Spanish Journal of Agricultural Research
21 (1), e0901, 13 pages (2023)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2023211-19835
INIA-CSIC

RESEARCH ARTICLE

OPEN ACCESS

Use of multivariate analysis and machine learning methods to characterize traits contributing to wheat yield diversity

Ali BEHPOURI, Sara FAROKHZADEH, Zahra ZINATI* and Zobeir KHOSRAVI

Department of Agroecology, College of Agriculture and Natural Resources of Darab, Shiraz University, Iran.

*Correspondence should be addressed to Zahra Zinati: zahrazinati@shirazu.ac.ir

Abstract

Aim of study: Regarding the third largest staple food crop in the world, determining the factors affecting wheat yield is of great importance. This study aimed to determine useful subsets of agronomic traits and evaluate the order of importance of traits in grain yield.

Area of study: Fars province, Iran.

Material and methods: In total, the data corresponding to 22 agronomic traits was collected from six different regions (Darab, Kavar, Marvdasht, Fasa, Lar, and Khonj) of 90 farms of Fars province, Iran as the most important wheat-growing regions. Multivariate statistical analysis (correlation, stepwise regression, and principal component analysis (PCA)) and machine learning modeling approaches, such as partial least squares regression (PLSR) and support vector regression (SVR) models, were applied to agronomic traits.

Main results: The findings, based on integrated approaches such as correlation, stepwise regression, and PCA, highlighted that number of spikes m^{-2}, grain number $spike^{-1}$, and thousand-grain weight had a major impact on the yield followed by awn length, spike length, narrow leaf herbicide, broadleaf herbicide, time to plant maturity (month), and soil salinity. Besides, PLSR with nine inputs (nine selected traits) displayed better prediction capability (R^2=85 %, RMSE=0.32, MSE=0.10, and BIAS=-0.05) than that with all twenty-two input traits.

Research highlights: Integrated multivariate statistical analyses and machine learning regression methods could be a powerful tool in determining traits that have a significant impact on yield. These achievements can be considered for future breeding programs.

Additional key words: *Triticum aestivum*; multivariate statistical analysis; partial least squares regression; support vector regression.

Abbreviation used: MSE (mean squared error); PCA (principal component analysis); PLSR (partial least squares regression); RMSE (root-mean-square error); SVR (support vector regression); TGW (thousand-grain weight); TOL (tolerance); VIF (variance inflation factor).

Citation: Behpouri, A; Farokhzadeh, S; Zinati, Z; Khosravi, Z (2023). Use of multivariate analysis and machine learning methods to characterize traits contributing to wheat yield diversity. Spanish Journal of Agricultural Research, Volume 21, Issue 1, e0901. https://doi.org/10.5424/sjar/2023211-19835.

Received: 16 Sep 2022. **Accepted:** 31 Jan 2023.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Following rice and maize, wheat (*Triticum aestivum* L.) is the third key food crop in the world and it is farmed in a variety of environments. Wheat yield is affected by agronomic, phenological, physiological, and climatic factors, nutrient storage, crop management such as fertilizer amount and infection to pests and diseases, land management, and land conditions (Farokhzadeh et al., 2020). Also, soil properties such as soil texture, percentage/concentration of nitrogen (N), potassium (K), phosphorus (P), percentage of organic matter, and electrical conductivity

(EC) affect wheat yield (Asseng et al., 2001; Takahashi & Anwar, 2007).

So far, many studies have been conducted to reveal the relationship between yields and other related traits and consequently identify the traits affecting the growth and development of wheat (Zhang et al*.,* 2016b; Farokhzadeh et al*.,* 2020). For instance, in a study conducted by Yang et al*.* (2022), according to a two-site multi-cultivar test via principal component analysis (PCA), structural equation model, and partial least squares model, 11 phenotypes were considered as the key phenotypes that contributed most to grain yield, including spike density, leaf area index, biomass, harvest index, net photosynthetic rate, leaf chlorophyll, canopy temperature, carboxylation efficiency, stomatal conductance, leaf nitrate reductase, and transpiration rate. Norouzi et al*.* (2010) applied artificial neural networks to predict dryland wheat yield in semi-arid and mountainous areas of western Iran. They stated that the sediment transport index was the most significant topographic factor in the yield of wheat. Barikloo et al*.* (2017) evaluated the performance of a neuro-genetic hybrid model to predict wheat yield based on land characteristics. Sensitivity analysis indicated that soil parameters such as available phosphorus, total nitrogen, gravel content, organic matter percentage, and soil reaction play the main role in determining wheat yield. They found that total soil organic matter and nitrogen had the highest and lowest correlations with the yield quantity and quality of wheat respectively. In addition, they argued that some of the chemical and physical properties of soil such as nitrogen content had an impact on soil fertility and water storage in the soil, which are the major factors in wheat yield.

Characterizing the traits that contributed most to yield diversity, could provide the basis for developing cultivars with high yields. Using multivariate analysis including PCA, stepwise regression and machine learning not only may extract the concealed pattern present in data but also facilitate ways of determining notable traits (Farokhzadeh et al*.,* 2021).

The PCA is a usual procedure to condense a larger set of correlated variables into smaller and effectively interpretable axes of variation. The PCA technique can help us to understand the main data structure and develop a smaller number of uncorrelated variables. The objective of the principal component analysis was to determine the highest variance with the lowest number of components possible (Farokhzadeh et al*.,* 2022).

In recent decades, various yield models, including linear and non-linear models, have been applied to modeling linear and non-linear relationships among variables. Yield modeling not only allows us to predict plant production but also contributes to understanding how the yield is affected by environmental factors and yield components.

Generally, partial least squares regression (PLSR) is accepted as one of the methods with highest efficiency in terms of extracting and creating reliable model to predict chemical composition in sunflower seed (Fassio & Gozzolino, 2003), predict grain yield in maize farm through drought tolerance traits (Shaibu & Adnan*,* 2015), examine growth of rice leaf and nitrogen level (Nguyen & Lee, 2006), determine the factors in rice yield and the yield of fields of winter wheat (Zhang et al*.,* 2020), and determine the priority of the factors in winter wheat yield (Hu et al*.,* 2018). PLSR, as an effective method, combines multiple linear regression and PCA to transform the data matrix efficiently and alleviate the collinearity issue of independent variables (Costa et al*.,* 2012). The support vector machine (SVM) was first introduced by Vapnik et al. (1995). It has gained much popularity as a machine learning tool in classification and regression analysis called SVM classification and support vector regression (SVR), respectively. The SVR technique provides users with high flexibility of underlying variables distribution, the relationship between the independent and dependent variables, and the control on the penalty term (Hu et al., 2018; Zhang et al., 2020). Similar to PLSR, SVR has been used in crop research including agricultural drought prediction (Tian et al., 2018), yield prediction in pepper (Wilson et al., 2021), and predicting the mass of ber fruits (Abdel-Sattar et al., 2021), but there are few studies regarding the application of SVR in crop yield prediction.

PLSR is also highly recommended for analyzing an immense array of pertinent predictor variables with a sample size that is not large enough in comparison with the number of independent variables (Carrascal et al*.,* 2009) and SVR has the capacity to process the data of high dimensionality and is less affected by sample size *(*Meng & Zhao, 2015). In addition to the aforementioned and also sample size in the present study, these two machine learning methods were used to fit the yield prediction model based on agronomic traits. Besides, to our best knowledge, no study has directly compared the performance of the aforementioned two methods, PLSR and SVR, in developing models for predicting wheat grain yield.

On the other hand, the relationship between agronomic traits and yield may be due to the influence of the environment. However, few studies examined traits related to yield based on data from multiple locations. For instance, in Gustavo et al.´s (2022) study, a large database included 367 papers published compiled to recognize the main determinants of the number of grains per unit in response to environmental and genetic factors. They suggested that the responsiveness of the number of grains per unit area was similarly explained by changes in both the number of spikes m^{-2} and the number of grains $spike^{-1}$. To fill this gap, in the current study, a comprehensive investigation of key traits that contribute to yield was conducted on 22 agronomic traits collected from six different regions of 90 farms to minimize the effect of the environment on the relationships between traits and yield. Multivariate statistical analyses were used to clarify and assess the underlying determinants of wheat yield diversity. In addition, two machine learning

Table 1. Descriptive statistics of the model dataset (total sample) used in the study.

Traits	Variable	Maximum	Minimum	Mean	Std Deviation	Correlation coefficients of traits with grain yield
Seeding rate (kg ha^{-1})	x1	350.00	220.00	285.17	3.95	0.228*
Awn length (cm)	x2	12.00	4.00	8.13	1.87	0.353**
Spike length (cm)	x3	13.00	5.00	10.59	1.35	0.316**
Plant height (cm)	x4	98.00	70.00	83.61	6.94	0.452**
Nitrogen fertilizer (N, kg ha^{-1})	x5	350.00	100.00	210.22	61.83	0.381**
Phosphorus fertilizer (P, kg ha^{-1})	x6	150.00	0.00	71.28	44.64	0.441**
Potassium chloride fertilizer (K, kg ha^{-1})	x7	100.00	0.00	33.17	30.99	0.254*
Narrow leaf herbicide	x8	2.00	0.00	1.34	0.47	0.278**
Broadleaf herbicide	x9	2.00	0.00	1.05	0.52	0.368**
Time to plant maturity (month)	x10	7.50	5.00	6.35	0.66	0.392**
Number of irrigation cycles	x11	10.00	5.00	7.16	1.34	0.230*
Animal manure application	x12	3000.00	0.00	111.11	507.13	0.375**
Pest infestation (%)	x13	18.00	0.00	5.76	3.20	-0.276**
Disease infestation (%)	x14	15.00	0.00	4.61	3.56	-0.433**
Number of weeds m^{-2}	x15	20.00	3.00	8.60	3.95	-0.302**
Rainfall (mm)	x16	104.70	66.60	93.32	12.87	-0.159
Planting depth (cm)	x17	5.00	2.00	3.41	0.68	-0.299**
Soil salinity (dS m^{-1})	x18	3.50	0.50	1.63	0.81	-0.585**
Number of spikes m^{-2}	x19	402.00	220.00	277.74	34.86	0.915**
Grain number spike^{-1}	x20	66.00	27.00	41.46	8.49	0.841**
Thousand grain weight (g)	x21	45.00	31.00	37.66	3.67	0.798**
Grain yield (t ha^{-1})	x22	8.00	3.80	5.66	0.91	1.00

* and **: Significant (α= 5%), and highly significant (α= 1%), respectively.

methods, PLSR, as a linear model, and SVR, as a non-linear model, were applied to assess the predicting power of two models of the relationship between 21 traits with grain yield as well as to assess selected traits related to yield.

Material and methods

Data collection

The data were collected from six different regions (Darab, Kavar, Marvdasht, Fasa, Lar, and Khonj) of 90 farms in Fars province, Iran, as the most important wheat-growing regions during 2020-2021. Repeated random sampling was performed on every farm using a 1-m^2 quadrat to measure the agronomic traits, including grain yield (t ha^{-1}), thousand seed weight (g), number of spikes m^{-2}, grain number spike^{-1}, awn length (cm), spike length (cm), plant height (cm), number of weeds m^{-2}, pest and disease infestations percentage. Soil EC (dS m^{-1}) was measured with an EC meter for a slurry consisting of 1:5 (w/v) soil/distilled water (Bao et al*.,* 2005). Rainfall (mm) was obtained from the weather stations. Other traits including the nitrogen fertilizer (N, kg ha^{-1}), phosphorus fertilizer (P, kg ha^{-1}), potassium chloride fertilizer (K, kg ha^{-1}), animal manure application, number of irrigation cycles, seed rate (kg ha^{-1}), use of herbicides (narrow leaf-herbicide and broadleaf herbicide), time to plant maturity (month) and planting depth (cm) were collected using a questionnaire on each farm. It should be noted that samples using a quadrat of 1 m^2 were taken in each farm. Then the pest damaged and diseased wheat plants were counted, separately and the percentage of disease and pest infestations were calculated via count of: (pest damaged or diseased plant count/total plant counts) × 100.

Multivariate statistical analysis

The Shapiro-Wilk test is a statistical test that was used to check if a variable follows a normal distribution. Pearson's correlation coefficient was used to measure the degree of linearity of the relationship between two variables.

Table 2. Stepwise regression analysis of grain yield as dependent and other traits as independent variables in wheat.

Traits	Unstandardized coefficients		Standardized coefficients	F	Partial R^2	R^2	Entering into model, respectively
	B	Std. error	Beta				
(Constant)	-2.603	0.573	-	-	-	-	-
Number of spikes m^{-2}	0.008	0.002	0.313	452.40**	0.8371	0.8371	1
Grain number $spike^{-1}$	0.030	0.006	0.278	15.37**	0.0245	0.8616	2
Spike length (cm)	0.064	0.023	0.095	6.52**	0.0098	0.8713	3
Broadleaf herbicide	0.162	0.062	0.092	4.88**	0.0070	0.8783	4
Soil salinity (dS m^{-1})	-0.184	0.052	-0.162	5.11**	0.0070	0.8853	5
Time to plant maturity (month)	0.209	0.055	0.152	10.01**	0.0123	0.8976	6
Narrow leaf herbicide	0.209	0.065	0.107	5.06*	0.0059	0.9036	7
Thousand-grain weight (g)	0.042	0.015	0.167	4.94*	0.0055	0.9091	8
Rainfall (mm)	0.007	0.002	0.095	5.75*	0.0061	0.9152	9
Awn length (cm)	0.048	0.02	0.098	6.05*	0.0060	0.9212	10

* and **: Significant (α= 5%), and highly significant (α= 1%), respectively; B: unstandardized coefficients; R^2: coefficient of determination.

Stepwise regression was applied to identify the worthiest effective features on grain yield in the regression model. Through a stepwise regression analysis, grain yield was considered as a dependent variable while the rest of the traits were considered as independent variables.

Since there was a correlation among independent variables, the multicollinearity test was performed to test regression assumptions through the computation of TOL (tolerance) and VIF (variance inflation factor) using SPSS software. The VIF index was smaller than 10, which indicates the absence of multicollinearity between variables. Also, the TOL index, which was greater than 0.1, indicates that there was no multicollinearity between variables.

The PCA was accomplished to distinguish new variables (principal components) containing the trait combinations that include the most variation. In this study, an association of some important traits with grain yield was estimated using PCA.

Data analysis in SAS (Statistical Analysis System v. 9.2) was used to check the normal distribution of data (Shapiro-Wilk test) and also to perform stepwise regression analysis and descriptive statistics, and in SPSS (Statistical Package for the Social Sciences, v. 24) for Pearson's correlation analysis. Pheatmap, factoextra, and ggplot2 packages in RStudio (v. 4.0.3) were used to plot correlation heatmap and PCA bi-plot.

Machine learning methods

Machine learning model analysis, including PLSR and SVR analysis, were conducted using PYTHON (multi-paradigm programming language, v. 3.10.5) for prediction and EXCEL software for statistical analyses and graphs drawing. For this aim, 80% of samples were used for the training stage, and the rest 20% of samples, for testing stage. Mathematical background on PLSR and SVR is as follows:

— SVR. The principles used in SVR are identical to those used in SVM classification. SVR is a modified form of SVM in which instead of categorized dependent variables, numerical ones are used. SVR enables optimal interpretation of the resulting model since it allows non-linear model construction without altering the explanatory variable outlines. Pertinent to SVR is the implementation principle of maximal margin, which allows its description as a convex optimization problem. Hence the method can proceed with predictions as long as the error does not exceed a certain set value. Moreover, regression is not over-fitted as the SVR method allows it to be penalized using a cost parameter.

— PLSR. This is a robust, efficient regression method for multivariate analysis over a wide data range (Martens & Martens, 2000). This technique reduces predictors to a smaller set of non-correlated components for least square regression. PLSR is uniquely useful for analyzing highly collinear predictors or for data that predictors exceed observations and the ordinary least square regression method would have failed completely or yielded coefficients with high standard errors. Additionally, PLSR outperforms the traditional regression method as it employs a linear multivariate model to correlate two data matrices, X and Y, and further progresses to model the corresponding structures. The ability of this technique to further analyze mul-

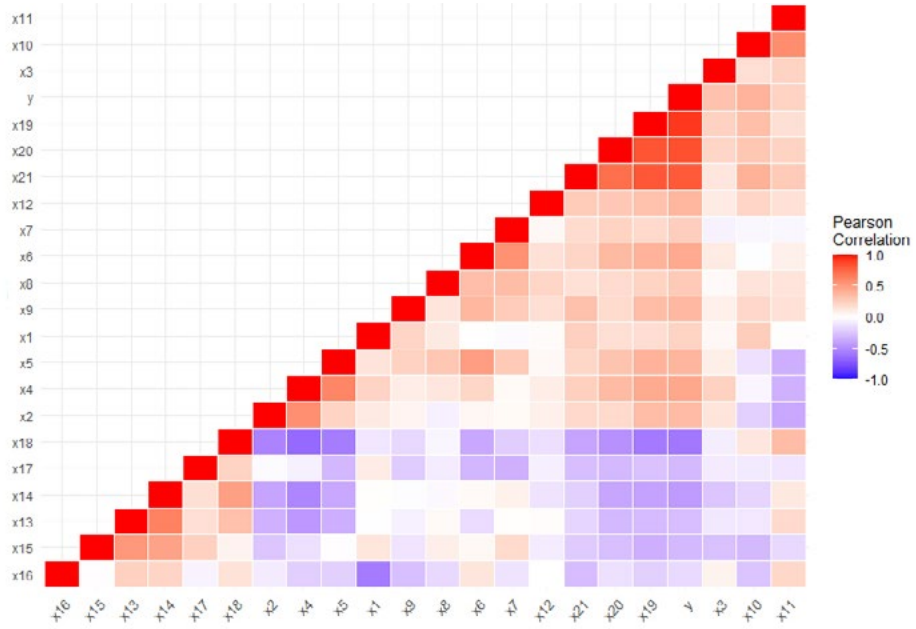

Figure 1. Heatmap correlation matrix among 22 agronomic traits: x1, seeding rate (kg ha^{-1}); x2, awn length (cm); x3, spike length (cm); x4, plant height (cm); x5, nitrogen fertilizer (n; kg ha^{-1}); x6, phosphorus fertilizer (P; kg ha^{-1}); x7, potassium chloride fertilizer (K; kg ha^{-1}); x8, narrow-leaf herbicide; x9, broadleaf herbicide; x10, time to plant maturity (month); x11, number of irrigation cycles; x12, animal manure application; x13, pest infestation (%); x14, disease infestation (%); x15, number of weeds m^{-2}; x16, rainfall (mm); x17, planting depth (cm); x18, soil salinity (ds m^{-1}); x19, number of spikes m^{-2}; x20, grain number spike^{-1}; x21, thousand-grain weight (g); and x22, grain yield (t ha^{-1}).

tiple, noisy, collinear, and incomplete variables in X and Y (Wold et al., 2001) describes its uniqueness and efficiency. PLSR theory, principles, and application have been extensively reviewed by Abdi (2010).

Model performance

The four statistical evaluation criteria were used to assess the model performance, including the coefficient of determination (R^2), the root-mean-square error (RMSE), the mean squared error (MSE), and BIAS for the training and testing datasets. These statistical indexes were calculated as:

$$R^2 = \frac{(n\sum_{i=1}^{n} X_{Pi}X_{mi} - \sum_{i=1}^{n} X_{Pi}\sum_{i=1}^{n} X_{mi})^2}{[n\sum_{i=1}^{n} X_{pi}^2 - (\sum_{i=1}^{n} X_{Pi})^2][n\sum_{i=1}^{n} X_{mi}^2 - (\sum_{i=1}^{n} X_{mi})^2]}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{pi} - X_{mi})^2}{n}}$$

$$MSE = \frac{\sum_{i=1}^{n}(X_{pi} - X_{mi})^2}{n}$$

$$BIAS = \frac{\sum_{i=1}^{n} X_{pi} - X_{mi}}{n}$$

where X_{mi} is the measured yield value in the field, X_{pi} is the predicted yield value, and n is the number of samples. A combination of the mentioned statistical parameters is sufficient for model evaluation. Sheikh Khozani et al. (2020) used these indices to examine the performance of models.

Results

The Shapiro-Wilk test of data collected for different traits indicated a normal distribution for variables. Statistical information related to 22 different traits is presented in Table 1.

Multivariate statistical analysis

Correlation

The phenotypic correlations (r_p) between grain yield and different traits are listed in Table 1. Also, heatmap correlation matrix among 22 agronomic traits is presented in Fig. 1. The correlation coefficients (r) of significant correlations ranged from 0.228 to 0.915. Grain yield indicated significant positive correlations with the majority of traits except for pest and disease infestations percentage, num-

Table 3. The results of the principal component analysis for different traits in wheat.

Traits	PC1	PC2	PC3	PC4	PC5	PC6
Seeding rate (kg ha^{-1})	0.103	0.046	0.022	0.585	0.010	0.052
Awn length (cm)	0.183	-0.320	-0.104	0.002	0.250	-0.071
Spike length (cm)	0.121	0.070	-0.188	-0.153	0.074	0.503
Plant height (cm)	0.257	-0.312	-0.026	0.077	0.110	0.246
Nitrogen fertilizer (N, kg ha^{-1})	0.238	-0.235	0.254	-0.008	-0.171	0.197
Phosphorus fertilizer (P, kg ha^{-1})	0.196	0.064	0.399	-0.233	-0.087	0.125
Potassium chloride fertilizer (K, kg ha^{-1})	0.122	0.063	0.451	-0.083	-0.207	-0.086
Narrow leaf herbicide	0.116	0.138	0.273	0.075	-0.097	0.521
Broadleaf herbicide	0.165	0.156	0.166	0.123	-0.224	-0.105
Time to plant maturity (month)	0.124	0.378	-0.237	0.174	-0.181	0.097
Number of irrigation cycles	0.017	0.500	-0.132	-0.113	-0.010	0.108
Animal manure application	0.131	0.167	-0.005	-0.035	0.315	0.231
Pest infestation (%)	-0.201	0.223	0.237	0.067	0.407	-0.010
Disease infestation (%)	-0.235	0.205	0.300	0.062	0.176	-0.149
Number of weeds m^{-2}	-0.151	-0.024	0.419	0.222	0.297	0.037
Rainfall (mm)	-0.123	0.041	0.016	-0.584	0.288	0.053
Planting depth (cm)	-0.156	-0.076	-0.120	0.296	0.354	0.261
Soil salinity (dS m^{-1})	-0.283	0.263	-0.101	0.036	-0.194	0.142
Number of spikes m^{-2}	0.357	0.111	-0.032	-0.012	0.183	-0.138
Grain number spike^{-1}	0.322	0.139	-0.034	-0.060	0.154	-0.153
Thousand-grain weight (g)	0.296	0.214	-0.035	0.102	0.162	-0.319
Grain yield (t ha^{-1})	0.367	0.141	-0.015	-0.024	0.187	-0.042
Eigenvalue	6.21	2.86	2.30	1.83	1.29	1.07
Proportional variance (%)	28.22	13.02	10.48	8.33	5.85	4.87
Cumulative variance (%)	28.22	41.25	51.72	60.05	65.90	70.77

ber of weeds m^{-2}, planting depth, and soil salinity. Also, there was no correlation between rainfall and grain yield.

Stepwise linear regression

Regression analysis showed that the number of spikes m^{-2}, grain number spike^{-1}, spike length, broadleaf herbicide, soil salinity, time to plant maturity, narrow leaf herbicide, thousand-grain weight (TGW), rainfall, and awn length had justified the maximum of grain yield changes (Table 2).

Principal component analysis (PCA)

The PCA showed six major principal components (with eigenvalues more than one), indicating 70.77 % of the total variance among 90 wheat samples (Table 3). The first three principal components explained 28.22, 13.02 and 10.48 % of the total variance, respectively. These components represented 51.72 % of the total variance components (Table 3). The greatest variability in the data is justified by the first PC, in which number of spikes m^{-2}, grain number spike^{-1}, TGW, and grain yield had the most positive contribution while soil salinity had the most negative contribution. In PC2, number of irrigation cycles and time to plant maturity had the most positive contribution while awn length and plant had the most negative contribution. In PC3, K, P, number of weeds m^{-2}, and disease infestation had the most positive contribution. In PC4, seeding rate and rainfall had the most positive and negative contributions, respectively. In PC5, Animal manure application, pest infestation, and planting depth had the most positive contributions. Finally, spike length and narrow leaf herbicide had the most positive contribution to variation justified by PC6.

Next, PCA-Biplot was drawn by R packages factoextra, and ggplot2. It illustrates the relationships of the samples as well as the measured traits in wheat. The sample size is based on the amount of yield. The cosine of the angles

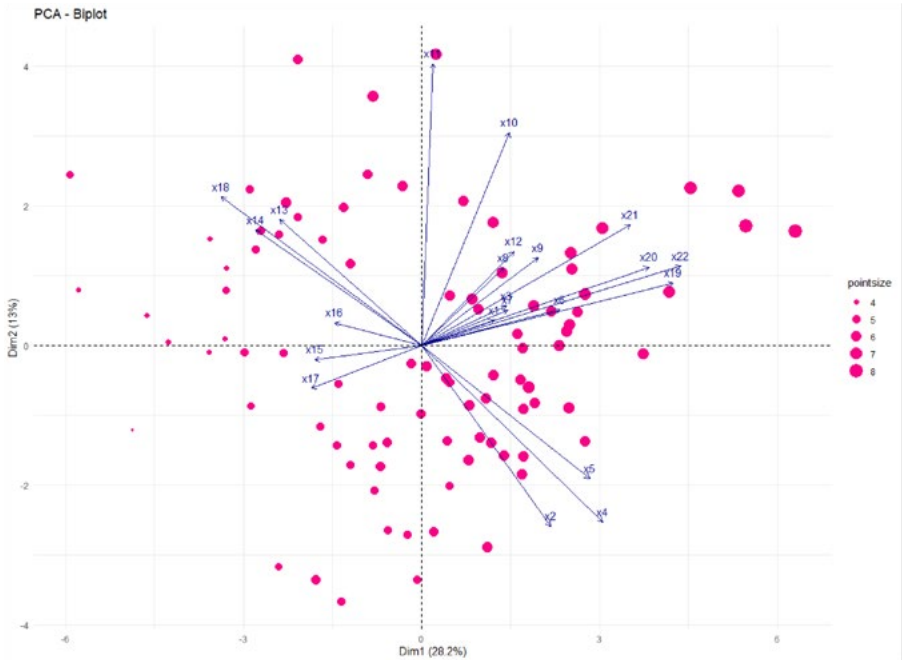

Figure 2. PCA-Bi-plot of the measured traits in wheat. Larger circles represent higher yield, and smaller circles show lower yield. Traits x1 to x22: see Figure 1.

between vectors shows the extent of correlation between traits. The angles between grain yield vector with planting depth (x17), rainfall (x16), and number of weeds (x8 and x9) m^{-2} were obtuse, indicating a negative correlation, while grain yield vector had acute angles with the rest of the traits, indicating a positive correlation (Fig. 2).

Finally, 9 traits were specified by integrated approaches including stepwise regression, correlation, and PCA. In summary, stepwise regression was applied to designate the most effective traits on grain yield in the regression model, which 10 traits with R^2 = 92.12% justified the maximum yield changes. Analysis of the correlation coefficient indicated that among these 10 traits, 9 traits were highly correlated with grain yield. In addition, PCA was conducted to identify significant traits related to grain yield and since loadings in the PC show that the variables how strongly influence the component and in other words reflect variables' significance, we enumerated the variables with large loadings in the PC, which 9 traits had large loadings in the PC1 and PC2. In other words, in all three methods (Stepwise linear regression, correlation, and PCA), these 9 traits including the awn length (cm), spike length (cm), narrow leaf herbicide, broadleaf herbicide, time to plant maturity (month), soil salinity (dS m^{-1}), number of spikes m^{-2}, grain number $spike^{-1}$, and TGW (g) were jointly identified as important and effective traits on grain yield.

Grain yield modeling by SVR for all traits

The results of the training (I) and testing (II) stages for all traits are shown in Fig. 3 as a scatter plot (A) and line-graph (B). According to the results shown in these figures, the SVR model applied to all traits, predicted the grain yield with R^2 of 0.7292 and 0.6239 for training and testing datasets, respectively. Moreover, based on the results of line-graph, this model could not estimate grain yield for all traits. As indicated in the line-graph, these models are not efficient in terms of predicting minimum or maximum grain yield.

Grain yield modeling by PLSR for all traits

The results of the training (I) and testing (II) stages for all traits are shown in Fig. 4 as a scatter plot (A) and line-graph (B). Taking all traits as inputs, the PLSR model predicted the grain yield with R^2 of 0.92 and 0.76 for training and testing datasets, respectively. Therefore, the PLSR model on all traits, as shown in the line-graphs, is strong in predicting the maximum and minimum grain yield. Accordingly, from two prediction models, PLSR was chosen for further evaluation of the selected traits associated with yield.

Grain yield modeling by SVR for the most important variables

The results of the training and testing stages for the 9 selected traits are shown in Table 4. The SVR model applying on all traits predicted the grain yield with R^2 of 0.73 and 0.62 for training and testing datasets, respectively. Moreover, applying the SVR model on 9 selected traits

Table 4. Grain yield estimation model results based on the SVR (support vector regression) and PLSR (partial least squares regression) models for training and testing datasets.

Models	Training dataset				Testing dataset			
	R^2	RMSE	MSE	BIAS	R^2	RMSE	MSE	BIAS
SVR (All data)	0.73	0.57	0.33	0.04	0.62	0.67	0.44	0.29
PLSR (All data)	0.92	0.26	0.07	0.00	0.76	0.46	0.21	0.27
SVR (9 traits)	0.88	0.35	0.12	-0.02	0.59	0.53	0.29	0.10
PLSR (9 traits)	0.92	0.26	0.07	0.00	0.85	0.32	0.10	-0.05

R^2: coefficient of determination; RMSE: root-mean-square error; MSE: mean squared error.

predicted the grain yield with R^2 of 0.88 and 0.59 for training and testing datasets, respectively. Accordingly, this model could not estimate satisfactorily grain yield neither for all traits nor 9 selected traits.

Grain yield modeling by PLSR for the most important variables

The results of the training (I) and testing (II) stages are illustrated in Fig. 5 as a scatter plot (A) and line-graph (B) for the 9 selected traits. The figures show a satisfactory agreement between the measured and predicted PLSR model values. Applying the PLSR model on 9 selected traits predicted the grain yield with R^2 of 0.92 and 0.85 for training and testing datasets, respectively.

Table 4 presents the R^2, RMSE, MSE, and BIAS values of the applied models for both training and testing stages for all traits and 9 selected traits by integrated approaches including stepwise regression, correlation, and PCA. As shown in Table 4, the SVR model demonstrated the least accurate results for estimating grain yield with the highest RMSE, MSE, BIAS, and the lowest value of R^2 for both training and testing stages for all data and 9 selected traits. Accordingly, this model could not estimate grain yield satisfactorily neither for all traits nor 9 selected traits. Table

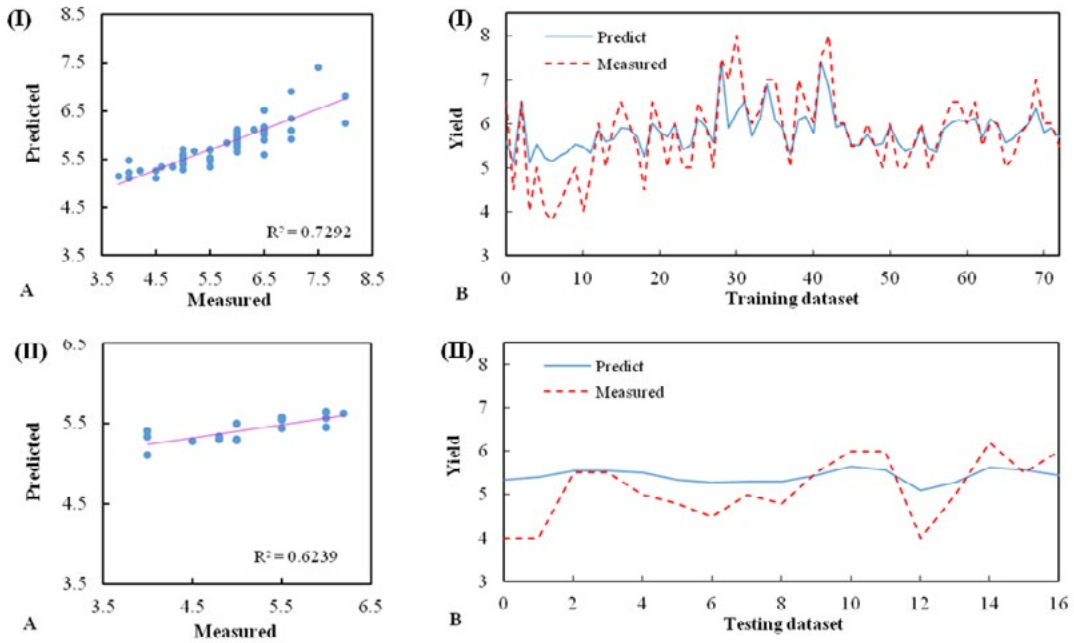

Figure 3. Scatter plot (A) and line-graph (B) of predicted vs. measured grain yield for the training (I) and testing (II) datasets of all traits, using SVR. Traits include seeding rate (kg ha^{-1}), awn length (cm), spike length (cm), plant height (cm), nitrogen fertilizer (N, kg ha^{-1}), phosphorus fertilizer (P, kg ha^{-1}), potassium chloride fertilizer (K, kg ha^{-1}), narrow-leaf herbicide, broadleaf herbicide, time to plant maturity (month), number of irrigation cycles, animal manure application, pest infestation (%), disease infestation (%), number of weeds m^{-2}, rainfall (mm), planting depth (cm), soil salinity (dS m^{-1}), number of spikes m^{-2}, grain number $spike^{-1}$, thousand-grain weight (g), and grain yield (t ha^{-1}).

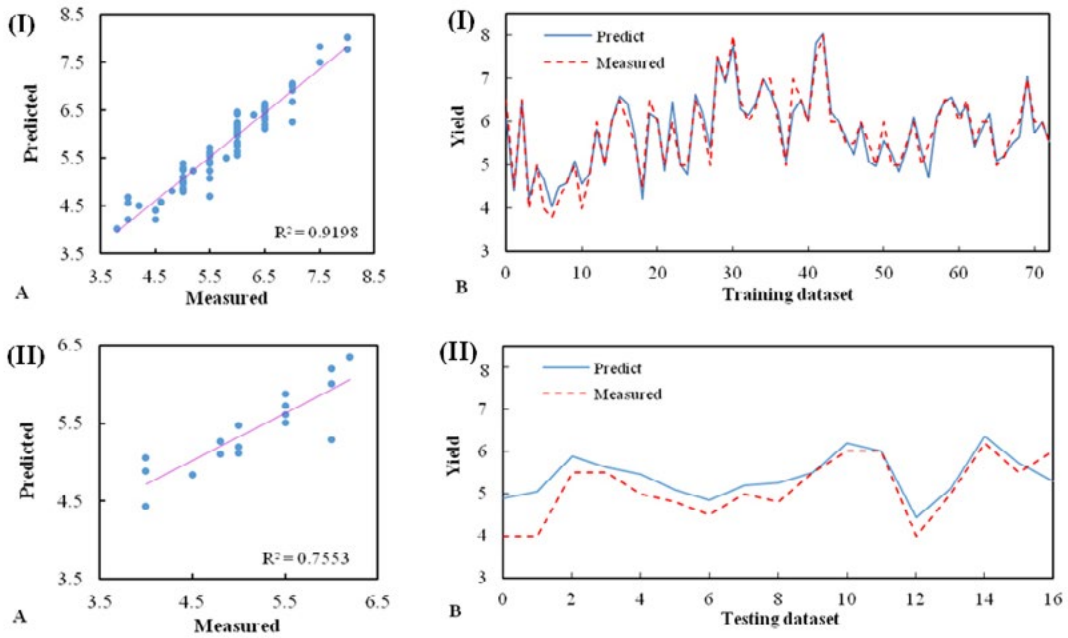

Figure 4. Scatter plot (A) and line-graph (B) of predicted vs. measured grain yield for the training (I) and testing (II) datasets of all traits, using PLSR. Traits as Figure 3.

4 also shows that the PLSR model outperforms SVR models given that for all the data and nine selected traits the value of RMSE, MSE, and BIAS were the lowest and the value of R^2 was the highest. Therefore, given the results, the PLSR model was implemented for nine selected traits as the best model to achieve the most reliable estimate of grain yield.

Discussion

Regarding the third main staple food crop and one of the most economically significant crops worldwide, determining the factors affecting wheat yield is of paramount importance to subsequent improve yield and ensure food security (Farokhzadeh et al*.,* 2021).

The factors in yield are usually correlated, meaning redundancies or potentially misleading outcomes when it comes to identifying the dominant variables that control the yield (Carrascal et al., 2009), and may misdirect the plant breeders to achieving their main breeding objectives. In this study also, Pearson correlation analysis revealed the relationship among 22 traits (Fig. 1). In this regard, to get a better understanding of key determinants of wheat yield, a comprehensive data analysis including multivariate statistical analysis (stepwise regression, correlation analysis, and PCA) combined with machine learning methods (PLSR and SVR) was done. Several studies have documented the relationship between yield and some yield components in wheat (Leilah & Al-Khateeb, 2005; Baye et al., 2020). In this context, to clarify the underlying determinants of wheat yield diversity, the yield components were also included in the input variables list.

Correlation and stepwise linear regression

Analysis of the correlation coefficient indicated that among traits, number of spikes m^{-2} was highly correlated with grain yield followed by grain number $spike^{-1}$ and TGW in wheat. These findings were in accordance with previous studies carried out to determine traits affecting grain yield (Farokhzadeh et al*.,* 2013; 2020).

Stepwise regression analysis was used to measure the effect of agronomic traits on wheat grain yield. Through this method, 92.12% of the total change in grain yield was attributed to 10 traits. All these traits except rainfall had a significant positive correlation with grain yield. Although rainfall trait had no significant correlation with grain yield, it had a direct effect on grain yield regarding the standardized coefficient of rainfall in the stepwise regression model. The lack of a significant correlation between rainfall and grain yield can be due to the fact that irrigation probably neutralizes the effect of rainfall effect. Farokhzadeh et al. (2022) using stepwise regression reported that wheat grain yield as a dependent variable has been modeled as a function of the independent variables grain number $spike^{-1}$, days to heading, and spikelet number $spike^{-1}$.

Principal component analysis (PCA)

PCA was performed and the first two components were utilized to make a bi-plot to visualize the relationships of the samples as well as the measured traits (Fig. 1). As seen, low or average yield samples are on the left side and high and very high yield samples are on the right side of the bi-plot.

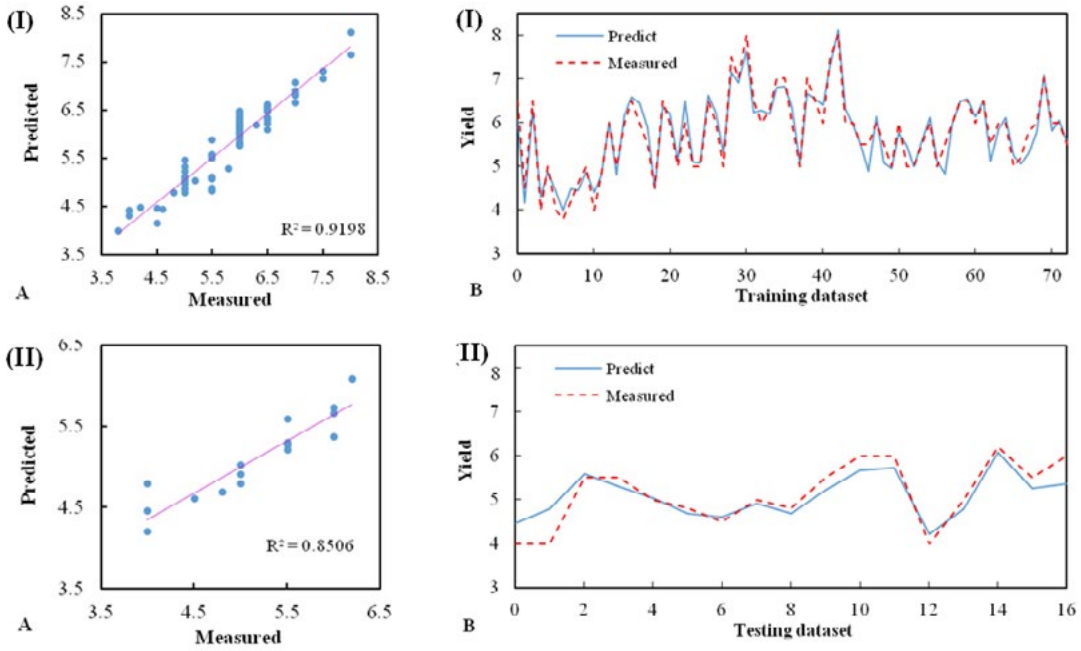

Figure 5. Scatter plot (A) and line-graph (B) of predicted vs. measured grain yield for the training (I) and testing (II) datasets of nine traits, using PLSR. Traits include awn length (cm), spike length (cm), narrow-leaf herbicide, broadleaf herbicide, time to plant maturity (month), soil salinity (dS m^{-1}), number of spikes m^{-2}, grain number $spike^{-1}$, and thousand-grain weight (g).

Samples with larger PCA1 and PCA2 scores showed a very high yield. As clearly revealed in the PCA bi-plot, number of spikes m^{-2}, grain number $spike^{-1}$, and TGW were the dominant and similar traits in samples with very high yield, while pest infestation (x13), disease infestation (x14), and soil salinity (x18) are the dominant traits of samples with a low or average yield. The results of PCA confirm the correlation coefficient and stepwise regression results expressing that the increase in the number of spikes m^{-2}, grain number $spike^{-1}$, and TGW results in higher yields. Different reports confirm the importance of the traits that are identified in this study. For example, the importance of awn length, spike length, time to plant maturity (month), weed percentage, soil salinity, number of spikes m^{-2}, grain number $spike^{-1}$, and TGW have been highlighted in some studies (Shamsi et al., 2011; Farokhzadeh et al., 2013; 2020; 2022). The main goal of wheat breeding programs is to improve grain yield potential. Grain yield and yield components, including the number of spikes, grain number $spike^{-1}$, and TGW are polygenic complex traits that are influenced by genetic background, environment, and the interaction between them. Usually, grain yield can be determined by combining two components: grain number and TGW (Zhang et al., 2016a). Physiological studies showed that increased grain yield is positively associated with an increase in grain number, attended by a negative association between grain number and grain weight (Miralles & Slafer, 2007). Grain weight is a yield component for breeders that have made considerable attempts to improve (Lopes et al., 2012). Larger seeds are not only directly related to grain yield, but also affect beneficially seedling vigor and early growth, thus promoting and stabilizing production ability (Botwright et al., 2002). Under current agronomic production systems, to enhance grain yield potential, improvement of spikes number or grains m^{-2} is more important than other yield components (Gaju et al., 2009).

It is important to highlight the awn length affects wheat yield. Abebe et al. (2009) revealed that the awn length influences the grain yield. The study showed that awn affects significantly spikes photosynthetic characteristics and tolerance to stress in barley. Seed weight and size were reduced due to the absence of awns and in turn reduction in starch content caused by it.

As expected, soil salinity was another key trait identified by the multivariate statistical analysis in this study. The soil salinity issues generally happen in arid/semi-arid areas and decrease the productivity of crops at different levels. Salinity is also a major factor limiting crop yield in defectively drained soils (Patel et al., 2002; Rogers, 2002). Rajpar et al. (2006) reported spike length, plant height, number of spikelet $spike^{-1}$ and grain yield progressively decreased with increasing soil salinity. Besides, with the increase in soil ECe, Na^+ and K^+ concentrations increase and decrease, respectively, which leads to reduce the K^+/Na^+ ratio in the flag leaf sap and grains (Rajpar et al., 2006). The percentage of weeds is one of the most important harmful and reducing factors in agricultural systems. It is the main consumer of nutrients such as chemical fertilizers that can increase their growth and development compared to crops (Jalilian et al., 2018). The percentage of weeds reduces grain yield and yield components (Mekonnen, 2022).

Yield estimation model results based on the SVR and PLSR models

In the current study, two analyzing methods, SVR and PLSR, were performed on agronomic traits obtained from 90 farms in six different areas for model development. PLSR performed better than SVR, in which R^2 and RSME were 0.76 vs. 0.62 and 0.46 vs. 0.67, respectively, and showed a good predictive ability and robustness compared with the SVR. This result is consistent with the conclusion of a previous study which stated that PLSR can prioritize the most important factors controlling winter wheat yield through a long-term experiment (Hu et al., 2018). Duan et al. (2020) argued that PLSR gives us a practical way to determine the factors in yield as it practically removes the correlation of the variable and modifies the bias of the factors' role in rice yield. These findings are consistent with the present work, which indicates that PLSR has a better performance in terms of predicting yield. Also, Zhang et al. (2020) estimated the yield of field-grown winter wheat by applying the PLSR model. These reports are in accordance with the present study which implies that PLSR displayed better prediction capability.

Taking all traits in predicting grain yield with the superior model (PLSR), the predicted R^2 and RSME of the validation (testing) dataset for grain yield prediction were 0.76 and 0.46, respectively. Whereas applying this model on a testing dataset with 9 selected traits as inputs improved R^2 and RSME to 0.85 and 0.32, respectively. PLSR model could attain a comparable accuracy by using the 9 selected traits compared with using all traits. This result indicated the substantial influence of nine selected traits in explaining the grain yield variation. In other words, multivariate statistical analyses could provide great potential to reduce the number of inputs (traits) of the model and consequently increase the prediction accuracy of PLSR.

The high accuracy of the PLSR model with 9 inputs (traits) implies that the selected traits are attributable to the improvement of grain yield. From a plant breeding perspective, our study recommends considering the number of spikes m^{-2}, grain number $spike^{-1}$, TGW , awn length, spike length, and time to plant maturity (month) in breeding programs for achieving improved grain yield in wheat. Farokhzadeh et al. (2022) using integrating results of multivariate statistics and supervised learning methods illustrated that spikelet number $spike^{-1}$ and grain number $spike^{-1}$ traits can be used to create a selection index for the high grain yield in wheat.

Conclusions

The grain yield of wheat is the most important economic part of the plant, which is the result of yield components and other related traits. Identifying the main components of grain yield and their relationship with grain yield can be essential in directing breeding and management programs to increase yield. In the current study, multivariate analyses including, correlation, stepwise regression, and principal component analysis (PCA) were applied to select useful subsets of agronomic traits from 90 farms and to evaluate the order of importance of traits influencing the grain yield. Moreover, partial least squares regression (PLSR) displayed better prediction capability with nine inputs (nine selected traits). In brief, integrated multivariate statistical analyses and machine learning regression methods could be a powerful tool in determining key contributing traits to wheat yield.

Acknowledgments

The authors are very grateful to Zohreh Sheikh Khozani for her help in the data modeling and analysis and Ehsan Bijanzadeh for valuable guidance throughout this study.

Authors' contributions

Conceptualization: Z. Zinati, S. Farokhzadeh.
Data curation: S. Farokhzadeh.
Formal analysis: S. Farokhzadeh, Z. Zinati.
Funding acquisition: A. Behpoori.
Investigation: A. Behpoori**,** Z. Khosravi.
Methodology: S. Farokhzadeh.
Project administration: Z. Zinati.
Resources: S. Farokhzadeh, Z. Zinati.
Software: S. Farokhzadeh, Z. Zinati.
Supervision: A. Behpoori.
Validation: Z. Zinati, S. Farokhzadeh, A. Behpoori.
Visualization: S. Farokhzadeh, Z. Zinati.
Writing – original draft: S. Farokhzadeh, Z. Zinati.
Writing – review & editing: Z. Zinati, S. Farokhzadeh, A. Behpoori.

References

Abdel-Sattar M, Aboukarima AM, Alnahdi BM, 2021. Application of artificial neural network and support vector regression in predicting mass of ber fruits (*Ziziphus mauritiana* Lamk.) based on fruit axial dimensions. PLoS ONE 16(1): e0245228. https://doi.org/10.1371/journal.pone.0245228

Abdi H, 2010. Partial least squares regression and projection on latent structure regression (PLS regression). Wiley Interdisciplinary Reviews: Comput Stat 2: 97-106. https://doi.org/10.1002/wics.51

Abebe T, Wise RP, Skadsen RW, 2009. Comparative transcriptional profiling established the awn as the major photosynthetic organ of the barley spike while the lemma and the palea primarily protect the seed. Plant Genome 2: 247-259. https://doi.org/10.3835/plantgenome.2009.07.0019

Asseng S, Turner NC, Keating BA, 2001. Analysis of water- and nitrogen-use efficiency of wheat in a Mediterranean climate. Plant Soil 233: 127-143. https://doi.org/10.1023/A:1010381602223

Bao SD (ed), 2005. Analysis of soil agrochemistry. China Agriculture Press, Beijing, China. 495 pp.

Barikloo A, Alamdari P, Moravej K, Servati M, 2017. Prediction of irrigated wheat yield by using hybrid algorithm methods of artificial neural networks and genetic algorithm. J Water Soil 31: 715-726.

Baye A, Berihun B, Bantayehu M, Derebe B, 2020. Genotypic and phenotypic correlation and path coefficient analysis for yield and yield-related traits in advanced bread wheat (*Triticum aestivum* L.) lines. Cogent Food Agric 6(1): 1752603. https://doi.org/10.1080/23311932.2020.1752603

Botwright TL, Condon AG, Rebetzke AG, Richards RA, 2002. Field evaluation of early vigour for genetic improvement of grain wheat. Aust J Agric Res 53(10): 1137-1145. https://doi.org/10.1071/AR02007

Carrascal LM, Galván I, Gordo O, 2009. Partial least squares regression as an alternative to current regression methods used in ecology. Oikos 118(5): 681-690. https://doi.org/10.1111/j.1600-0706.2008.16881.x

Costa C, Menesatti P, Spinelli R, 2012. Performance modelling in forest operations through partial least square regression. Silva Fenn 46(2): 241-252. https://doi.org/10.14214/sf.57

Duan L, Xie H, Li Z, Yuan H, Guo Y, Xiao X, Zhou Q, 2020. Use of partial least squares regression to identify factors controlling rice yield in Southern China. Agron J 112(3): 1502-1516. https://doi.org/10.1002/agj2.20161

Farokhzadeh S, Shahsavand-Hassani H, Mohammadi-Nejad GH, 2013. Evaluation of genetic diversity of primary tritipyrum, triticale and bread wheat genotypes. Iran J Agron Sci 5: 93-112.

Farokhzadeh S, Fakheri BA, Mahdinejad N, Tahmasebi S, Mirsoleimani A, Heidari B, 2020. Mapping QTLs associated with grain yield and yield-related traits under aluminum stress in bread wheat. Crop Pasture Sci 71: 429-444. https://doi.org/10.1071/CP19511

Farokhzadeh S, Fakheri BA, Zinati Z, Tahmasebi S, 2021. New selection strategies for determining the traits contributing to increased grain yield in wheat (*Triticum aestivum* L.) under aluminum stress. Genet Resour Crop Evol 68: 2061-2073. https://doi.org/10.1007/s10722-021-01117-4

Farokhzadeh S, Shahsavand-Hassani H, Zinati Z, Rajaei M, 2022. Evaluation of triticale lines compared to wheat cultivars in terms of agronomic traits using supervised learning methods and multivariate statistics. Philipp Agric Sci 105(4): 369-389.

Fassio A, Cozzolino D, 2003. Non-destructive prediction of chemical composition in sunflower seeds by near infrared spectroscopy. Indust Crops Prod 20: 321-329. https://doi.org/10.1016/j.indcrop.2003.11.004

Gaju O, Reynolds MP, Sparkes DL, Foulkes MJ, 2009. Relationships between large-spike phenotype, grain number, and yield potential in spring wheat. Crop Sci 49: 961-973. https://doi.org/10.2135/cropsci2008.05.0285

Gustavo AS, Guillermo AG, Roman AS, Daniel JM, 2022. Physiological drivers of responses of grains per m2 to environmental and genetic factors in wheat. Field Crops Res 285: 108593. https://doi.org/10.1016/j.fcr.2022.108593

Hu Y, Wei X, Hao M, Fu W, Zhao J, Wang Z, 2018. Partial least squares regression for determining factors controlling winter wheat yield. Agron J 110: 281-292. https://doi.org/10.2134/agronj2017.02.0108

Jalilian A, Mondani F, Khoramivafa M, Bagheri A, 2018. Evaluation of Clipest model in simulation of winter wheat (*Triticum aestivum* L.) and wild oat (*Avena ludoviciana* L.) competition in Kermanshah. Iran J Agroeco 10: 248-266.

Leilah AA, Al-Khateeb SA, 2005. Statistical analysis of wheat yield under drought conditions. J Arid Environ 61(3): 483-496. https://doi.org/10.1016/j.jaridenv.2004.10.011

Lopes MS, Reynolds MP, Manes Y, Singh RP, Crossa J, Braun HJ, 2012. Genetic yield gains and changes in associated traits of CIMMYT spring bread wheat in a "historic" set representing 30 years of breeding. Crop Sci 52(3): 1123-1131. https://doi.org/10.2135/cropsci2011.09.0467

Martens H, Martens M, 2000. Modified Jack-knife estimation of parameter uncertainty in bilinear modeling by partial least squares regression (PLSR). Food Qual Prefer 11: 5-16. https://doi.org/10.1016/S0950-3293(99)00039-7

Mekonnen G, 2022. Wheat (*Triticum aestivum* L.) yield and yield components as influenced by herbicide application in Kaffa Zone, Southwestern Ethiopia. Int J Agron 2022: 3202931. https://doi.org/10.1155/2022/3202931

Meng M, Zhao C, 2015. Application of support vector machines to a small-sample prediction. Adv Petrol Explord Dev 10(2): 72-75.

Miralles DJ, Slafer GA, 2007. Sink limitations to yield in wheat: How could it be reduced? J Agric Sci 145: 139-149. https://doi.org/10.1017/S0021859607006752

Nguyen HT, Lee BW, 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. Eur J Agron 24: 349-356. https://doi.org/10.1016/j.eja.2006.01.001

Norouzi M, Ayoubi S, Jalalian A, Khademi H, Dehghani AA, 2010. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. Acta Agric Scand - B Soil Plant Sci 60(4): 341-352. https://doi.org/10.1080/09064710903005682

Patel R, Prasher S, Bonnell R, Boughton R, 2002. Development of comprehensive soil salinity index. J Irrig Drain Eng 128: 185-188. https://doi.org/10.1061/(ASCE)0733-9437(2002)128:3(185)

Rajpar I, Khanif YM, Soomro FM, Suthar JK, 2006. Effect of NaCl salinity on the growth and yield of Inqlab wheat (*Triticum aestivum* L.) variety. Am J Plant Physiol 1: 34-40. https://doi.org/10.3923/ajpp.2006.34.40

Rogers ME, 2002. Irrigating perennial pasture with saline water: Effects on soil chemistry, pasture production and

composition. Aust J Exp Agric 42: 265-272. https://doi.org/10.1071/EA00128

Shaibu AS, Adnan AA, 2015. Predicting grain yield of maize using drought tolerance traits. Afr J Agric Res 10(33): 3332-3337. https://doi.org/10.5897/AJAR2015.9561

Shamsi K, Petrosyan M, Noor-Mohammadi G, Haghparas A, Kobraee S, et al., 2011. Differential agronomic responses of bread wheat cultivars to drought stress in the west of Iran. Afr J Biotechnol 10: 2708-2715. https://doi.org/10.5897/AJB10.1133

Sheikh Khozani Z, Khosravi KH, Torabi M, Mosavi A, Rezaei B, Rabczuk T, 2020. Shear stress distribution prediction in symmetric compound channels using data mining and machine learning models. Front Struct Civ Eng 14: 10971109. https://doi.org/10.1007/s11709-020-0634-3

Takahashi S, Anwar MR, 2007. Wheat grain yield, phosphorus uptake and soil phosphorus fraction after 23 years of annual fertilizer application to an Andosol. Field Crops Res 101: 160-171. https://doi.org/10.1016/j.fcr.2006.11.003

Tian Y, Xu YP, Wang G, 2018. Agricultural drought prediction using climate indices based on support vector regression in Xiangjiang River basin. Sci Total Environ 622: 710-720. https://doi.org/10.1016/j.scitotenv.2017.12.025

Vapnik V (ed), 1995. The nature of statistical learning theory. Springer, NY. https://doi.org/10.1007/978-1-4757-2440-0

Wilson A, Hemalatha N, Sukumar R, 2021. Computational prediction model for pepper yield prediction using support vector regression. AgriRxiv 10310468. https://doi.org/10.31220/agriRxiv.2021.00069

Wold S, Sjostrom M, Eriksson L, 2001. PLS-regression: A basic tool of chemometrics. Chemometr Intell Lab Syst 58: 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yang Y, Li N, Wu Y, Liu B, Li S, Tao L, et al., 2022. Key phenotypes related to wheat grain yield in a two-site multi-cultivar test. Agron J 114(5): 2874-2885. https://doi.org/10.1002/agj2.21098

Zhang H, Chen J, Li R, Deng Z, Zhang K, Liu B, Tian J, 2016a. Conditional QTL mapping of three yield components in common wheat (*Triticum aestivum* L.). Crop J 4: 220-228. https://doi.org/10.1016/j.cj.2016.01.007

Zhang PP, Zhou XX, Wang ZX, Mao W, Li WX, Yun F, et al., 2020. Using HJ-CCD image and PLS algorithm to estimate the yield of field-grown winter wheat. Sci Rep 10: 5173. https://doi.org/10.1038/s41598-020-62125-5

Zhang Y, Xu W, Wang W, Dong H, Qi X, Zhao M, et al., 2016b. Progress in genetic improvement of grain yield and related physiological traits of Chinese wheat in Henan Province. Field Crops Res 199: 117-128. https://doi.org/10.1016/j.fcr.2016.09.022