

CLASSIFICATION OF WINTER WHEAT GENOTYPES
BASED ON GRAIN YIELD COMPONENTS
USING K-MEANS CLUSTERING
IN MULTI-ENVIRONMENT TRIALS

Georgi Raykov 

Received on April 9, 2026

Presented by H. Najdenski, Member of BAS, on June 30, 2026

Abstract

Understanding the interaction between grain yield and its main components is essential for improving wheat productivity and guiding breeding strategies. In this study, K-means clustering was applied to classify 36 winter wheat genotypes evaluated in multi-environment trials across four locations in Bulgaria over four growing seasons. The analysis was based on grain yield (GY), number of grains m^{-2} (NGM), and thousand grain weight (TGW), representing key processes of yield formation.

The optimal number of clusters was determined using the Elbow method and Silhouette coefficient, which identified $k = 2$ as the most robust solution, while $k = 4$ provided additional biological resolution. The two-cluster classification revealed a clear separation of genotypes into contrasting yield formation strategies defined by the trade-off between grain number and grain weight.

The four-cluster solution further refined genotype classification, identifying intermediate and high-performing groups, including genotypes combining high grain number and high yield. This suggests that the classical trade-off between yield components can be partially overcome.

Overall, the results confirm the dominant role of grain number in yield determination, while highlighting the compensatory contribution of grain weight. The application of K-means clustering provides a robust framework for identifying biologically meaningful genotype groups and supports its use as a practical

tool for selection in wheat breeding programmes under variable environmental conditions.

Key words: K-means clustering, grain yield, yield components, grain number, grain weight, wheat breeding

Introduction. Grain yield in wheat (*Triticum aestivum* L.) is a complex trait determined by the interaction among several components, including number of productive tillers, grains per spike, number of grains m^{-2} (NGM), and thousand grain weight (TGW) [1]. These components are hierarchically structured, with upstream traits influencing yield primarily through their effect on number of grains m^{-2} , which acts as an integrative variable. Grain number is widely recognized as the main determinant of yield, reflecting sink capacity established during pre-anthesis development [2–4], whereas grain weight is largely influenced during grain filling by assimilate availability and environmental conditions [5, 6]. A characteristic feature of yield formation is the partial compensation between grain number and grain weight, resulting from source–sink limitations. Consequently, similar yield levels may be achieved through alternative strategies based either on increased grain number or improved grain filling [7, 8]. Recent studies confirm that this trade-off remains a central constraint in wheat productivity, particularly under variable environmental conditions, where the balance between sink capacity and assimilate supply determines final yield [3, 4, 9]. Moreover, advances in physiological research indicate that although partial decoupling of this relationship is possible, the compensation between grain number and grain weight continues to constrain their simultaneous improvement, particularly under source–sink limitations during grain filling [4, 9]. Despite well-established physiological relationships, the combined expression of yield components across genotypes and environments often leads to complex patterns that are difficult to interpret using conventional statistical approaches. Multivariate techniques such as principal component analysis provide dimensionality reduction but do not explicitly classify genotypes. In contrast, K-means clustering enables grouping based on multivariate similarity without prior assumptions [10–12], and has been increasingly applied in recent studies to identify patterns in crop performance, genotype classification, and yield-related traits under diverse environmental conditions [13, 14].

The objective of this study was to identify and interpret yield formation strategies in winter wheat by classifying genotypes based on grain yield, number of grains m^{-2} , and thousand grain weight using K-means clustering under multi-environment conditions.

Material and methods. Experimental design and plant material. The study was based on multi-environment trials (MET) conducted across four locations in Bulgaria over four growing seasons (2017, 2018, 2019, and 2021), resulting in a total of 16 environments (4 locations $\times$ 4 years), representing contrasting agro-ecological conditions typical of major wheat-producing regions.

Multi-environment testing is widely used to evaluate genotype performance and to quantify genotype $\times$ environment ($G \times E$) interactions [15].

The experimental sites included Paskalevo (Dobrich), Trastenik (Rousse), Straldzha (Yambol), and the experimental field of the Agricultural University in Plovdiv, covering diverse climatic conditions characterized by variability in temperature and precipitation. These conditions provide a suitable framework for assessing genotype performance under contrasting environments.

A total of 36 winter wheat (*Triticum aestivum* L.) genotypes were evaluated. Genotypes were tested across all environments within a multi-environment trial structure, allowing the assessment of genotype performance and the analysis of genotype $\times$ environment interactions.

Data structure and trait evaluation. The dataset comprised multi-environment observations of these genotypes, evaluated for grain yield (GY), thousand grain weight (TGW), and number of grains m^{-2} (NGM). These variables were selected as key indicators of yield formation, capturing both sink capacity (NGM) and grain filling processes (TGW), while avoiding redundancy associated with upstream traits. NGM was considered an integrative trait reflecting the combined influence of productive tillers and grains per spike, and the inclusion of only these variables ensured a robust clustering structure by minimizing multicollinearity.

For each environment, genotype mean values were used for the analysis. Prior to clustering, data were standardized using Z -score transformation:

$$Z = (X - \mu) / \sigma,$$

where X is the observed value, μ is the mean, and σ is the standard deviation. This transformation eliminates scale effects and ensures equal contribution of variables to the clustering procedure [11].

Clustering analysis. K-means clustering was applied to classify genotypes based on their performance for GY, TGW, and NGM, with the aim of identifying distinct yield formation strategies. The algorithm partitions observations into k clusters by minimizing the within-cluster sum of squares (WCSS), thereby maximizing homogeneity within clusters and separation between clusters [10].

The optimal number of clusters was determined using the Elbow method and the Silhouette coefficient. The Elbow method evaluates the reduction in WCSS with increasing k , while the Silhouette coefficient quantifies cluster compactness and separation, with higher values indicating more distinct clustering [16].

Post hoc mixed-model analysis was conducted to test whether the clusters identified by the unsupervised classification differed significantly for grain yield and yield-related traits.

All analyses were conducted using the R statistical environment [17].

Results and discussion. Determination of the optimal number of clusters. The Elbow method showed a sharp decrease in within-cluster sum of

squares (WCSS) up to $k = 3$, followed by a gradual reduction, with a secondary inflection point at $k = 4$. The Silhouette coefficient reached its maximum at $k = 2$ (0.447), with a secondary local maximum at $k = 4$ (0.438), while lower values were observed for higher cluster numbers, indicating reduced cluster separation. These results suggest that $k = 2$ provides the most distinct partition of the data, whereas $k = 4$ offers a more detailed classification with slightly reduced separation. Therefore, both clustering solutions were retained for further analysis. This approach is widely recommended for validating clustering solutions in multivariate datasets and remains a standard procedure in cluster analysis and unsupervised classification [16, 18, 19].

Clustering of genotypes based on yield formation traits ($k = 2$). K-means clustering ($k = 2$) separated the 36 wheat genotypes into two distinct groups based on grain yield (GY), thousand grain weight (TGW), and number of grains m^{-2} (NGM) (Fig. 1). The two clusters exhibited contrasting yield formation patterns, reflecting alternative physiological strategies.

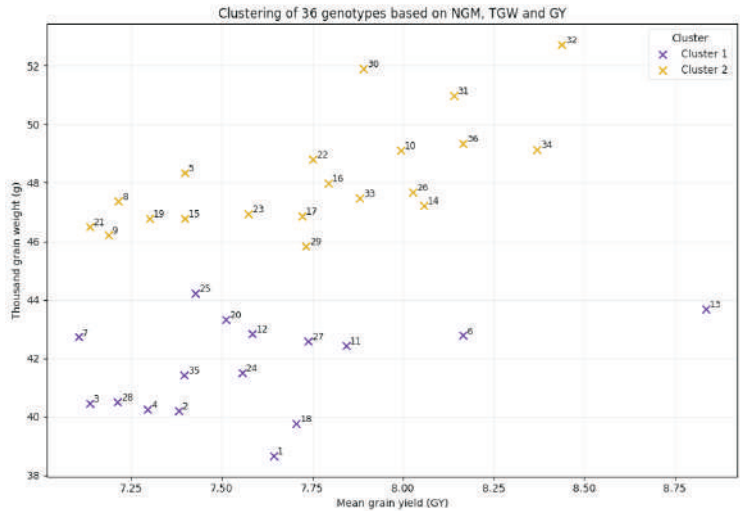


Fig. 1. Two-dimensional visualization of K-means clusters ($k = 2$) based on grain yield (GY) and thousand grain weight (TGW). Cluster assignment was derived from multivariate classification using grain yield (GY), thousand grain weight (TGW), and number of grains m^{-2} (NGM)

Cluster 1 was characterized by significantly higher NGM (18153.69) and lower TGW (41.71), whereas Cluster 2 showed lower NGM (16038.32) and higher TGW (48.20) (Table 1). Despite the reduced grain number, Cluster 2 exhibited significantly higher grain yield (7.76 t ha^{-1}) compared to Cluster 1 (7.60 t ha^{-1}), indicating that enhanced grain filling compensated for the lower sink capacity.

T a b l e 1

Cluster characteristics based on NGM, TGW and GY ($k = 2$)

Cluster	Number of genotypes	NGM	TGW	GY
1	16	18153.69 a	41.71 b	7.60 b
2	20	16038.32 b	48.20 a	7.76 a

Note: Different letters indicate significant differences at $p \leq 0.05$ (Tukey-adjusted LMM; cluster fixed, environment random)

These results demonstrate a clear trade-off between grain number and grain weight, consistent with source-sink relationships in wheat [3,8]. This relationship has been extensively confirmed under high-yielding conditions, where increasing grain number often results in reduced grain weight due to assimilate limitations [4, 9].

Clustering of genotypes based on yield formation traits ($k = 4$). K-means clustering ($k = 4$) further refined the classification of genotypes, separating them into four distinct groups based on grain yield (GY), thousand grain weight (TGW), and number of grains m^{-2} (NGM) (Fig. 2). This higher-resolution partition revealed more detailed variation in yield formation strategies compared to the two-cluster solution, as increasing the number of clusters allows the identification of more subtle structure within multivariate datasets [11,12].

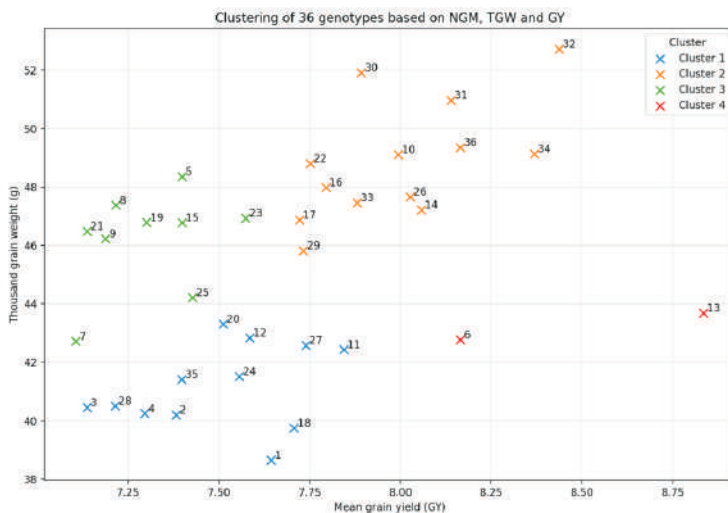


Fig. 2. Two-dimensional visualization of K-means clusters ($k = 4$) based on grain yield (GY) and thousand grain weight (TGW). Cluster assignment was derived from multivariate classification using grain yield (GY), thousand grain weight (TGW), and number of grains m^{-2} (NGM)

Clear differences among clusters were observed for all traits (Table 2). Cluster 4 was characterized by the highest NGM (19555) and the highest grain yield (8.5 t ha⁻¹), indicating a highly productive sink-driven strategy. In contrast, Cluster 2 exhibited the highest TGW (48.85) combined with high grain yield (8.0 t ha⁻¹), reflecting a grain filling-driven strategy.

T a b l e 2

Cluster characteristics based on NGM, TGW and GY (*k* = 4)

Clusters	Number of genotypes	NGM	TGW	GY
1	9	18167 b	41.16 c	7.5 b
2	12	16319 c	48.85 a	8.0 a
3	13	15774 c	46.21 b	7.3 c
4	2	19555 a	43.23 c	8.5 a

Note: Means sharing a letter are not significantly different at *p* ≤ 0.05 (Tukey-adjusted LMM; cluster fixed, environment random).

Clusters 1 and 3 showed intermediate and lower performance, with Cluster 1 characterized by relatively high NGM but lower TGW, and Cluster 3 by lower values for both traits, resulting in reduced yield. These patterns suggest that high yield can be achieved through different trait combinations, but the most favourable genotypes are those combining either very high grain number or efficient grain filling.

Overall, the four-cluster solution highlights the heterogeneity of yield formation strategies and identifies genotype groups with contrasting combinations of grain number and grain weight. Cluster 4 exhibited the highest number of grains m⁻² (NGM) and the highest grain yield, suggesting a favourable combination of high sink capacity and productive performance. However, because this cluster contained only two genotypes, the observed pattern should be interpreted with caution and requires validation in larger germplasm collections. Nevertheless, the results are consistent with previous studies indicating that the combination of high sink capacity and efficient grain filling represents an important strategy for achieving high yield potential under variable environmental conditions [3, 4, 20].

Conclusions. The application of K-means clustering to multi-environment data of 36 winter wheat genotypes revealed clear and biologically meaningful patterns of yield formation. Grain yield was achieved through distinct physiological strategies defined by the balance between grain number and grain weight. The comparison between clustering resolutions showed that the two-cluster solution captured the principal contrast between yield components, whereas the four-cluster solution provided a more detailed representation of genotypic variability. This higher resolution enabled the identification of superior genotypes combining high yield with favourable trait combinations. Notably, the detection of genotypes exhibiting both high grain number and high yield suggests that the

classical trade-off between yield components may be partially mitigated. These findings confirm that K-means clustering is an effective exploratory tool for classifying genotypes and disentangling complex trait interactions in multi-environment studies. The approach provides a robust framework for identifying yield formation patterns and supports more targeted selection of high-performing genotypes under variable environmental conditions.

Acknowledgments. The author would like to express sincere gratitude to Prof. Dr. Nikolay Tsenov for providing the experimental dataset and for his contribution to the design and implementation of the multi-environment trials on which this study is based.

REFERENCES

- [1] CARRERA C. S., R. SAVIN, G. A. SLAFER (2023) Critical period for yield determination across grain crops, *Trends Plant Sci.*, **29**(3), 329–342, <https://doi.org/10.1016/j.tplants.2023.08.012>.
- [2] FISCHER R. A. (1985) Number of kernels in wheat crops and the influence of solar radiation and temperature, *J. Agric. Sci.*, **105**(2), 447–461, <https://doi.org/10.1017/S0021859600056495>.
- [3] SLAFER G. A., R. SAVIN, V. O. SADRAS (2014) Coarse and fine regulation of wheat yield components in response to genotype and environment, *Field Crops Res.*, **157**, 71–83, <https://doi.org/10.1016/j.fcr.2013.12.004>.
- [4] XIE Q., D. L. SPARKES (2021) Dissecting the trade-off of grain number and size in wheat, *Planta*, **254**(1), 3, <https://doi.org/10.1007/s00425-021-03658-5>.
- [5] WARDLAW I. F., C. BLUMENTHAL, O. LARROQUE et al. (2002) Contrasting effects of chronic heat stress and heat shock on kernel weight and flour quality in wheat, *Funct. Plant Biol.*, **29**(1), 25–34, <https://doi.org/10.1071/PP00147>.
- [6] BRINTON J., C. UAUY (2019) A reductionist approach to dissecting grain weight and yield in wheat, *J. Integr. Plant Biol.*, **61**(3), 337–358, <https://doi.org/10.1111/jipb.12741>.
- [7] MIRALLES D. J., G. A. SLAFER (2007) Sink limitations to yield in wheat: How could it be reduced?, *J. Agric. Sci.*, **145**(2), 139–149, <https://doi.org/10.1017/S0021859607006752>.
- [8] SADRAS V. O. (2007) Evolutionary aspects of the trade-off between seed size and number in crops, *Field Crops Res.*, **100**(2–3), 125–138, <https://doi.org/10.1016/j.fcr.2006.07.004>.
- [9] BUSTOS D. V., A. K. HASAN, M. P. REYNOLDS et al. (2013) Combining high grain number and weight through a DH-population to improve grain yield potential of wheat in high-yielding environments, *Field Crops Res.*, **145**, 106–115, <https://doi.org/10.1016/j.fcr.2013.01.015>.
- [10] MACQUEEN J. (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 281–297.

- [11] KAUFMAN L., P. J. ROUSSEEUW (2009) Finding Groups in Data: An Introduction to Cluster Analysis, Hoboken, John Wiley & Sons.
- [12] JAIN A. K. (2010) Data clustering: 50 years beyond K-means, Pattern Recognit. Lett., **31**(8), 651–666, <https://doi.org/10.1016/j.patrec.2009.09.011>.
- [13] MARINO S., A. ALVINO (2021) Vegetation indices data clustering for dynamic monitoring and classification of wheat yield crop traits, Remote Sens., **13**(4), 541, <https://doi.org/10.3390/rs13040541>.
- [14] BONNEU F., D. MAKOWSKI, J. JOLY et al. (2024) Machine learning based on functional principal component analysis to quantify the effects of the main drivers of wheat yields, Eur. J. Agron., **159**, 127254, <https://doi.org/10.1016/j.eja.2024.127254>.
- [15] YAN W., M. S. KANG (2002) GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists, Boca Raton, CRC Press.
- [16] ROUSSEEUW P. J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., **20**, 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [17] R Core Team (2024) R: A Language and Environment for Statistical Computing, Vienna, R Foundation for Statistical Computing, <https://www.r-project.org/>.
- [18] ARBELAITZ O., I. GURRUTXAGA, J. MUGUERZA et al. (2013) An extensive comparative study of cluster validity indices, Pattern Recognit., **46**(1), 243–256, <https://doi.org/10.1016/j.patcog.2012.07.021>.
- [19] CHARRAD M., N. GHAZZALI, V. BOITEAU et al. (2014) NbClust: An R package for determining the relevant number of clusters in a data set, J. Stat. Softw., **61**(6), 1–36, <https://doi.org/10.18637/jss.v061.i06>.
- [20] REYNOLDS M. P., M. J. FOULKES, G. A. SLAFER et al. (2009) Raising yield potential in wheat, J. Exp. Bot., **60**(7), 1899–1918, <https://doi.org/10.1093/jxb/erp016>.

Department of Crop Science, Agricultural University - Plovdiv,

12 Mendeleev Blvd, 4000 Plovdiv, Bulgaria

e-mail: georgi.raikov.dzi@gmail.com, <https://orcid.org/0009-0000-4564-7279>