Association between the Yield and the Main Agronomic Traits of Tartary

Buckwheat Evaluated Using the Random Forest Model

Shan Feng, Jing Li, Guoqi Qian and Baili Feng*

S. Feng, College of Statistics, Xi'an University of Finance and Economics, Xi'an, Shaanxi, 710100, China; J. Li and B. Feng, College of Agronomy, Northwest A & F University, State Key Laboratory of Crop Stress Biology for Arid Areas, Yangling, Shaanxi, 712100, China; G. Qian, School of Mathematics and Statistics, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia. S. Feng and J. Li contributed equally to this work. Received

_____. *Corresponding author (fengbaili@nwsuaf.edu.cn).

ABSTRACT

The popularity of Tartary buckwheat (*Fagopyrum tataricum* (L.) Gaertn), as medicinal and food crop, has been increasing in recent years. However, its low yield seriously restricts its industrial development. Amongst the various studies conducted to enhance the productivity of Tartary buckwheat, the association between yield and main agronomic traits has formed the foundation for the breeding and cultivation of high-yielding varieties, becoming the primary interest of breeders. The commonly used methods are often restricted by sample size,

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/csc2.1243](#).

This article is protected by copyright. All rights reserved.

distribution assumptions and trait properties and confined to the linear relationship. In this paper, the random forest regression model was used to obtain a comprehensive and reliable evaluation. The phenotypic data of 200 Tartary buckwheat landraces with 15 quantitative and 2 qualitative agronomic traits for two consecutive years were used. Results were compared between planting seasons and with those from classical methods, such as the correlation analyses and the multiple linear regression model. The random forest model distinguished the number of grains per plant, plant height and 1000-grain weight as the most influential agronomic traits in both seasons. The main and interactive effects were explored using the accumulated local effects plot and showed great conformity between the two seasons. The robustness and reliability of the random forest model make it a desirable methodology for breeding new varieties and germplasm innovation of Tartary buckwheat and other crops.

INTRODUCTION

Tartary buckwheat (*Fagopyrum tataricum* (L.) Gaertn), which belongs to the family Polygonaceae, is traditionally regarded as a medicinal and food crop (Li et al., 2019, 2020) and has been gaining increasing popularity in recent years. However, its low yield restricts its industrial development. Various studies have been conducted to enhance the productivity of Tartary buckwheat. The association between yield and main agronomic traits has formed the foundation for the breeding and cultivation of high-yielding varieties, thus become the primary interest of breeders.

Most studies have adopted the correlation coefficients to assess the importance of the main agronomic traits to the yield of crops. However, Yang et al. (2008) have shown that the simple correlation coefficient may give misleading results as the effect of a trait on the yield

may be confounded by the indirect effect from other traits. Thus, the partial correlation coefficient is preferred. Multiple linear regression has been conducted by Yang et al. (2008) for a comprehensive assessment of the association between yield and agronomic traits of Tartary buckwheat. However, these methods focus on the linear or monotonic relationship between yield and agronomic traits and need assumptions. For example, data are required to be approximately normal for standard statistical tests, and the sample size should be bigger than the number of traits for model solvability and accuracy of the results. In addition, the correlation analyses are only suitable to quantitative traits. The explanation of the linear regression models with the common standardisation procedures remains to be examined when qualitative traits exist.

The random forest model (Breiman, 2001) is widely used in genetics (Diaz-Uriarte and de Andres, 2005), metabolomics (Li et al., 2012), entomology (Li, 2013) and ecology (Zhang et al., 2014). In agricultural science, Qin et al. (2016) have adopted the random forest model to analyse the effect of climate factors on the yield of oilseed rape. Everingham et al. (2016) have constructed the sugarcane yield prediction model with observed climate indices, and Smidt et al. (2016) have identified the important field attributes that explain the soybean yield by using the random forest regression model. As an ensemble learning algorithm, the random forest model requires less assumptions compared with the classical classification and regression models. The ‘bootstrap resampling’ and the ‘feature bagging’ techniques have greatly improved the accuracy and robustness of the prediction results (Hastie et al., 2009, Li et al., 2012). In most cases, the random forest model has outperformed the traditional methods (Everingham et al., 2016).

In this paper, we have comprehensively evaluated the association between the yield and the main agronomic traits of Tartary buckwheat by using the random forest regression model to determine whether the random forest model can offer new insights into the trait–yield association of Tartary buckwheat. Through the mean decrease in node impurity (MDI) (Breiman, 2001), we have identified the most important agronomic traits that contribute to the yield of Tartary buckwheat. The accumulated local effects (ALE) plot (Apley, 2016) is used as a supplementary tool to visualise and explore the main and the second-order interaction effects of the important traits. The model is constructed on the basis of the Tartary buckwheat data for two consecutive years. Comparisons are made with the commonly used methods, such as the correlation analysis and the multiple linear regression model and between two planting seasons. The application of classical methods should be based on the related assumptions and conditions. Some shining advantages of the random forest model make it a desirable methodology for breeding new varieties and germplasm innovation of Tartary buckwheat and other crops.

MATERIALS AND METHODS

Plant materials

A total of 200 Tartary buckwheat landraces were selected as experimental materials. All Tartary buckwheat seeds were provided by the Minor Grain Research Centre of the College of Agronomy, Northwest A & F University. The experiment was conducted in the Experimental Station of Qishan County, Baoji Academy of Agricultural Sciences, Shaanxi, China (107° 33' E, 34° 07' N; altitude, 850 m) in 2018 and 2019. The area is characterised by warm temperate semihumid climate with annual average precipitation of approximately 626.1

mm, average annual temperature of 11.9 °C and annual sunshine duration of 2066.6 h. In the experimental site, 0–40 cm soil was loess soil with 18.18 g kg⁻¹ organic carbon, 4.02 mg kg⁻¹ available nitrogen, 11.79 mg kg⁻¹ available phosphorus and 395.7 mg kg⁻¹ available potassium. The pH value of the soil was 8.37.

Collection of agronomic traits

The experiment involved a completely randomised block design with two replications. The area of each block was 4 m² containing four rows with row distance of 40 cm. The distance of the two germplasm materials was 80 cm. On average, the plant density was approximately 300 000 plants per hectare. The Tartary buckwheat landraces were sown on July 5, 2018 and June 25, 2019 and harvested in batches in accordance with the maturity time of different landraces.

At the maturity stage, five plants from each landrace in each replication were randomly selected for phenotyping. Seventeen agronomic traits of the Tartary buckwheat (Table 1), which can be classified as yield-related traits, plant phenotypes and grain-related traits, were investigated. The yield-related traits consisted of yield per hectare, number of grains per plant (NGP) and 1000-grain weight (TGW). The plant phenotypes included plant height (PH), stem diameter (SD), number of nodes (NN), number of branches (NB) and branch height (BH). The grain-related traits were grain area (GA), grain perimeter (GP), grain aspect ratio (GAR), grain length (GL), grain width (GW), grain diameter (GD), grain roundness (GR), grain colour (GC) and grain shape (GS). For the experiment in 2018, the data of three Tartary buckwheat landraces were subjected to severe missingness and omitted, leaving the complete data of 197 landraces.

Random forest regression model

In this paper, the random forest regression model was conducted on the original Tartary buckwheat data to explore the importance and the effects of the main agronomic traits to the yield of Tartary buckwheat. The mean decrease in node impurity (MDI) or mean decrease in accuracy (MDA) obtained through the model fitting process are natural indices for the ranking of the explanatory variables (Breiman, 2001). As the MDA tends to spread the importance of predictors uniformly (Hastie et al., 2009), to make the ranking results more distinguishable, MDI was used. The computation results of MDI were standardised using the Min–Max method to transform into a range of 0 to 1 and maintain the rankings and relative distances between adjacent traits.

The random forest regression model is a nonparametric regression technique (Li et al., 2012). Thus, some supplementary tools, like the partial dependence (PD) plot (Friedman, 2001), were often used to describe the pattern of predictors that influence response. Compared with the PD plot, the ALE plot can overcome the ‘problem of extrapolation’ and can be computed efficiently. Thus, in this paper, instead of the PD plot, the ALE plot was used to explore the main and the interaction effects of the substantial agronomic traits.

In the ALE main effect plot, the bootstrap main effect curve produced from 300 replicates was added. The discrepancy between the original and the bootstrap main effect curves formed the nonparametric bootstrap estimation of the bias of the ALE main effect estimator, which should be zero theoretically. The bootstrap confidence interval based on the normal approximation (Davison and Hinkley, 2013) was also plotted to depict the variability of the

ALE main effect estimator. In the ALE interaction plot, the effects were scaled to the range of 0 to 1 using the Min–Max standardisation method for legend consistency.

All statistical analyses were performed in the R software (version 3.6.3). The ‘randomForest’ package (Liaw and Wiener, 2002) was loaded to build the random forest regression model. By default, the model was built using 500 trees derived from 500 bootstrapped datasets, and the size of the random subset of variables for each split was 5, which was approximately one-third of all available predictor variables. The ‘ALEPlot’ package (Apley, 2016) was loaded to realise the ALE main and interaction plots of the significant traits. The default number of intervals the range of each predictor is split in the algorithm is 40. Considering the sample size of our data, the number of intervals was specified as 10 for the ALE main effect and 6 for the interactive effect plots to overcome overfitting problems, which may increase the variability of the results. The yield function of the main effect was calculated on each splitting point of the intervals and the interaction effects on 49 crossings over the two-dimensional space. The length of intervals was adjusted internally in accordance with the number of sample points in each interval or square.

Comparison with classical methods

Comparisons were made with the commonly used correlation analysis and the multiple linear regression model. In the correlation analysis, the Pearson’s linear correlation (Sedgwick, 2012), often known as the simple correlation, and partial correlation (Baba et al., 2004) were considered. The latter excluded the confounding effects from the other traits and was expected to give a more reliable assessment than the former. The significance of the correlation coefficient was evaluated using the Student’s *t*-test statistics assuming a two-sided

alternative with the distribution assumption that the data had multivariate normal distribution. As the correlation coefficients were not comparable amongst different correlation measures, the evaluation was based on the p -values.

Besides the Pearson's linear correlation, the Spearman's rank correlation was considered. In the Spearman's correlation, data were converted to ranks before calculation (Fieller et al., 1957). The Spearman's rank correlation was used to measure the monotonic relationship between a pair of variables. Thus, the significance of the Pearson's correlation implies the significance of the Spearman's correlation (Hauke and Kossowski, 2011).

For the multiple linear regression model, the stepwise and the best subset selection techniques are typically conducted to solve the multicollinearity problems (Lu, 2007, You and Yan, 2017, Zhu et al., 2008) and for model simplification and variable selection. As the dimension of the trait space was moderate compared with the total number of landraces in our study, the best subset selection was implemented to obtain the 'best' linear regression model between the yield and the main agronomic traits. The model fitting was realised using the 'leaps' package in R (Miller, 2002).

During the best subset selection, the Akaike (AIC) and the Bayesian (BIC) information criterion statistics were computed for each linear regression model with different numbers of predictors. The final model was located by finding the smallest value of AIC or BIC over these models. The values of the BIC and the AIC statistics were mapped to the interval of [0,1] through the Min–Max method to facilitate comparison between criteria. The BIC and the AIC were suitable for the comparison of models with different numbers of traits, and the BIC statistic placed a heavier penalty on models with more variables, generally resulting in

the selection of smaller models (James et al., 2013). Considering the dimensional discrepancy of predictors, data standardisation was implemented on each quantitative trait. The significance of the coefficient was evaluated using the Student's *t*-test statistic with normal distribution assumption. The grain-related traits GC and GS were qualitative. GC of a landrace was categorised as grey, brown and black, and GS included heart-shaped, long-tapered and short-tapered. These traits were replaced with the dummy variables during the model fitting.

All statistical analyses were conducted on the Tartary buckwheat data in a comparable manner between the 2018 and 2019 planting seasons.

RESULTS

Random forest regression analysis

The random forest regression model was quite successful in explaining the association between the yield and the main agronomic traits of Tartary buckwheat. On average, 85.72% and 82.58% variances of the out-of-bag data were explained by the fitted model for 2018 and 2019 planting seasons, respectively.

The rankings of the main agronomic traits in 2018 and 2019 based on the MDI are shown in Figure 1. Sixteen traits were arranged in descending order of importance in each plot. The rankings of the main agronomic traits in 2018 and 2019 were quite similar with each other. The yield-related traits and the plant phenotypes ranked highly, and the grain-related traits showed generally less importance. The top five traits, including the yield-related traits NGP and TGW and plant phenotypes PH, SD and NN, were the same in the two seasons. The first

three traits, namely, NGP, PH and TGW, in the two seasons were distinguished from the remaining traits. PH showed relatively larger effects than TGW in the yield of Tartary buckwheat. NGP had the largest influence amongst all agronomic traits. The MDIs of the remaining traits were relatively small and exhibited little differences.

The frequency distributions of the three remarkable traits, namely, NGP, PH and TGW, in 2018 and 2019 are demonstrated in Figure 2. The kernel and the normal density curves with mean and standard deviation as the sample mean and standard deviation of the corresponding trait were plotted for each trait and enlarged to match the scheme of the frequency distribution of the corresponding traits. The kernel density curves of TGW were symmetrically distributed and close to the normal density curves in both seasons, where the sample points were concentrated in the middle of the distribution. The distribution of NGP was skewed to the right in both seasons, as evidenced by the relatively sparse data points on the right. The distribution of PH was skewed to the left in 2018, as evidenced by the sparse data points on the left.

The ALE main effect plots exploring how the significant traits, namely, NGP, PH and TGW, influenced the yield of the Tartary buckwheat in 2018 and 2019 planting seasons are shown in Figure 3. In general, the three most important agronomic traits positively affected the yield of Tartary buckwheat in 2018 and 2019 planting seasons, and the overall effects were not linear. Except for PH in 2018, the main effect curves all exhibited an approximate logistic function shape, which kept increasing until thresholds were reached and then decreased in enhancements to the yield as the curves tended to be stable in their final stages. Comparing the results in 2018 and 2019, the main effect curves for NGP in the two seasons

shared similar shape, and the curves for traits PH and TGW in the 2018 planting season were apparently steeper than those in the 2019 planting season.

The original and bootstrap main effect curves were close to each other in each plot with almost the same variation trend, which indicated that the main effect estimators were unbiased. The width of the bootstrap confidence intervals for GNP, PH and TGW was generally small, indicating that the ALE main effect estimators were efficient. The width of the intervals increased, and the distances between the bootstrap and the original main effect curves increased at the two margins of distribution, which implied that the main effect estimations for the three important traits were subject to increased variability in these regions. These observations showed great conformity with the distributions of the corresponding traits in Figure 2. When the distribution of the sample points became sparse, the width of the confidence intervals grew large.

The ALE interactive effects amongst the three most important agronomic traits to the yield of Tartary buckwheat in the 2018 and the 2019 planting seasons are shown in Figure 4. In each plot, the deeper colour represented a higher yield of Tartary buckwheat. As shown in Figures 4A and 4D, the interactive effects between NGP and PH in 2018 and 2019 shared a similar pattern within the common area of the two traits. Generally, high NGP and PH led to high yield. When PH was small, the NGP should also be small to maintain the desired yield. However, from Figure 4A, the compatible NGP became small when PH was large. This finding was not clear in 2019, which had smaller ranges of PH and NGP than in 2018. Comparing Figures 4B and 4E, the pattern of the interaction effects of NGP and TGW was generally stable from the 2018 planting season to the 2019 planting season. In general, the colour became deep at the diagonal of the two-dimensional predictor space. The

combinations of large NGP and TGW corresponded to high yield of Tartary buckwheat. However, from Figure 4B, as NGP reached its maximum, increasing TGW cannot increase the yield. This result was not apparent in Figure 4E, which showed that the range of NGP in 2019 was smaller and that the range of TGW in 2019 shifted left compared with those in 2018. The interactive effects of TGW and PH in 2018 and 2019 are shown in Figures 4C and 4F. The interactive effects varied greatly in 2018. When PH was small, increasing TGW decreased the yield of Tartary buckwheat. As PH increased, increasing TGW improved the yield, but a sharp decrease in the yield was observed when PH increased further. The pattern of the effect in 2019 was similar to that in 2018 within the common range of TGW and PH.

Evaluation using classical methods

Correlation analysis

The Pearson's linear correlation, Spearman's rank correlation and the partial correlation coefficients between the yield and main agronomic traits were calculated (Figure 5). The yield-related traits NGP and TGW were highly significant with all the correlation measures and in both seasons ($p < 0.01$). For the 2018 planting season, the plant phenotypes (i.e., PH, SD, NN and BH) were highly significant with the Pearson's and Spearman's correlations ($p < 0.01$), and neither of their partial correlations was significant ($p > 0.05$). For the 2019 planting season, PH and NN were highly significant with the Pearson's correlation ($p < 0.01$), and PH, NN, GP and GAR were highly significant with the Spearman's correlations ($p < 0.01$) but became not significant with the partial correlation ($p > 0.05$). Overall, the colours were lighter in the partial correlation than in the other two correlations for both seasons, indicating that the association structures between the yield and the main agronomic traits

were hindered to some extent by the confounding effects. For the 2018 or the 2019 season, the significance of the Pearson's and Spearman's correlations were generally in a similar pattern. However, in 2018, the NB was significant when evaluated by the Spearman's correlation, whereas the GL became not significant ($p > 0.05$). In 2019, GP, GAR and GR became significant in the Spearman's correlation ($p < 0.01$).

Under the Pearson's or the Spearman's correlation, the traits SD and BH were highly significant in 2018 ($p < 0.01$) and became not significant in 2019 ($p > 0.05$), whereas the nonsignificant traits GAR and GR in 2018 ($p > 0.05$) became significant in 2019 ($p < 0.05$). Under the partial correlation, the relationships between yield and main agronomic traits varied slightly between the two years except GD, which changed from not significant to significant in 2019 ($p < 0.05$).

Multiple linear regression

For the multiple linear regression method, the best subset selection technique was used to simplify the model and identify the most influential agronomic traits for each planting season. Figure 6 shows the scaled values of the AIC and the BIC statistics for the 'best' regression models over different numbers of predictors. For the 2018 planting season, the BIC and the AIC statistics reached their lowest values when the numbers of predictors were 2 and 4, respectively. For the 2019 planting season, the BIC and the AIC statistics had the lowest values when the numbers of predictors were 4 and 7, respectively. Compared with the AIC statistic, the BIC statistic placed a heavier penalty on the model size, thereby resulting in the selection of smaller linear model for each season. The estimation results of the 'best'

regression models under the AIC and the BIC statistics for the two seasons are summarised in Table 2.

These multivariate linear regression models fit the data well. The adjusted R-squared values under AIC and BIC statistics were 0.9821 and 0.9817, respectively, in 2018, and 0.9497 and 0.9481, respectively, in 2019. For the 2018 season, NGP, TGW, GD and grey GC were selected as the important predictors in the AIC model. As shown by the regression coefficients, a grey GC had negative effect on the yield, and raising the other three traits increased the yield linearly. Compared with GD and GC, NGP and TGW exhibited the most significant effects on the yield of Tartary buckwheat. In the BIC model, only NGP and TGW were preserved. Similar to that under the AIC model, the linear effect of NGP on the yield of Tartary buckwheat in the BIC model was apparently larger than the linear effect of TGW. The grain-related traits were generally gaining significance in the 2019 planting season. GA and GD entered into the BIC model, and GA, GP, GD and brown GC into the AIC model in 2019. NGP and TGW were still highly significant under the AIC and the BIC models. Comparing the coefficients in the BIC model of 2019, GD and GA had larger effects on the yield than the yield-related traits NGP and TGW. The negative effect of GA and positive effect of GD were significant at the level of 0.01.

The two BIC models included no qualitative trait, and their intercepts were not significant. The AIC models in 2018 and 2019 contained the qualitative trait GC. The intercept of the AIC model in 2019 was significantly not equal to zero ($p < 0.05$). The intercept of the AIC model in 2018 was not significant, but the standard error of the intercept was almost in the same scale with the estimates.

The multiple linear regression models for the 2018 and the 2019 seasons with the three remarkable traits ranked highly by the random forest model were fitted. Results are shown in Table 3. NGP and TGW were still highly significant in the linear regression models of 2018 and 2019 with positive effects on the yield ($p < 0.001$). However, the effect of PH was not significant in the 2018 model ($p > 0.1$) and weakly significant in the 2019 model ($p < 0.1$).

Table 4 shows the results of the fitted multiple regression models for the 2018 and the 2019 planting seasons with additional consideration on the second-order interaction terms of the three remarkable agronomic traits from the random forest model. In the 2018 season, only the interaction effects of NGP and TGW were significant ($p < 0.001$). Increasing NGP or TGW, the positive linear effect of the other on the yield of Tartary buckwheat may be enhanced linearly. In the 2019 season, the interaction effects of NGP and TGW were still highly significant ($p < 0.001$). In addition, the interaction effects of TGW and PH became slightly significant ($p < 0.1$). The positive linear effect of TGW was enhanced as PH increased.

DISCUSSION

Model comparison

NGP and TGW are the most significant agronomic traits for both planting seasons as evaluated by the random forest regression model, the correlation analyses and multiple linear regression. NGP and TGW had positive and monotonic effects on the yield of Tartary buckwheat. However, the random forest model shows that this effect may not be linear. The other important traits distinguished by the three methods are generally different. Some grain-related traits regarded as significant by the linear regression model and the Pearson's and

Spearman's correlation analyses in the 2018 and 2019 planting seasons show weak importance in the random forest model. PH is evaluated as important in the random forest models but is not distinguished by the best subset selection and the partial correlation analysis in 2018 and 2019. In addition, the evaluation results by the random forest model show great conformity between the two seasons even though the distributions of the traits have changed. The results in the linear regression model and the Pearson's and Spearman's correlations generally show less consistency between the two planting seasons, which indicate to some extent the robustness and reliability of the random forest model (Everingham et al., 2016, Zou et al., 2017).

The Pearson's and the partial correlation analyses can only evaluate the linear effects of the traits, but the Spearman's correlation is suitable for all monotonic relationships. In our study, the results given by the Pearson's and the Spearman's correlation analyses are synchronous generally. The significance of the former should imply the significance of the latter, but exceptions remain in our study. From Figure 5, the Pearson's coefficient between GL and yield of Tartary buckwheat is significant ($p < 0.05$), whereas the Spearman's coefficient is not significant ($p > 0.05$). An inspection on the frequency distribution of GL shows that the underlying Gaussian distribution assumption may not be satisfied, which may be one cause of the abnormal pattern. In the partial correlation, NGP and TGW are significantly correlated with the yield of Tartary buckwheat in 2018 and 2019. The other traits significant in the Pearson's and Spearman's correlation analyses have become not significant. By excluding the confounding effects from the Pearson's correlation, the partial correlation should give a more reliable assessment of the linear effects.

Compared with the correlation analyses, the qualitative traits GC and GS are evaluated together with other quantitative traits in the multiple linear regression model. GC has significant effects on the yield of Tartary buckwheat in the AIC models of both seasons. However, the association evaluation by the linear regression model is still a measurement of the linear relationship. In our study, the regression models through the best subset selection in both seasons do not distinguish PH, which is ranked highly by the random forest models after NGP in both seasons, as significant. Table 3 shows that the linear regression coefficients of PH are not significant in the linear regression model of 2018 ($p > 0.1$) with the three most important traits selected by the random forest model and weakly significant in that of 2019 ($p < 0.1$). When the linear regression model only contains the quantitative traits, the intercept of the model based on the normalised data is expected to be zero. This phenomenon is observed in the BIC models for the two seasons, where the intercepts are not significant. However, in the AIC model in 2019, the intercept is significantly not zero with the qualitative trait GC involved, which may influence the interpretation of the regression model. Generally, the normalisation procedures for the linear regression model when the qualitative traits are involved need careful consideration.

In a multiple linear regression model, the second-order interaction term shows the moderation effect of one predictor to the linear effect of the other to the response (Morgan-Lopez and Mackinnon, 2006, Villavicencio and Bernardo, 2013). Table 4 shows that the interaction term of NGP and TGW is highly significant in both seasons. The positive linear effects of NGP to yield are enhanced linearly by increasing TGW. As shown in Figures 4B and 4E, given a certain level of TGW, the effects of NGP may not be linear. When fixing on a certain level of NGP, the yield can increase and then decrease as TGW increases. In

addition, adding interaction terms as new predictors to the model raises the degree of freedom, which may result in increased model variability and impaired reliability of results with limited sample size.

Association between yield and main agronomic traits

The association between the yield and the main agronomic traits of Tartary buckwheat evaluated by the random forest models in the two planting seasons shows that NGP, PH and TGW have remarkably major effects on the yield of Tartary buckwheat. The ALE main effect plots show that NGP, PH and TGW have positive effects on the yield of Tartary buckwheat. Their main effects on yield generally exhibit a logistic function shape in the two seasons, which indicate an appropriate period for each trait when the enhancement to the yield is relatively apparent and efficient. The ALE second-order interaction plots show that the three traits should be improved accordingly for yield enhancement. The PH of Tartary buckwheat should be optimised simultaneously with increased NGP or TGW, and the compatible NGP and TGW should be modified as PH increases.

These observations coincide with the growth and developmental habits of crops. Higher plants can provide more potential for efflorescence and seeding. Several studies have shown that the yield may be reduced when plants are shortened too much (Berry et al., 2007, Li et al., 1996, Wang et al., 2008). However, plant breeders often recommend landraces with moderate height to reduce the lodging risk. The yield of crops is greatly dependent on the lodging resistance capacity of the plant (Stapper and Fischer, 1990). Increasing the NGP and TGW of crops and ignoring the improvement of the lodging resistance capacity of the plant may harm the yield. Some internal and external factors may stimulate the vegetative growth

of the crops, leading to abnormally high plants and intensive inhibition of the reproductive growth of plants under limited resources and nutrition (Huett, 1996, Leonard, 1962). Hence, the compatible NGP and TGW have decreased when the PH is large, as observed in the ALE interaction plots.

CONCLUSION

In this paper, we applied the random forest regression model to evaluate the association between the yield and the 16 main agronomic traits of Tartary buckwheat on the basis of the data from two planting seasons. The random forest model showed that NGP, PH and TGW were the most important traits to the yield of Tartary buckwheat in both seasons. NGP had overwhelmingly large effects on the yield of Tartary buckwheat, but the effects from PH and TGW were also distinguishable from the remaining traits.

The association structures between yield and main agronomic traits evaluated using correlation analyses and multiple linear regression for the two seasons were generally different from the random forest regression model. The latter method, which required less assumptions and was not confined to detecting a linear relationship, was more robust and reliable compared with the former methods. The effects of agronomic traits on the yield were visualised using the ALE main effect and interaction plots, which showed great conformity with the growth and developmental habits of crops.

The key findings of this paper support the use of data mining and machine learning technologies for scientific guidance on various agricultural practices by considering the expanded volume and increased dimensionality of the data coming with the innovation and advancement occurring within the agricultural system.

ACKNOWLEDGEMENTS

Jing Li and Baili Feng conceived the project and designed the research plans; Jing Li collected the data; Shan Feng, Jing Li and Guoqi Qian analyzed the data; Shan Feng and Jing Li wrote the article. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [31671631, 31501365]; Young Scientists Fund of the Natural Science Foundation of Shaanxi Province [2018JQ3045]; Shaanxi Province Modern Crops Seed Industry Project [20171010000004]; and Technical System Project of Minor Grain Crops Industry in Shaanxi Province [NYKJ-2018-YL19].

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

- Apley, D.W. 2016. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv:1612.08468 [stat.ME]*. <http://arxiv.org/abs/1612.08468>.
- Baba, K., R. Shibata and M. Sibuya. 2004. Partial correlation and conditional correlation as measure of conditional independence. *Australian & New Zealand Journal of Statistics* 46: 657-664. doi:10.1111/j.1467-842X.2004.00360.x.
- Berry, P.M., R. Sylvester-Bradley and S. Berry. 2007. Ideotype design for lodging-resistant wheat. *Euphytica* 154: 165-179. doi:10.1007/s10681-006-9284-3.
- Breiman, L. 2001. Random forests. *Machine learning* 45: 5-32. doi:10.1023/A:1010933404324.

- Davison, A.C. and D.V. Hinkley. 2013. *Bootstrap Methods and their Application*. Cambridge University Press, New York, United States.
- Diaz-Uriarte, R. and S.A. de Andres. 2005. Variable selection from random forests: application to gene expression data. *arXiv:q-bio/0503025*. <http://arxiv.org/abs/q-bio/0503025>.
- Everingham, Y., J. Sexton, D. Skocaj and G. Inman-Bamber. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development* 36: 27. doi:10.1007/s13593-016-0364-z.
- Fieller, E.C., H.O. Hartley and E.S. Pearson. 1957. Test for rank correlation coefficients. I. *Biometrika* 44: 470-481. doi:10.2307/2003183.
- Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29: 1189-1232. doi:10.1214/aos/1013203451.
- Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer Science+Business Media.
- Hauke, J. and T. Kossowski. 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae* 30: 87-93. doi:10.2478/v10117-011-0021-1.
- Huett, D.O. 1996. Prospects for manipulating the vegetative-reproductive balance in horticultural crops through nitrogen nutrition: a review. *Australian Journal of Agricultural Research* 47: 47-66. doi:10.1071/AR9960047.
- James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer Science+Business Media, New York.
- Leonard, E.R. 1962. Inter-relations of vegetative and reproductive growth, with special reference to indeterminate plants. *Botanical Review* 28: 353-410. doi:10.1007/BF02868688.
- Li, J., M. S. Hossain, H. Ma, Q. Yang, X. Gong, P. Yang, et al. 2020. Comparative metabolomics reveals differences in flavonoid metabolites among different coloured buckwheat flowers. *Journal of Food Composition and Analysis* 85: 103335. doi:10.1016/j.jfca.2019.103335.
- Li, J., P. Yang, Q. Yang, X. Gong, H. Ma, K. Dang, et al. 2019. Analysis of flavonoid metabolites in buckwheat leaves using UPLC-ESI-MS/MS. *Molecules* 24: 1310. doi:10.3390/molecules24071310.
- Li, R., T. Jiang, T. Qiu, C. Cui and Z. Gong. 1996. Study on effect of lodging to yield and relationship between lodging and plant height in rice. (In Chinese, with English abstract.) *HeiLongjiang Agriculture Science*: 13-17.
- Li, X. 2013. Using "random forest" for classification and regression. (In Chinese, with English abstract.) *Chinese Journal of Applied Entomology* 50: 1190-1197. doi:10.7679/j.issn.2095-1353.2013.163.
- Li, Z., T. Zhang, X. Wu and K. Li. 2012. Methodology of regression by random forest and its application on metabolomics. (In Chinese, with English abstract.) *Chinese Journal of Health Statistics* 29: 158-160,163. doi:10.3969/j.issn.1002-3674.2012.02.001.

- Liaw, A. and M. Wiener. 2002. Classification and regression by random forest. *R News* 2: 18-22.
- Lu, M. 2007. Comparison analysis of several approaches for handling with the multicollinearity. (In Chinese, with English abstract.) *Statistics and Decision*: 8-10. doi:10.3969/j.issn.1002-6487.2007.13.004.
- Miller, A. 2002. *Subset Selection in Regression*. 2nd ed. Chapman and Hall/CRC United States.
- Morgan-Lopez, A.A. and D.P. Mackinnon. 2006. Demonstration and evaluation of a method for assessing mediated moderation. *Behavior Research Methods* 38: 77-87. doi:10.3758/BF03192752.
- Qin, P., Z. Liu, S. Wan, R. Su and J. Huang. 2016. Identification of yield limiting factors for oilseed rape in Hubei province based on CART and random forest model. (In Chinese, with English abstract.) *Chinese Journal of Agrometeorology* 37: 691-699. doi:10.3969/j.issn.1000-6362.2016.06.009.
- Sedgwick, P. 2012. Pearson's correlation coefficient. *BMJ (online)* 345: e4483. doi:10.1136/bmj.e4483.
- Smidt, E.R., S.P. Conley, J. Zhu and F.J. Arriaga. 2016. Identifying field attributes that predict soybean yield using random forest analysis. *Agronomy Journal* 108: 637-646. doi:10.2134/agronj2015.0222.
- Stapper, M. and R.A. Fischer. 1990. Genotype, sowing date and plant spacing influence on high-yielding irrigated wheat in Southern New South Wales. II growth, yield and nitrogen use. *Australian Journal of Agricultural Research* 41: 1021-1041. doi:10.1071/ar9901021.
- Villavicencio, F.T. and A.B.I. Bernardo. 2013. Positive academic emotions moderate the relationship between self-regulation and academic achievement. *British Journal of Educational Psychology* 83: 329-340. doi:10.1111/j.2044-8279.2012.02064.x.
- Wang, Q., R. Zhen, Y. Ding and S. Wang. 2008. Strong stem effect and physiological characteristics of rice plant under rice-duck farming. (In Chinese, with English abstract.) *Chinese Journal of Applied Ecology* 19: 2661-2665. <http://www.cjae.net/EN/Y2008/V19/I12/2661>.
- Yang, Y., W. Wu, Y. Zheng, J. Wang, J. Li and C. Wu. 2008. The correlation and path analysis on the seed yield per plant and main agronomic traits in tartary buckwheat (*Fagopyrum tataricum*). (In Chinese, with English abstract.) *Journal of Anhui Agricultural Sciences* 36: 6719-6721, 6746. doi:10.3969/j.issn.0517-6611.2008.16.050.
- You, S. and Y. Yan. 2017. Stepwise regression analysis and its application. (In Chinese, with English abstract.) *Statistics and Decision*: 31-35. doi:10.13546/j.cnki.tjyj.2017.14.007.
- Zhang, L., L. Wang, X. Zhang, S. Liu, P. Sun and T. Wang. 2014. The basic principle of random forest and its applications in ecology: a case study of *Pinus yunnanensis*. (In Chinese, with English abstract.) *Acta Ecologica Sinica* 34: 650-659. doi:10.5846/stxb201306031292.

Zhu, J., W. Yang and W. Su. 2008. Analysis of the driving influence mechanism of Linli county based on the multiple regression analysis law and the latus rectum analysis. (In Chinese, with English abstract.) *Economic Geography* 28: 488-491,507.

Zou, Y., R. Wang and W. Li. 2017. Application of the random forest algorithm in wheat breeding evaluation. (In Chinese, with English abstract.) *Computer Systems & Applications* 26: 181-185. doi:10.15888/j.cnki.csa.006162.

Figure captions:

Figure 1. The importance rankings of the main agronomic traits to yield of Tartary buckwheat in planting seasons 2018 (A) and 2019 (B) based on the mean decrease in node impurity (MDI) of the random forest model. The most important traits, namely, the number of grains per plant (NGP), plant height (PH) and 1000-grain weight (TGW) are presented in red.

Figure 2. Frequency distributions of the remarkable traits, namely, the number of grains per plant (NGP), plant height (PH) and 1000-grain weight (TGW) in the 2018 (A, B and C) and the 2019 planting seasons (D, E and F). The enlarged kernel and the normal density curves with mean and standard deviation as the sample mean and standard deviation of the corresponding trait are drawn in black and red, respectively. The red bars below indicate the distribution of sample points.

Figure 3. The main effects of the remarkable traits, namely, the number of grains per plant (NGP), plant height (PH) and 1000-grain weight (TGW), calculated by the accumulated local effects (ALE) method for the 2018 (A, B and C) and the 2019 (D, E and F) planting seasons. The original main effect curve, bootstrap main effect curve and the 95% bootstrap confidence intervals produced from 300 replicates are drawn in black, red and grey, respectively, in each panel. The ALE function values of the yield at the y-axis is meaningful in a relative sense because a centralisation process is involved. The red bars at the bottom of each plot show the distribution of sample points.

Figure 4. ALE second-order interaction plots for the number of grains per plant (NGP), plant height (PH) and 1000-grain weight (TGW) in the 2018 (A, B and C) and the 2019 (D, E and F) planting seasons. A and D: the interactive effects of NGP and PH; B and E: NGP and TGW; C and F: TGW and PH.

Figure 5. Comparison of the significance of Pearson's linear correlation, Spearman's rank correlation and partial correlation between yield and the quantitative traits of Tartary buckwheat during the 2018 and the 2019 planting seasons. Red and blue colours represent positive and negative correlations, respectively. The colour depth is proportional to the value of $1 - p$ -value. * and ** indicate significance at the levels of 0.05 and 0.01, respectively.

Figure 6. Variability of the scaled AIC and BIC statistics for the ‘best’ multiple linear regression models of different sizes for the 2018 (A) and the 2019 (B) planting seasons. The subplot on the upper right corner of each plot is an enlarged version around the part with the smallest BIC and AIC values.

Tables:

Table 1. Agronomic traits of Tartary buckwheat investigated in the experiment.

Classification	Agronomic traits	Units
Yield-related traits	Yield per hectare (Y)	kg ha ⁻¹
	Number of grains per plant (NGP)	-
	1000-grain weight (TGW)	g
Plant phenotypes	Plant height (PH)	cm
	Stem diameter (SD)	mm
	Number of nodes (NN)	-
	Number of branches (NB)	-
	Branch height (BH)	cm
Grain-related traits	Grain area (GA)	mm ²
	Grain perimeter (GP)	mm
	Grain aspect ratio (GAR)	-
	Grain length (GL)	mm
	Grain width (GW)	mm
	Grain diameter (GD)	mm
	Grain roundness (GR)	-
	Grain color (GC)	-
	Grain shape (GS)	-

Table 2. The ‘best’ regression models obtained using the best subset selection under AIC and BIC for the 2018 and the 2019 planting seasons.

Year	AIC				BIC			
	Traits	Est.	SE	<i>p</i> -value	Traits	Est.	SE	<i>p</i> -value
2018	(Intercept)	0.013	0.012	0.268	(Intercept)	0.000	0.010	1.000
	NGP	0.969	0.010	0.000***	NGP	0.962	0.010	0.000***
	TGW	0.201	0.013	0.000***	TGW	0.219	0.010	0.000***
	GD	0.027	0.014	0.058‡				
	GCgrey†	-0.037	0.020	0.067‡				
2019	(Intercept)	-0.096	0.039	0.015*	(Intercept)	0.000	0.016	1.000
	NGP	0.935	0.017	0.000***	NGP	0.936	0.017	0.000***
	TGW	0.171	0.020	0.000***	TGW	0.178	0.019	0.000***
	GA	-0.673	0.420	0.110	GA	-1.095	0.381	0.005**
	GP	-0.081	0.050	0.103	GD	1.133	0.382	0.003**
	GD	0.791	0.402	0.050‡				
	GCbrown†	0.116	0.046	0.013*				

GCgrey 0.115 0.047 0.016*

Est. represents the estimates of the regression coefficient, and SE is the standard error of the coefficient estimator.

† GCgrey and GCbrown represent the dummy variables for GC: GCgrey (GCbrown) is equal to 1 only if GC is grey (brown), otherwise equal to 0.

‡, *, ** and *** represent significance at the levels of 0.1, 0.05, 0.01 and 0.001, respectively.

Table 3. Linear regression models for the 2018 and the 2019 planting seasons with the three most remarkable traits from the random forest model.

Year	Traits	Est.	SE	p-value
2018	(Intercept)	0.000	0.010	1.000
	NGP	0.966	0.011	0.000 ***
	PH	-0.007	0.011	0.535
	TGW	0.218	0.010	0.000 ***
2019	(Intercept)	0.000	0.016	1.000
	NGP	0.928	0.018	0.000***
	PH	0.035	0.018	0.055‡
	TGW	0.204	0.016	0.000***

Est. represents the estimates of the regression coefficient, and SE is the standard error of the coefficient estimator.

‡ and *** represent significance at the levels of 0.1 and 0.001, respectively.

Table 4. Linear regression models for the 2018 and the 2019 planting seasons with the three most remarkable traits from the random forest model and their second-order interaction terms.

Year	Traits	Est.	SE	<i>p</i> -value
2018	(Intercept)	0.002	0.008	0.809
	NGP	0.967	0.009	0.000***
	PH	-0.010	0.009	0.276
	TGW	0.219	0.008	0.000***
	NGP*TGW	0.105	0.010	0.000***
	NGP*PH	-0.010	0.008	0.190
	TGW*PH	-0.006	0.009	0.482
2019	(Intercept)	-0.003	0.014	0.836
	NGP	0.947	0.015	0.000 ***
	PH	0.022	0.015	0.145
	TGW	0.253	0.014	0.000 ***
	NGP*TGW	0.122	0.013	0.000 ***
	NGP*PH	-0.009	0.013	0.486
	TGW*PH	0.027	0.014	0.053‡

Est. represents the estimates of the regression coefficient, and SE is the standard error of the coefficient estimator.

‡ and *** represent significance at the levels of 0.1 and 0.001, respectively.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Feng, S;Li, J;Qian, G;Feng, B

Title:

Association between the yield and the main agronomic traits of Tartary buckwheat evaluated using the random forest model

Date:

2020-09-01

Citation:

Feng, S., Li, J., Qian, G. & Feng, B. (2020). Association between the yield and the main agronomic traits of Tartary buckwheat evaluated using the random forest model. *Crop Science*, 60 (5), pp.2394-2407. <https://doi.org/10.1002/csc2.20243>.

Persistent Link:

<http://hdl.handle.net/11343/276122>