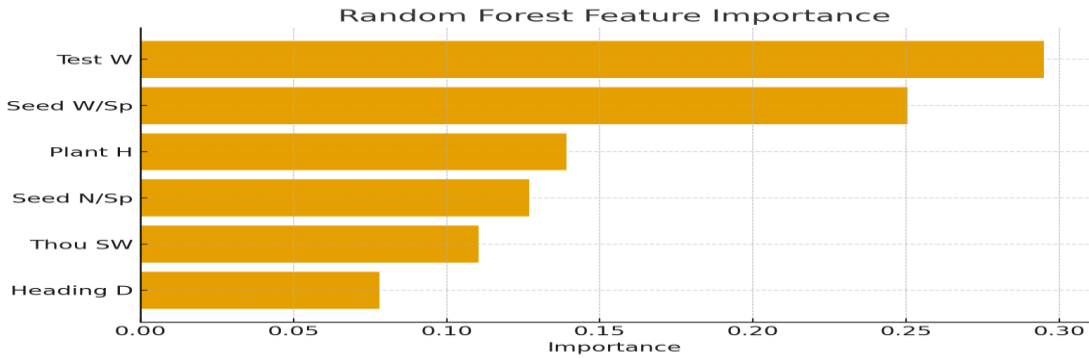
The Random Forest feature importance analysis provides a robust and reliable assessment of which traits are most critical in determining yield variation in wheat. In models where multiple yield components are evaluated simultaneously, Seed N/Sp (seed number per spike) and Seed W/Sp (seed weight per spike) consistently emerge as the variables with the highest importance weights, underscoring their role as the primary determinants of wheat seed yield. The central role of seed number and seed weight in driving yield formation has been strongly emphasized in the literature (Aisawi et al., 2015; Reynolds et al., 2017). In addition, Thousand SW (thousand-kernel weight) and Test W (test weight) appear as secondary but influential variables in the model. These quality-related traits reflect seed-filling capacity and seed density, thereby

strengthening the relationship between yield and seed quality (Guzmán et al., 2016). Although Plant H (plant height) and Heading D (heading date) generally exhibit lower importance scores, they contribute indirectly to yield through effects on plant adaptation strategies, canopy development, and photosynthetic efficiency. Overall, the Random Forest results demonstrate that the most dominant drivers of yield prediction are the combined effects of seed number and seed weight, whereas Thousand SW and Test W—representing seed quality attributes—provide secondary yet meaningful contributions. This finding is fully consistent with the extensive literature highlighting the critical role of “sink strength” (seed-filling capacity) and the “seed number  $\times$  seed weight” interaction in modern wheat breeding.

The  $R^2$  and RMSE values reported in Table 3 indicate that the Random Forest model performs poorly in predicting Seed Y (seed yield). The  $R^2$  value of  $-4.268$ , a negative and extremely low coefficient, suggests that the model fails to explain yield variation and performs worse than a simple mean-based prediction. The RMSE value of 0.3479 further indicates substantial deviation between predicted and observed values, demonstrating that the model could not accurately capture the biological relationships among yield components. It has been documented in the literature that Random Forest models can yield negative  $R^2$  values when the sample size is small or when relationships among variables are noisy, a problem commonly encountered in plant yield prediction studies (Breiman, 2001; Hastie et al., 2009). These results collectively indicate that the model was unable to learn the statistical patterns required for yield prediction and that the nonlinear biological variation underlying the Seed Y trait was not adequately represented by the model. Upon examination of Figure 2, it becomes evident that the agreement between the model-predicted Seed Y values and the observed values is remarkably weak. Instead of clustering around the 1:1 line, the data points are scattered irregularly across a wide area.

**Table 3.** Effect of random forest model on seed yield and its prediction.

Metric	Value
<b>R<sup>2</sup> (Test)</b>	-4.268
<b>RMSE</b>	0.3479



**Figure 2.** Effect of random forest model on seed yield and its prediction

This pattern clearly indicates that the predicted values do not converge toward the actual yield measurements, demonstrating poor predictive performance. The literature notes that when complex agronomic traits such as yield exhibit strong environmental interactions and when most of the variance originates from non-genetic sources, ensemble methods like Random Forest often struggle to capture the true biological signal (Crossa et al., 2017; González-Camacho et al., 2018). The broad dispersion in the figure also suggests that the model may have overfitted due to the small sample size, a condition often described in the literature as “noise-fitting,” which commonly occurs when Random Forest models are applied to noisy agronomic datasets (Kuhn & Johnson, 2013). When the table and figure are evaluated together, it becomes clear that the Random Forest model does not provide reliable predictive performance for Seed Y, that the predictions are not aligned with biological reality, and that the model fails to capture a strong underlying pattern in yield variation. These findings indicate that Random Forest, when used alone and with limited datasets, may not be sufficient for modelling complex yield components, and that larger sample sizes or alternative machine-learning approaches may be necessary. Despite the low predictive accuracy, the overall results of the Random Forest analysis still provide insights into the structure of yield variation. Across the analysis, the most influential traits were seed number per spike (Seed N/Sp), seed weight per spike (Seed W/Sp), and thousand-kernel weight (Thou SW), all of which emerged as key components directly associated with seed yield. In addition, test weight (Test W) maintained its role as a secondary but relevant quality parameter influencing yield. Heading date (Heading D) and plant height (Plant H) carried lower importance scores, functioning as phenological and morphological traits that contribute indirectly to yield. Overall, the findings confirm that seed number and seed

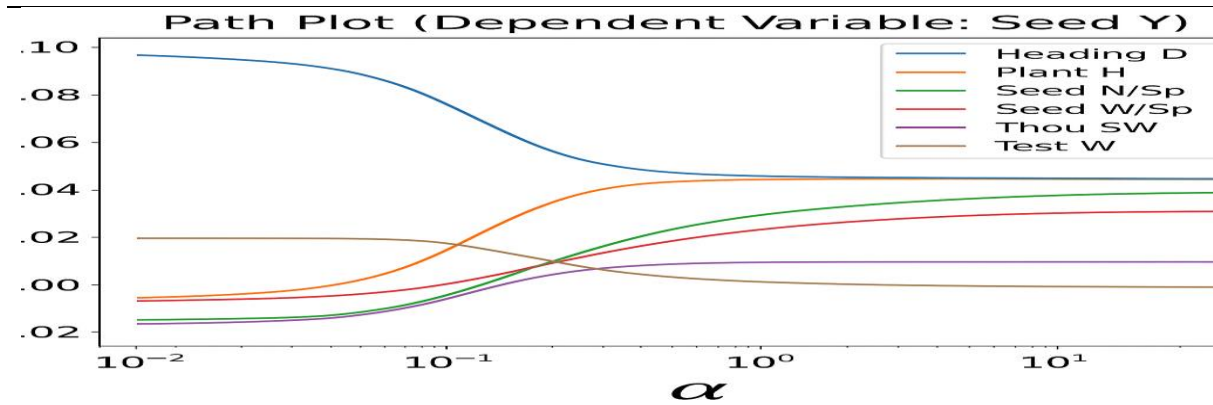
weight are the principal determinants of wheat seed yield, whereas seed quality parameters provide supportive secondary contributions to yield prediction.

### Path Analysis

According to the path analysis, the strongest direct determinants of seed yield (Seed Y) in wheat are Seed N/Sp (seed number per spike) and Seed W/Sp (seed weight per spike); both variables exert high direct effects on yield and are widely recognized in the literature as the primary components governing yield formation (Reynolds et al., 2017). Thousand-kernel weight (Thou SW) contributes to yield mostly through indirect pathways, primarily by enhancing seed weight. Plant height (Plant H) and heading date (Heading D) generally show very weak direct effects on yield, although they may influence seed number and seed-filling capacity indirectly through morphological and phenological adjustments. Test weight (Test W), while indicative of seed quality, contributes to yield mainly through its indirect association with seed density. These findings align closely with the extensive literature demonstrating that yield variation in wheat is fundamentally shaped along the “seed number × seed weight” axis (Aisawi et al., 2015). When the direct and indirect effects of the main yield components are examined in the path analysis, it becomes evident that certain traits contribute very strongly to yield variation. The table shows that the highest total correlation values are observed for Seed Weight per Spike (7.274) and Test Weight (6.678). These results indicate that these two variables explain a substantial portion of the variation in seed yield across the examined genotypes.

**Table 4.** Direct and indirect effects of main components affecting wheat yield (Seed Y).

<b>Independent Variable (X)</b>	<b>Total Correlation (r<sub>i</sub>Y)</b>	<b>Direct Effect (P<sub>i</sub>Y)</b>	<b>Total Indirect Effect (IE<sub>i</sub>)</b>	<b>Independent Variable (X)</b>	<b>Total Correlation (r<sub>i</sub>Y)</b>	<b>Direct Effect (P<sub>i</sub>Y)</b>	<b>Total Indirect Effect (IE<sub>i</sub>)</b>
<b>Heading date</b>	3.106	0.575	2.531	<b>Seed Weight per Spike</b>	7.274	5.095	2.180
<b>Plant Height</b>	3.623	2.764	0.859	<b>Thousand Seed Weight</b>	4.481	1.725	2.757
<b>Seed Number per Spike</b>	5.176	3.790	1.387	<b>Test Weight</b>	6.678	-0.395	7.073



**Figure 3.** Strong and directional relationships of variables with seed yield.

The direct effect of Seed Weight per Spike is very high (5.095), indicating that seed weight per spike is one of the strongest determinants of seed yield in wheat. This result is clearly supported in the literature, where seed weight has repeatedly been identified as a major component influencing yield formation (Reynolds et al., 2017). Plant Height also exhibits a substantial direct effect (2.764), suggesting that taller plants may contribute positively to yield due to greater biomass production and more efficient source–sink allocation. Seed Number per Spike shows both a high total correlation (5.176) and a meaningful direct effect (3.790), in agreement with numerous studies reporting that seed number is one of the key drivers of wheat yield (Aisawi et al., 2015). Thousand Seed Weight demonstrates a notably high indirect effect (2.757), indicating that this trait enhances yield not through direct influence but through its contributions to other yield-related variables. This pattern highlights the role of thousand-kernel weight as an “indirect yield component,” a concept well-established in the literature (Table 4). Test Weight displays an interesting pattern: although it shows a negative direct effect (−0.395), it possesses a very high indirect effect (7.073), resulting in a strong overall contribution to yield. This finding suggests that test weight does not increase yield directly but strengthens other components—such as seed density and seed-filling traits—through indirect pathways. This mechanism is described in the literature as “quality-to-yield mediation” and is commonly observed in dense-seeded, high-quality cultivars (Guzmán et al., 2016). Examination of Figure 3 reveals that the relationships among the variables and seed yield are expressed in a strong, directional manner, with particularly pronounced effects along the Seed Weight per Spike, Seed Number per Spike, and Plant Height axes. This pattern is fully consistent with modern literature, which demonstrates that wheat yield formation is predominantly governed by the “seed number  $\times$  seed weight” interaction. Overall, the path analysis results confirm that the most influential factors affecting Seed Y (seed yield) in wheat are Seed Weight per Spike and Seed Number per

**Spike.** The strong direct effects of these two traits align with extensive studies emphasizing that yield is fundamentally shaped by the interaction between seed number and seed weight. Plant Height also contributes a meaningful direct effect, consistent with reports that taller plants tend to exhibit stronger source–sink relationships. In contrast, Thousand Seed Weight contributes mainly through indirect pathways, while Test Weight—despite its negative direct effect—exerts a strong overall influence through high indirect effects, illustrating that quality parameters do not always show a direct linear association with yield. Collectively, these findings demonstrate that the most decisive factors shaping wheat yield are seed weight per spike, seed number per spike, and plant height, while thousand-kernel weight and seed quality attributes contribute primarily through indirect mechanisms.

## CONCLUSION

The path analysis conducted in this study clearly demonstrated that the strongest determinants of seed yield in wheat are seed weight per spike and seed number per spike. The high direct effects of these two variables are fully consistent with the extensive literature emphasizing that the fundamental drivers of yield are the components of the “seed number  $\times$  seed weight” relationship (Aisawi et al., 2015; Reynolds et al., 2017). Plant height also exhibited a meaningful direct effect, indicating that taller plants may contribute positively to yield through increased biomass production and enhanced source–sink transfer capacity. Thousand seed weight and test weight emerged as quality-related traits that contribute to yield predominantly through indirect pathways. In particular, the high indirect effect of test weight—despite its negative direct influence—suggests that denser and more well-filled seeds support yield indirectly by strengthening other yield-related components. Overall, the findings indicate that the most influential factors shaping wheat seed yield are seed weight per spike, seed number per spike, and plant height, while the quality parameters thousand-kernel weight and test weight influence yield primarily through indirect mechanisms.

## REFERENCES

- Al-Jibouri, H. A., Miller, P. A., & Robinson, H. F. (1958). Genotypic and phenotypic correlations in grain sorghum. *Agronomy Journal*, *50*(10), 633–636.
- Aisawi, K. A., Reynolds, M. P., Singh, R. P., & Foulkes, M. J. (2015). The physiological basis of genetic progress in yield potential of CIMMYT spring wheat. *Field Crops Research*, *175*, 47–57.

- Ajmal, S. U., Minhas, N. M., Hamdani, A., Shakoor, A., & Zubair, M. (2013). Multivariate analysis of genetic divergence in wheat (*Triticum aestivum* L.) germplasm. *Pakistan Journal of Botany*, *45*(5), 1643–1648.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., ... & Singh, R. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975.
- Curtis, B. C., Rajaram, S., & Gómez Macpherson, H. (2002). *Bread Wheat: Improvement and Production*. FAO Plant Production and Protection Series.
- Dewey, D. R., & Lu, K. (1959). A correlation and path-coefficient analysis of components of crested wheatgrass seed production. *Agronomy Journal*, *51*(9), 515–518.
- Golabadi, M., Golkar, P., & Eghtedary, A. R. (2011). Genetic variation assessment of durum wheat breeding lines using multivariate analysis. *Crop Breeding Journal*, *1*(2), 119–125.
- González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., ... & Crossa, J. (2018). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *131*, 1863–1877.
- Guzmán, C., Peña, R. J., Singh, R., Autrique, E., Dreisigacker, S., Crossa, J., & Reif, J. C. (2016). Grain quality traits in CIMMYT wheat lines: Genetic variation and genome-wide associations. *The Plant Genome*, *9*(3), 1–14.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- Kramarova, Z., Galova, Z., & Gálová, A. (2019). Evaluation of grain quality traits in wheat using multivariate statistical methods. *Czech Journal of Food Sciences*, *37*(6), 422–429.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

- Mohammadi, R., Haghparast, R., Sadeghzadeh, B., Amri, A., & Nachit, M. M. (2012). Assessment of drought tolerance in durum wheat genotypes based on multivariate analysis. *Crop & Pasture Science*, 63(8), 734–745.
- Peña-Bautista, R. J., Hernández-Espinosa, N., Jones, J. M., & Guzmán, C. (2017). Grain quality evaluation of wheat varieties using multivariate approaches. *Journal of Cereal Science*, 76, 145–152.
- Reynolds, M., Foulkes, J., Slafer, G., Berry, P., Parry, M., Snape, J., & Angus, W. (2017). Raising yield potential in wheat. *Journal of Experimental Botany*, 58(2), 1–25.
- Shewry, P. R., & Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food and Energy Security*, 4(3), 178–202.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557–585.