remote sensing

MDPI

Technical Note

Identifying Key Traits for Screening High-Yield Soybean Varieties by Combining UAV-Based and Field Phenotyping

Chen Yang [1,2], Guijun Yang [1,2,3], Haorang Wang [4], Simeng Li [4], Jiaoping Zhang [5], Di Pan [2], Pengting Ren [2], Haikuan Feng [2] and Heli Li [2,*]

1 Research Institute of Quantitative Remote Sensing and Smart Agriculture, School of Surveying and Mapping Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China; 212204010040@home.hpu.edu.cn (C.Y.); yanggj@nercita.org.cn (G.Y.)
2 Key Laboratory of Quantitative Remote Sensing in Agriculture of Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China; gqt2400205007@student.cumtb.edu.cn (D.P.); renpt@sxau.edu.cn (P.R.); fenghk@nercita.org.cn (H.F.)
3 College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China
4 Jiangsu Xuhuai Regional Institute of Agricultural Sciences, Xuzhou 221131, China; 20211009@jaas.ac.cn (H.W.); lisimeng@jaas.ac.cn (S.L.)
5 State Key Laboratory of Crop Genetics and Germplasm Enhancement, MARA National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing 210095, China; zjp@njau.edu.cn
* Correspondence: lihl@nercita.org.cn

Abstract: The breeding of high-yield varieties is a core objective of soybean breeding programs, and phenotypic trait-based selection offers an effective pathway to achieve this goal. The aim of this study was to identify the key phenotypic traits of high-yield soybean varieties and to utilize these traits for screening high-yield soybean varieties. In this study, the UAV (unmanned aerial vehicle)- and field-based phenotypic data were collected from 1923 and 1015 soybean breeding plots at the Xuzhou experimental site in 2022 and 2023, respectively. First, the soybean varieties were grouped by using a self-organizing map and K-means clustering to investigate the relationships between various traits and soybean yield and to identify the key ones for selecting high-yield soybean varieties. It was shown that the duration of canopy coverage remaining above 90% (Tcc90) was a critical phenotypic trait for selecting high-yield varieties. Moreover, high-yield soybean varieties typically exhibited several key phenotypic traits such as rapid development of canopy coverage (Tcc90r, the time when canopy coverage first reached 90%), prolonged duration of high canopy coverage (Tcc90), a delayed decline in canopy coverage (Tcc90d, the time when canopy coverage began to decline below 90%), and moderate-to-high plant height (PH) and hundred-grain weight (HGW). Based on these findings, a method for screening high-yield soybean varieties was proposed, through which 87% and 72% of high-yield varieties (top 5%) in 2022 and 2023, respectively, were successfully selected. Additionally, about 9% (in 2022) and 10% (in 2023) of the low-yielding (bottom 60%) were misclassified as high-yielding. This study demonstrates the benefit of high-throughput phenotyping for soybean yield-related traits and variety screening and provides helpful insights into identifying high-yield soybean varieties in breeding programs.

Keywords: UAV; phenotyping; varieties screening; cluster; soybean traits

check for updates

Academic Editor: Thomas Alexandridis

Received: 2 January 2025
Revised: 28 January 2025
Accepted: 17 February 2025
Published: 18 February 2025

Citation: Yang, C.; Yang, G.; Wang, H.; Li, S.; Zhang, J.; Pan, D.; Ren, P.; Feng, H.; Li, H. Identifying Key Traits for Screening High-Yield Soybean Varieties by Combining UAV-Based and Field Phenotyping. *Remote Sens.* **2025**, *17*, 690. https://doi.org/10.3390/rs17040690

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soybean (*Glycine max* L.), an important legume crop, is a main source of vegetable oil and plant proteins [1]. High-yield soybean variety screening is a focus of many soybean

breeding and research programs [2]. In order to obtain excellent soybean varieties, it is usually necessary to evaluate multiple phenotypic traits related to the target traits like yield [3]. However, traditional phenotypic data collection mainly relies on manual measurements, which are time-consuming, laborious, and difficult to support large-scale variety screening [4].

The rapid development of high-throughput phenotyping (HTP) provides new possibilities for soybean phenotypic research and excellent variety screening [5,6]. High-throughput phenotyping technology can collect and process phenotypic data of a large number of crops in a short time through image processing and data analysis techniques [7–9]. As an important high-throughput phenotyping platform, UAV technology has been widely used in crop phenotypic research because of its flexibility and efficiency [10], such as yield prediction [11,12], canopy coverage extraction [13,14], and lodging detection [15,16], etc. This dynamic and non-destructive monitoring method provides great convenience for the phenotypic data collection of soybean at different growth stages, helping breeders to screen out high-yield varieties in natural environments.

The number of pods per plant (NPP) and the hundred-grain weight (HGW) are two of the yield factors for soybean [17]. More NPP and larger HGW usually lead to higher yield. In recent years, many studies have shown that soybean yield was not only related to NPP and HGW, but also closely related to other phenotypic traits [18,19], and different phenotypic traits would have combined effects on the growth and yield of soybean. For example, Casagrande et al. [20] reported that varieties with higher canopy coverage usually have greater light interception and higher photosynthetic efficiency, which favors crop yield accumulation. Chen et al. [21] investigated the effect of plant height (PH) on soybean yield and found that the increase in PH and the number of nodes per plant (NNP) had positive roles in the increase in yield. Gao et al. [22] reported that increasing the number of branches per plant (NBP) can increase NPP to increase soybean yield. Besides, lodging resistance is also an important factor for yield stability. Shan et al. [23] found that the fifth internode of soybean was the key internode related to lodging. Therefore, understanding the relationships between soybean phenotypic traits and yield would help breeders to select key phenotypic features and improve the screening efficiency of high-yield varieties.

In addition, clustering analysis provides an effective approach for identifying and grouping soybean varieties with similar traits [24]. The self-organizing map (SOM) is a nonlinear dimensionality reduction method that maps high-dimensional data into a lower-dimensional space, intuitively revealing the topological structure of the data, though its clustering boundaries may be somewhat ambiguous [25,26]. K-means is a fast distance-based clustering method that groups samples through iterative optimization. It is characterized by simplicity and high efficiency but is sensitive to the initial cluster centers [27]. The combination of SOM with K-means can integrate the topological preservation ability of SOM and the grouping precision of K-means, significantly improving clustering stability and resolution [28]. This combined approach is highly effective for identifying key differences in traits in complex multidimensional phenotypic data.

Therefore, the main goal of this study was to provide a variety-screening framework by combining UAV-based and field phenotyping and to identify the key phenotypic features related to high soybean yield. Firstly, this study conducted a cluster analysis on numerous test varieties based on UAV and field phenotypic data. Then, the relationships between various traits and yield were investigated, and the key traits of high-yield soybean variety screening were determined. Finally, the selection strategy and results based on the key phenotypic traits for high-yield soybean varieties were provided.

2. Materials and Methods

The workflow of this study is presented in Figure 1, mainly including three parts: (1) UAV- and field-based data collection and preprocessing; (2) clustering and phenotypic analysis; and (3) screening methods of high-yield soybean varieties.

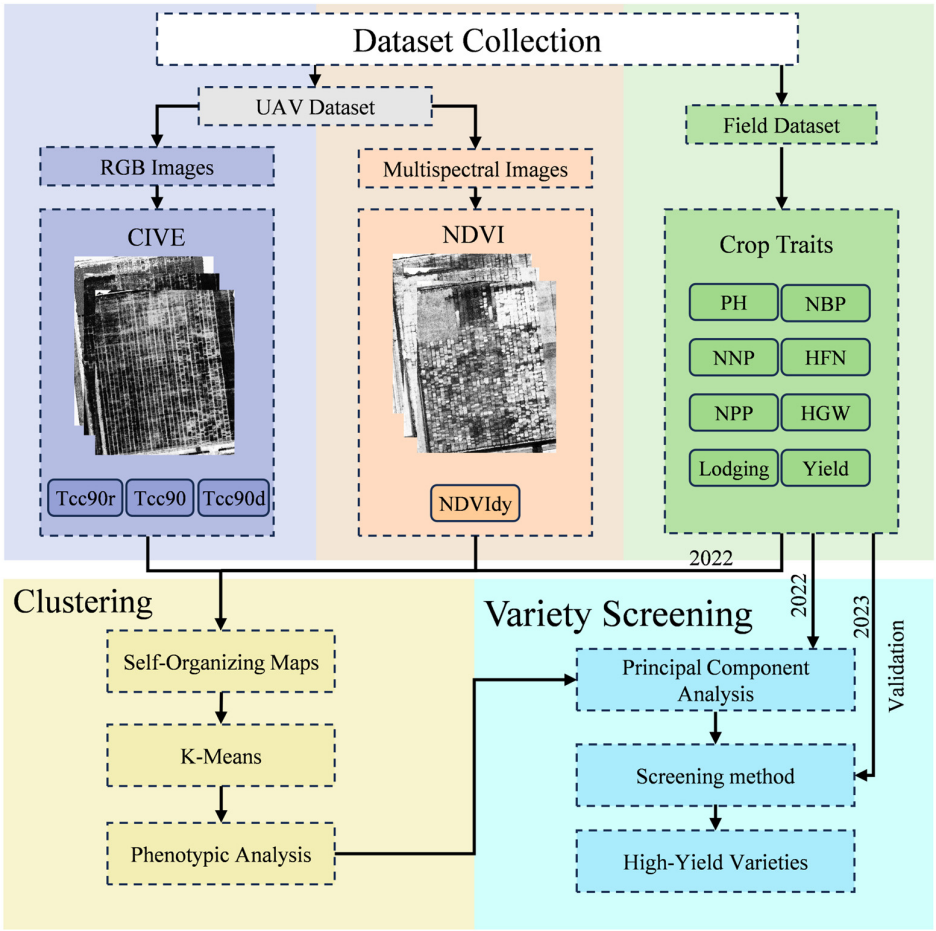

Figure 1. A workflow diagram of this study. Abbreviations in the figure: Tcc90r (time when canopy coverage first reached 90%), Tcc90d (time when canopy coverage began to decline below 90%), and Tcc90 (the duration of canopy coverage above 90%); PH (plant height), HFN (height of the fifth node), NBP (number of branches per plant), NNP (number of nodes per plant), NPP (number of pods per plant), and HGW (hundred-grain weight).

2.1. Field Experiments

The datasets used in this study were collected from soybean breeding trials conducted at the Xuzhou experimental site (117°24′35″E, 34°18′41″N) during the 2022 and 2023 growing seasons (Figure 2). Located in the southeastern part of the North China Plain, the region has a temperate monsoon climate, with an average annual temperature of 14 °C and annual precipitation of 800 mm. Soybean in this area is typically sown in late June and harvested in early October.

There were a total of 2280 different varieties, planted in two rows with a row length of 2 m and a row spacing of 0.5 m in both of the years. However, in the 2022 season, only 1923 varieties were investigated in this study, with the ones that were affected by poor germination or pests and diseases excluded. In 2023, the data from 1015 varieties were used in this work because the canopy appearance of the remaining ones was artificially intervened by the support bars in the plots. Yield measurements were made at matu-

rity after harvest. All plots were managed according to local field practices suitable for the region.

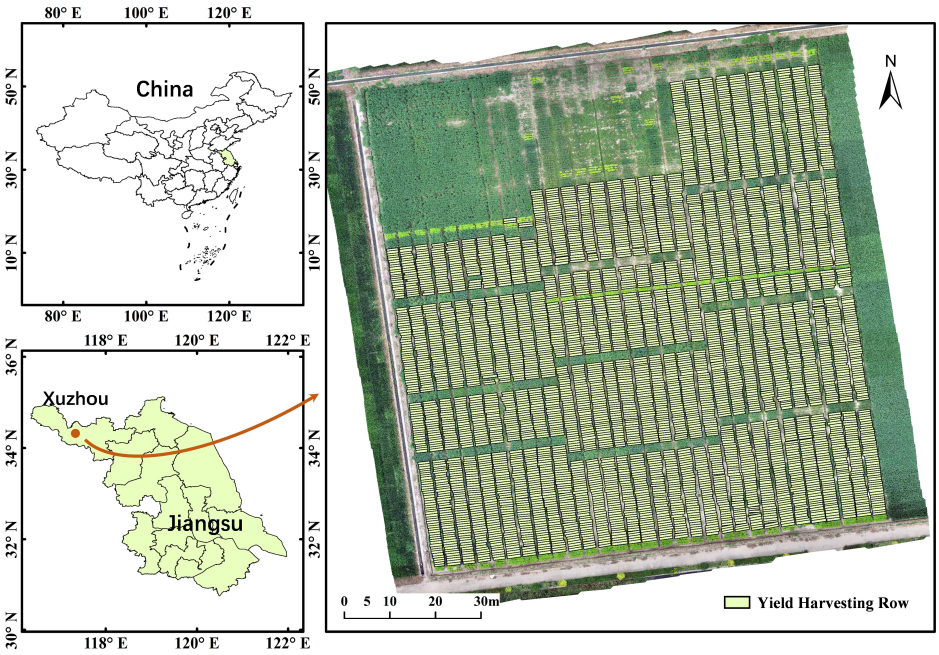

Figure 2. Field experiment in this study. The UAV image shown in the figure is from the experiment in 2022.

2.2. UAV Data Acquisition and Preprocessing

The UAV images used in this study included RGB images acquired by DJI P4Pro (SZ DJI Technology Co., Shenzhen, China) and multi-spectral images acquired by DJI P4M (SZ DJI Technology Co., Shenzhen, China). UAV missions were conducted every three days under windless and clear conditions between 11:00 a.m. and 1:00 p.m. To enhance the quality and accuracy of UAV data, both the heading overlap rate and side overlap rate were set to 80%. DJI Terra (SZ DJI Technology Co., Shenzhen, China) was used for image stitching, and Arc GIS 10.8 (Environmental Systems Research Institute, Inc., Redlands, CA, USA) was used for image correction and vegetation index calculation. RGB images were used to extract the color index of vegetation (CIVE) [29], and multispectral images were used to extract the normalized difference vegetation index (NDVI).

According to Maimaitijiang et al. [11], the best window period for soybean yield prediction was from blooming to seed filling. An investigation of the growth period in 2022 revealed that most field varieties were in this stage from DAS50 to DAS90 (i.e., DAS50, the subscript meant the day number after sowing). The NDVI, which reflects the photosynthetic potentiality and growth vigor of soybeans, has been widely utilized in yield estimation [30]. Given the diverse varieties in this study, the comparison of time-point data was heavily influenced by the difference in growth period. A previous study [5] also indicated that time-point data was not suitable for comparing cultivars. To address this, NDVI polylines were constructed based on the selected time points (from DAS50 to DAS90), and the area under these polylines was used to describe the NDVI of the key growth period of soybean, denoted as NDVIdy [3]. The calculation of NDVIdy was as follows:

$$NDVIdy = \sum_{i=1}^{n-1} \frac{(NDVI_i + NDVI_{i+1})}{2} (DAS_{i+1} - DAS_i) \quad (1)$$

where the $NDVI_i$ and DAS_i refer to the value of NDVI and DAS (days after sowing) obtained during the *i*-th measurement, respectively.

Canopy coverage was extracted using a threshold method based on CIVE [29,31]. A P-spline dynamic model was applied to analyze canopy coverage, identifying key time points. The canopy coverages were set at intervals of 5%, ranging from 50% to 95%. The time points included the time when canopy coverage first reached the value (T_{reach}), when it began to decline below the value ($T_{decline}$), and the duration of it above the value ($T_{duration}$). Pearson correlation coefficients between time-related traits at different canopy coverage levels and yield were calculated. The optimal time-related traits were observed when canopy coverage was 90% (Figure 3), such as when canopy coverage first reached 90% (Tcc90r), when it began to decline below 90% (Tcc90d), and the duration of canopy coverage above 90% (Tcc90). The high frequency of flights (i.e., every 3 days) made the extracted time points with an error less than or equal to 1 day.

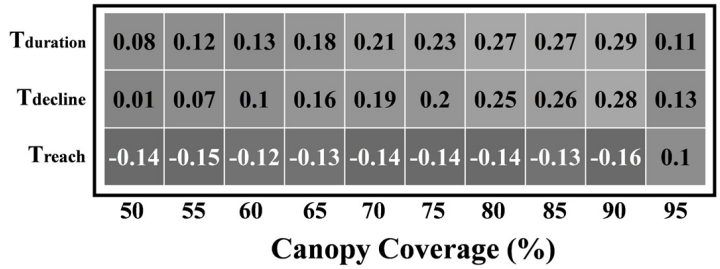

Figure 3. Pearson correlation coefficients between time-related traits at different canopy coverage levels and yield.

2.3. Field Data Acquisition

At maturity, 10 typical plants were randomly selected from each plot. Plant height (PH) and height of the fifth node (HFN) were measured using a steel ruler, and the number of branches per plant (NBP), number of nodes per plant (NNP), and number of pods per plant (NPP) were also recorded. The seeds were placed into corresponding labeled paper bags, first dried at 105 °C for 30 min, then dried at 80 °C to a constant weight, and the hundred-grain weight (HGW) was measured. Finally, the mean values of these traits were calculated as the field phenotypic traits of the soybean. Besides, the yield of each variety was also measured. The test varieties were categorized by yield level into high-yield varieties (top 5%), low-yield varieties (bottom 60%), and medium-yield varieties (other).

The lodging degree was another critical trait recorded in this study. Specifically, lodging was rated into five levels based on the proportion of the lodging area, with a step size of 25%. Level 1 indicates no lodging, where the plants remain upright and exhibit strong lodging resistance, and level 5 means there are more than 75% of the plants lodging.

2.4. Self-Organizing Map and K-Means Clustering Analysis

In this study, we combined self-organizing maps (SOM) with the K-means clustering algorithm to establish a two-step clustering framework for grouping the test soybean varieties based on their phenotypic traits. SOM is an unsupervised learning algorithm that can effectively capture the topological structure of data, allowing data points to naturally cluster into different regions on a two-dimensional plane. Firstly, SOM was used as pre-clustering in this study to gradually map high-dimensional data into two-dimensional space through adaptive learning. The parameters of the SOM were set based on the report by Han et al. [3]: the SOM size was set to 16 × 8, the learning rate was 0.5, and the number of iterations was 10,000. Then, the output from SOM was used as the initial input for the K-means algorithm, and the Dunn index was applied to determine the optimal number of clusters, further refining the preliminary clusters. In this way, the category structure can be fine-tuned, ensuring that similar data points are clearly grouped into the same category.

The two-step clustering strategy can generate a more precise grouping, aiming to divide the test soybean varieties into distinct categories based on their phenotypic traits. Both clustering methods were built in Python, with the Self-Organizing Map (SOM) algorithm using the MiniSom library and the K-means algorithm using the scikit-learn library.

2.5. Principal Component Analysis

In this study, principal component analysis (PCA) was used to analyze the phenotypic data of soybean to identify the main trends and relationships between different traits. After classifying the test soybean varieties, we standardized the data of multiple phenotypic traits to ensure that different traits had equal weight in the analysis. Then, the projection of each trait on the principal component was calculated by PCA, and the biplot was drawn to show the distribution of different soybean varieties in the principal component space. The biplot can clearly show the phenotypic differences and similarities between varieties, and can also help explain which traits contribute more to each principal component. Finally, the potential effects of soybean phenotypic traits on yield can be more intuitively understood by the results. The PCA algorithm was performed using the Origin 2021 software program (OriginLab Inc., Northampton, MA, USA)

3. Results

3.1. Analysis of Phenotypic Traits

A multidimensional statistical analysis of soybean phenotypic traits was conducted using the data from the 2022 season, and the results are shown in Figure 4.

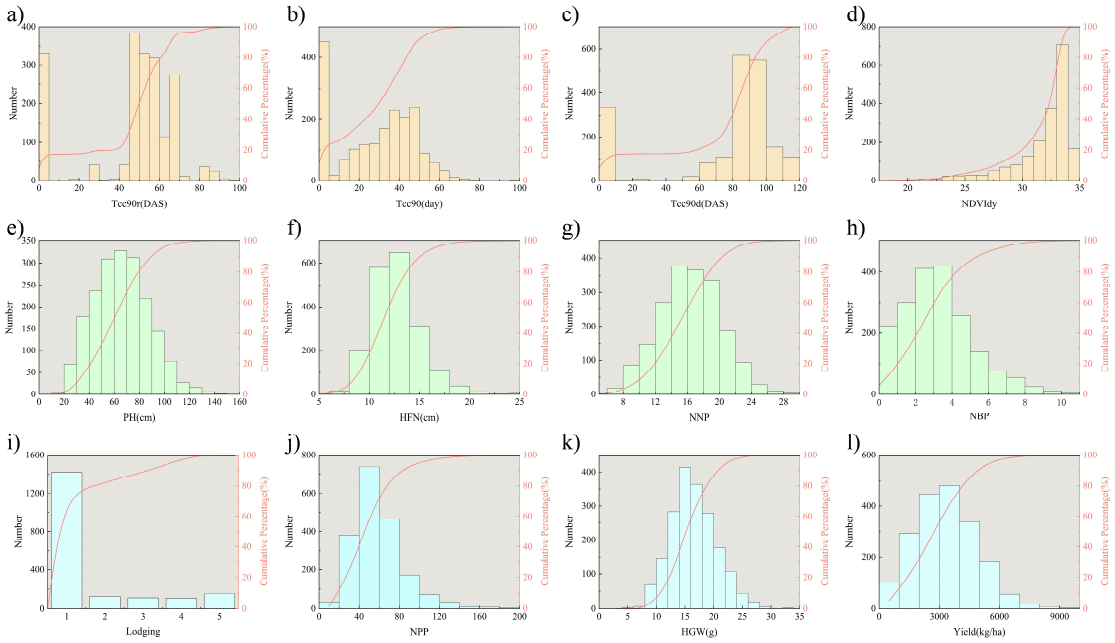

Figure 4. Statistics of phenotypic traits. Sub-panels (**a**–**d**) present UAV-derived phenotypic traits, including Tcc90r, Tcc90d, Tcc90, and NDVIdy; sub-panels (**e**–**h**) present plant-type traits, including PH (plant height), NBP (number of branches per plant), NNP (number of nodes per plant), and HFN (height of the fifth node); and sub-panels (**i**–**l**) present the traits of lodging, NPP (number of pods per plant), HGW (hundred-grain weight), and yield.

The top sub-panels (Figure 4a–d) were the UAV-based phenotyping traits, including Tcc90r, Tcc90, Tcc90d, and NDVIdy. The maximum canopy coverage of some varieties was less than 90%, and Tcc90r, Tcc90, and Tcc90d were recorded as 0. These traits were about the dynamic changes in soybean canopy coverage and spectral reflectance. As non-

destructive indicators of soybean growth, they could provide valuable information for assessing yield potential. The middle sub-panels (Figure 4e–h) were the traits related to plant morphological structure, including plant height (PH), height of the fifth node (HFN), number of branches per plant (NBP), and number of nodes per plant (NNP), which are related to productivity, resource allocation, and final yield [18]. The bottom sub-panels (Figure 4i–l) presented the yield and the traits that directly influence soybean productivity, such as lodging, number of pods per plant (NPP), and hundred-grain weight (HGW).

These traits were classified into polytomous and numerical categories based on data types. Lodging was a categorical variable, and the overall performance of test varieties in this regard was better. Most varieties had no lodging, and a few varieties had serious lodging. The numerical traits in this study were normal distribution. This distribution pattern reflected the diversity and randomness of population genetics and environment and also indicated that the test soybean varieties allowed the screening of high-yield varieties based on phenotypic traits.

3.2. Clustering of the Test Soybean Varieties Based on Phenotypic Traits

By using the SOM and K-means clustering, the 1923 soybean varieties from the 2022 season were divided into five categories (Cluster 1 to Cluster 5) based on their phenotypic traits. Figure 5 shows the yield distribution of the five categories, and the yield range and mean value of each category were significantly different. The performance of Cluster 3 was particularly significant, with an average yield of only 1700 kg/ha, and the yield of only a few varieties reached the average level of all varieties (3300 kg/ha). The mean yield of Cluster 4 was 4200 kg/ha, which was the highest among the five categories. The yield distribution of Cluster 4 showed a normal distribution near its mean, while the yield distributions of the other four categories were concentrated in the mean and below. It was feasible to use a clustering algorithm to screen low- and high-yield soybean varieties based on phenotypic traits.

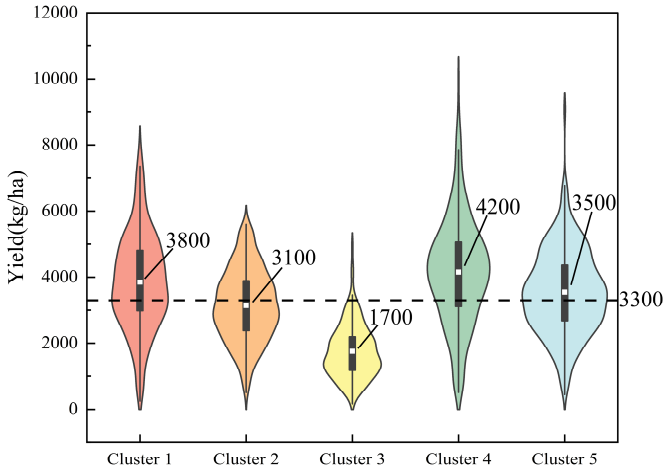

Figure 5. The yield distribution of five types of soybeans obtained by clustering. The average yield of all soybean varieties was 3300 kg/ha. The black box indicates the lower quartile to the upper quartile, and the white square represents the mean of each cluster. Larger contours indicate a higher concentration of data in that area, while smaller contours suggest fewer data in that region.

The 11 investigated phenotypic traits of the five soybean clusters, including Tcc90r, Tcc90d, Tcc90, NDVIdy, PH, HFN, NBP, NNP, Lodging, NPP, and HGW, are illustrated in Figure 6. To facilitate comparison and visualization, the traits of the five soybean clusters were normalized by dividing the mean value of each cluster by the mean value of all the soybean varieties. There was little difference in the performance of the five

categories on Tcc90r, NDVIdy, HFN, and HGW. The performance of Tcc90, Tcc90d, and NBP showed the same change as the yield, indicating that extending the duration of high canopy coverage and delaying its decline, along with increasing the number of branches, might have enhanced soybean yield. However, the performance of PH, NNP, and NPP differed from their yield trends. The performance of Cluster 1 outperformed Cluster 4 in these three phenotypic traits, yet its yield was lower. This might have been due to excessively high PH, which led to greater NNP and NPP but also increased the risk of lodging [32]. Lodging typically had a negative impact on yield [33].

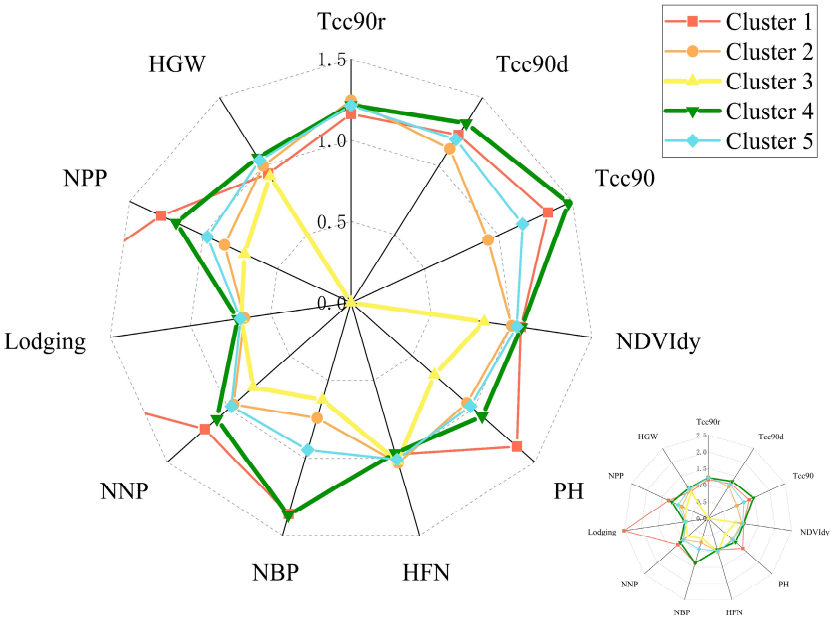

Figure 6. Comparison of 11 phenotypic traits across the five categories of soybeans. To facilitate analysis and visualization, the values of the traits were normalized by dividing the mean value of each cluster by the overall mean value of all soybean varieties.

The results showed that Cluster 3, a category of low-yield variety, was screened out, and its yield and various traits were significantly lower than the average. However, its most significant difference compared to the other four categories was its low canopy coverage (Tcc90r, Tcc90d, and Tcc90 were all 0). The persistence of high canopy coverage not only enhanced photosynthetic efficiency but also provided a stable physiological foundation for yield [19]. This highlighted the critical role of canopy coverage in soybean yield. Furthermore, Tcc90r, Tcc90d, and Tcc90 demonstrated significant effects when low-yield varieties were screened.

3.3. Key Traits for Screening High-Yield Varieties

As mentioned above, Cluster 4 showed relatively higher yields compared to the other categories, but this clustering did not screen out the high-yield varieties effectively. It was noted that the high-yield varieties (i.e., yields beyond 6000 kg/ha, top 5% of the test varieties) were found in Clusters 1, 4, and 5. To determine the traits of high-yield varieties, the three clusters were redivided into three yield-level groups: low-yield (<3000 kg/ha, bottom 60% of the test varieties), medium-yield (3000–6000 kg/ha), and high-yield (>6000 kg/ha, top 5% of the test varieties).

To facilitate comparison and visualization, the traits of the three yield-level groups were normalized by dividing the mean value of each cluster by the mean value of all the soybean varieties (Figure 7). It was observed that among these traits, high-yield varieties (>6000 kg/ha) exhibited superior performance in NPP and HGW, which directly

contributed to their high yield. These varieties showed lower Tcc90r and higher Tcc90 and Tcc90d values. This indicated that high-yield soybean varieties exhibited rapid canopy development, prolonged duration of high canopy coverage, and a delayed decline in canopy coverage. Additionally, the high-yield group demonstrated notable advantages in traits such as PH, NBP, and NNP. The medium-yield group (3000–6000 kg/ha) displayed trait performance that fell between the high- and low-yield groups, and the low-yield group (0–3000 kg/ha) performed poorly in most traits. Among all tested traits, Tcc90 exhibited the greatest variation, suggesting it may be one of the key factors influencing yield. These results indicated that canopy-related traits should be paid more attention in soybean variety screening.

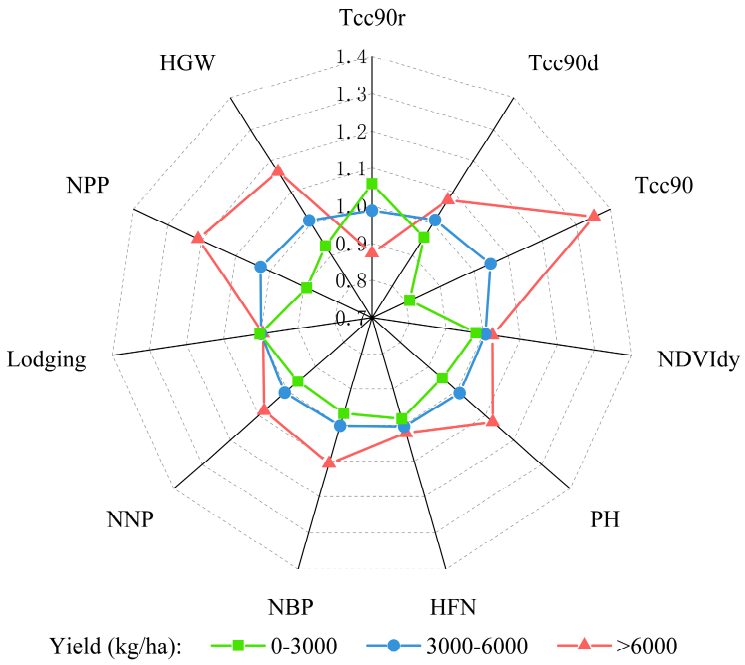

Figure 7. Comparison of the 11 examined phenotypic traits among high-yield (>6000 kg/ha), medium-yield (3000–6000 kg/ha), and low-yield (>6000 kg/ha) soybean varieties. To facilitate comparison and visualization, the traits of the three yield-level groups were normalized by dividing the mean value of each cluster by the mean value of all the soybean varieties.

To further determine the traits that are critical for the selection of high-yield varieties, PCA was used to analyze phenotypic correlations among three different yield groups (Figure 8). PC1 explained 28% of the variation and PC2 explained 19.7% of the variation. A group of highly correlated traits emerged, including canopy coverage-related traits (Tcc90r, Tcc90d, Tcc90) and NDVIdy. NDVIdy showed a strong positive correlation with Tcc90 and Tcc90d but a negative correlation with Tcc90r, which we attributed to the background effects of the soybean canopy and its developmental stage. PH exhibited positive correlations with NBP, NNP, NPP, and lodging. The main reason was that the increase in PH usually leads to increases in NBP [22] and NNP [21], ultimately leading to the increase of NPP [22]. As reported in a past study [34], the increase in PH would also increase the risk of lodging. HFN was negatively correlated with lodging, consistent with previous research [23], which found the fifth internode of soybean as a critical factor influencing lodging.

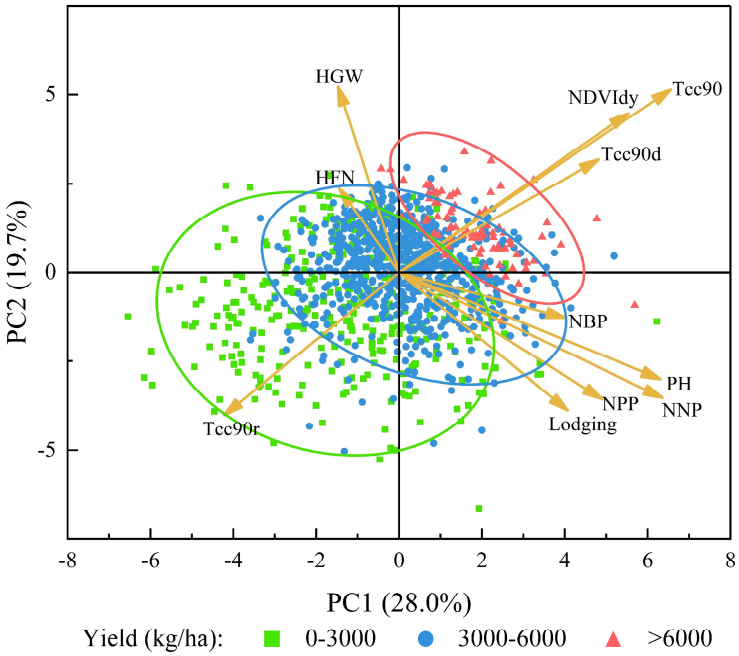

Figure 8. Biplot analysis for phenotypic similarity. The cosine of the angle between vectors represents the correlation between phenotypic traits. The highly correlated phenotypic traits pointed in roughly the same direction.

3.4. Utility of the Identified Key Traits in Screening High-Yield Soybean Varieties

Based on the differences and correlations of phenotypic traits, we proposed a screening method for high-yielding soybean varieties with canopy coverage-related traits (Tcc90r, Tcc90, Tcc90d), PH, and HGW. High-yield soybean varieties typically exhibited rapid canopy development, prolonged duration of high canopy coverage, a delayed decline in canopy coverage, and moderate-to-high plant height and hundred-grain weight. For threshold setting, we used the median as the criterion to define the ranges for Tcc90r, Tcc90d, PH, and HGW. Considering that excessively high PH can lead to severe lodging, the top 5% of values were excluded. Tcc90, which reflects traits related to soybean growth vigor and photosynthetic capacity, was set to the upper quartile to select varieties with strong growth potential.

Based on the proposed approach, 1923 soybean varieties from 2022 were tested with the following phenotypic parameters (Table 1). To validate the method's applicability, 2023 data with the same phenotypic parameters (Table 1) were used (Figure 9). Since the data only represents conditions without artificial support interference, and the parameters were not redefined as percentages, the performance parameter settings from 2022 were retained.

Among the 1923 varieties, 248 were selected, with 74 varieties having yields greater than 6000 kg/ha (top 5%), accounting for 29.8%. Only 23 varieties had yields below 3000 kg/ha (bottom 60%), making up 9.3% (Figure 10a). In total, there were 85 varieties with yields greater than 6000 kg/ha, of which 87% were successfully selected (Figure 10b). From 1015 varieties, 140 were selected, and 44 varieties had yields greater than 5000 kg/ha (top 5%), accounting for 31.4%. Among them, only 15 varieties had yields below 3000 kg/ha (bottom 60%), representing 10.7% (Figure 10c). Additionally, when analyzing the high-yield varieties, 61 varieties had yields greater than 5000 kg/ha, with 72% of them successfully selected (Figure 10d). These results clearly demonstrated that the method showed high accuracy and reliability in selecting high-yield soybean varieties.

Table 1. Phenotypic parameter settings.

	Tcc90r (DAS)	Tcc90d (DAS)	Tcc90 (day)	PH (cm)	HGW (g)
Range	<50	>80	>40	60–120	>15

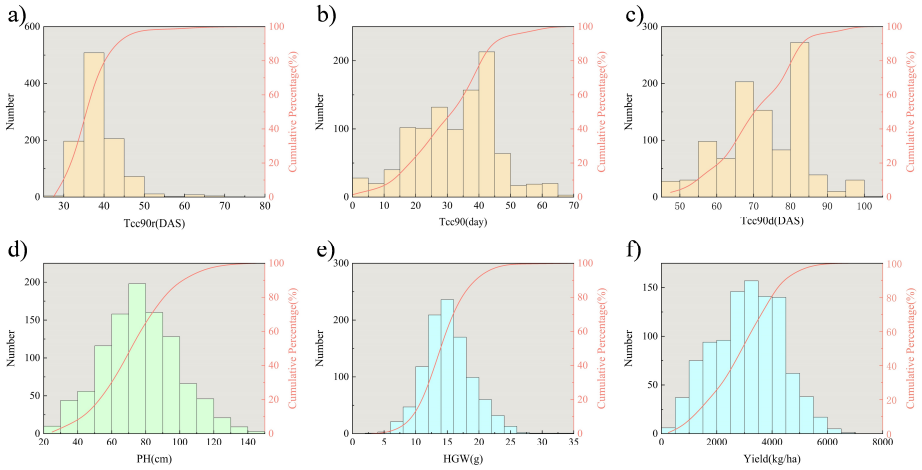

Figure 9. Phenotypic traits of soybean in 2023. (**a**–**c**) UAV-based phenotypic traits, including Tcc90r, Tcc90, and Tcc90d. (**d**–**f**) Field phenotypic traits, including PH, HGW, and Yield.

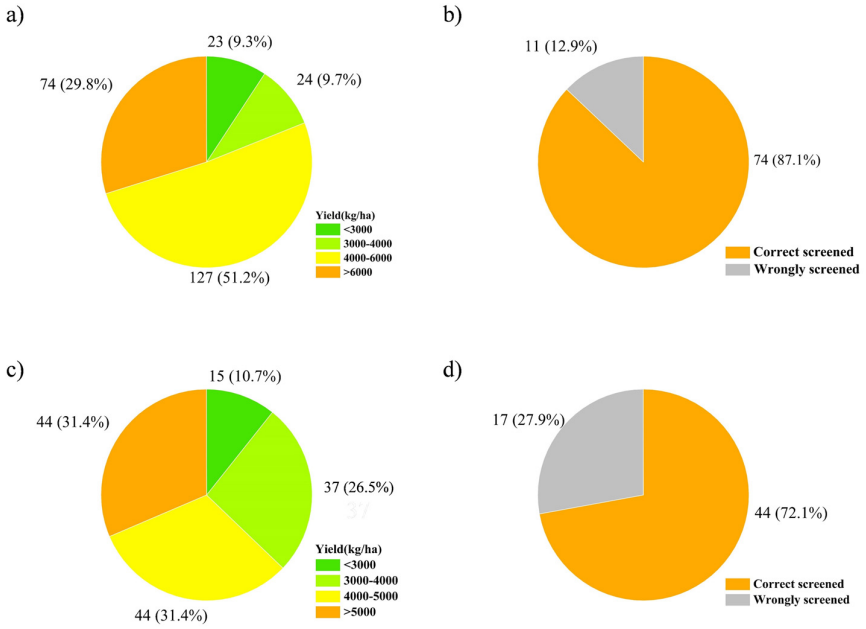

Figure 10. Screening results of soybean in 2022 and 2023. (**a**) The number of screened varieties with different yields in 2022. (**b**) The screening accuracy of varieties with yields greater than 6000 kg/ha (top 5%) in 2022. (**c**) The number of screened varieties with different yields in 2023. (**d**) The screening accuracy of varieties with yields greater than 5000 kg/ha (top 5%) in 2023.

4. Discussion

This study employed the SOM and K-means to conduct a two-step clustering analysis of numerous soybean varieties and effectively grouped the tested soybean varieties into five categories based on multidimensional phenotypic traits. The clustering results revealed significant differences in several key traits among different categories, such as canopy coverage indicators (Tcc90, Tcc90d), PH, and NPP. These traits in the high-yield category consistently outperformed those of the other categories, while the low-yield category exhibited deficiencies in most of the examined traits. These findings suggested that although

clustering analysis might not serve as a direct tool to finally select high-yield varieties, it could effectively favor identifying the key phenotypic traits related to high yield, providing a useful basis for further screening of high-yield varieties.

The differences in soybean yield were primarily reflected in NPP and HGW. High-yield soybean varieties typically combined a high NPP with a large HGW, both of which jointly determine the level of yield [17], as illustrated in Figure 7. In addition to yield components, canopy coverage was another critical factor influencing yield [20]. The high-yield group exhibited rapid canopy development, prolonged duration of high canopy coverage, and a delayed decline in canopy coverage (Figure 7). Rapid canopy establishment enhanced early light-use efficiency and suppressed weed competition [35]. Sustained high canopy coverage ensured efficient photosynthesis and carbon assimilation throughout the growing season, which was particularly critical for yield during the seed-filling period [19]. A delayed canopy decline extended the effective period of photosynthesis, contributing to increased grain dry matter and improved yield stability [36]. Furthermore, the high-yield soybean group also exhibited high PH, NBP, and NNP (Figure 7). These traits expanded the podding space, increased NPP, and ultimately boosted yield [18].

For identifying key traits in variety screening, this study categorized them into three groups: UAV-related traits (Tcc90r, Tcc90, Tcc90d, and NDVIdy), PH-related traits (PH, HFN, NBP, NNP, Lodging, and NPP), and HGW. UAV-related traits showed high correlations (Figure 8) but were relatively easy to obtain, with Tcc90r, Tcc90, and Tcc90d all being utilized for variety screening. NDVI has been widely demonstrated to be a good indicator of crop yield. However, the NDVIdy used in this study was not included in the screening process due to minimal differences between the three yield-level groups. PH-related traits were classified based on PCA results (Figure 8), and their performance aligned with findings from previous studies. Considering that these traits are manually acquired, only PH, which is relatively easy to obtain and has obvious differences between different groups, was selected for variety screening.

For high-yield variety selection, the method proposed in this study, which combined the UAV-based phenotypic traits (Tcc90r, Tcc90, and Tcc90d) with the field-based traits (PH and HGW), performed well in identifying high-yield varieties. Additionally, there are previous studies that have demonstrated the feasibility of extracting PH from UAV-based phenotyping [37,38]. Thus, in the future, PH could be integrated into UAV-based phenotypic traits, reducing the reliance on traditional field measurements and thereby making the screening of large-scale germplasm resources more efficient and easier.

The method proposed in this study had demonstrated its potential for screening of high-yield soybean varieties. However, its applicability may vary across different regions and crops. For example, environmental factors such as soil type, climate conditions, and management practices could influence the expression of key phenotypic traits, potentially reducing the method's robustness in regions with highly variable conditions. Additionally, while the UAV-based traits such as Tcc90r, Tcc90d, and Tcc90 were effective for testing soybean varieties in this study, their relevance and predictive power may differ for other crops with distinct canopy structures. Moreover, the frequent UAV data collection (e.g., every 3 days) may limit its applicability. However, reducing the frequency of data collection could compromise data accuracy. Future research should focus on adapting this method to other crops and testing its efficiency under diverse environmental conditions.

5. Conclusions

This study explored the relationships between multiple UAV- and field-based phenotypic traits and the yield of numerous soybean varieties using a two-step clustering analysis and identified the key traits that were valuable for rapid screening of high-yield varieties.

Firstly, we obtained UAV-based phenotypic traits (Tcc90r, Tcc90d, Tcc90, and NDVIdy) by high-throughput UAV phenotyping technology and field-based phenotypic traits (plant architecture traits and yield factors) by manual measurements. Then, a two-step clustering method using SOM and K-means was employed to group the test soybean varieties into different yield-level categories based on their phenotypic traits. It was found that the duration of canopy coverage remained above 90% (Tcc90) was a critical phenotypic trait for selecting high-yield varieties. Moreover, high-yield soybean varieties typically exhibited several key phenotypic traits such as rapid development of canopy coverage (Tcc90r, the time when canopy coverage first reached 90%), prolonged duration of high canopy coverage (Tcc90), a delayed decline in canopy coverage (Tcc90d, the time when canopy coverage began to decline below 90%), and moderate-to-high plant height (PH) and hundred-grain weight (HGW). Finally, the utility of the identified key traits in screening high-yield soybean varieties was demonstrated. In the 2022 and 2023 seasons, it successfully selected 87% and 72% of high-yield varieties (i.e., the top 5% of the test varieties), respectively. The proportion of low-yield varieties (i.e., the bottom 60% of the test varieties) involved mistakenly was only 9% and 10%, respectively. Overall, this study provided useful insights into the key traits that were valuable for identifying the high-yield soybean varieties from numerous testing materials and demonstrated the benefit of integrating the UAV-based high-throughput phenotypic traits into soybean high-yield variety screening. In the future, it needs more data from multi-environmental breeding trials to further test the efficiency and stability of high-yield soybean variety selection.

Author Contributions: Conceptualization, H.L. and G.Y.; methodology, C.Y.; software, C.Y.; validation, P.R., D.P. and H.F.; formal analysis, C.Y.; investigation, H.F.; resources, D.P., P.R., H.W., S.L. and J.Z.; data curation, D.P. and P.R.; writing—original draft preparation, C.Y.; writing—review and editing, H.L. and G.Y.; visualization, C.Y.; supervision, G.Y.; project administration, H.F.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2021YFD1201601) and the National Natural Science Foundation of China (41671411).

Data Availability Statement: Data are contained within the article.

Acknowledgments: We thank Lijuan Qiu, Yinghui Li from the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, and all the kind people for their assistance with the field data collection.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liang, Q.; Chen, L.; Yang, X.; Yang, H.; Liu, S.; Kou, K.; Fan, L.; Zhang, Z.; Duan, Z.; Yuan, Y.; et al. Natural variation of Dt2 determines branching in soybean. *Nat. Commun.* **2022**, *13*, 6429. [CrossRef] [PubMed]
2. Skobalski, J.; Sagan, V.; Alifu, H.; Akkad, O.A.; Lopes, F.A.; Grignola, F. Bridging the gap between crop breeding and GeoAI: Soybean yield prediction from multispectral UAV images with transfer learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *210*, 260–281. [CrossRef]
3. Han, L.; Yang, G.; Dai, H.; Yang, H.; Xu, B.; Li, H.; Long, H.; Li, Z.; Yang, X.; Zhao, C. Combining self-organizing maps and biplot analysis to preselect maize phenotypic components based on UAV high-throughput phenotyping platform. *Plant Methods* **2019**, *15*, 57. [CrossRef]
4. Ampatzidis, Y.; Partel, V. UAV-Based High Throughput Phenotyping in Citrus Utilizing Multispectral Imaging and Artificial Intelligence. *Remote Sens.* **2019**, *11*, 410. [CrossRef]
5. Roth, L.; Barendregt, C.; Bétrix, C.-A.; Hund, A.; Walter, A. High-throughput field phenotyping of soybean: Spotting an ideotype. *Remote Sens. Environ.* **2022**, *269*, 112797. [CrossRef]
6. Farshad, B.; Yildirim, K.; Irish, L.B.P.; Glen, L.R.; Hamed, S.-S. CottonSense: A high-throughput field phenotyping system for cotton fruit segmentation and enumeration on edge devices. *Comput. Electron. Agric.* **2024**, *216*, 108531.

7. Xu, R.; Li, C. A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots. *Plant Phenomics* **2022**, *2022*, 9760269. [CrossRef] [PubMed]
8. Zhao, D.; Xu, T.; Henke, M.; Yang, H.; Zhang, C.; Cheng, J.; Yang, G. A method to rapidly construct 3D canopy scenes for maize and their spectral response evaluation. *Comput. Electron. Agric.* **2024**, *224*, 109138. [CrossRef]
9. Blekanov, I.; Molin, A.; Zhang, D.; Mitrofanov, E.; Mitrofanova, O.; Li, Y. Monitoring of grain crops nitrogen status from UAV multispectral images coupled with deep learning approaches. *Comput. Electron. Agric.* **2023**, *212*, 108047. [CrossRef]
10. Bhandari, M.; Chang, A.; Jung, J.; Ibrahim, A.M.H.; Rudd, J.C.; Baker, S.; Landivar, J.; Liu, S.; Landivar, J. Unmanned aerial system-based high-throughput phenotyping for plant breeding. *Plant Phenome J.* **2023**, *6*, e20058. [CrossRef]
11. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Hartling, S.; Esposito, F.; Fritschi, F.B. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* **2020**, *237*, 111599. [CrossRef]
12. Ren, P.; Li, H.; Han, S.; Chen, R.; Yang, G.; Yang, H.; Feng, H.; Zhao, C. Estimation of Soybean Yield by Combining Maturity Group Information and Unmanned Aerial Vehicle Multi-Sensor Data Using Machine Learning. *Remote Sens.* **2023**, *15*, 4286. [CrossRef]
13. Shu, M.; Li, Q.; Ghafoor, A.; Zhu, J.; Li, B.; Ma, Y. Using the plant height and canopy coverage to estimation maize aboveground biomass with UAV digital images. *Eur. J. Agron.* **2023**, *151*, 126957. [CrossRef]
14. Borra-Serrano, I.; Swaef, T.D.; Quataert, P.; Aper, J.; Saleem, A.; Saeys, W.; Somers, B.; Roldán-Ruiz, I.; Lootens, P. Closing the Phenotyping Gap: High Resolution UAV Time Series for Soybean Growth Analysis Provides Objective Data from Field Trials. *Remote Sens.* **2020**, *12*, 1644. [CrossRef]
15. Sarkar, S.; Zhou, J.; Scaboo, A.; Zhou, J.; Aloysius, N.; Lim, T.T. Assessment of Soybean Lodging Using UAV Imagery and Machine Learning. *Plants* **2023**, *12*, 2893. [CrossRef] [PubMed]
16. Wang, C.; Xu, S.; Yang, C.; You, Y.; Zhang, J.; Kuai, J.; Xie, J.; Zuo, Q.; Yan, M.; Du, H.; et al. Determining rapeseed lodging angles and types for lodging phenotyping using morphological traits derived from UAV images. *Eur. J. Agron.* **2024**, *155*, 127104. [CrossRef]
17. Roekel, R.J.V.; Purcell, L.C.; Salmerón, M. Physiological and management factors contributing to soybean potential yield. *Field Crops Res.* **2015**, *182*, 86–97. [CrossRef]
18. Todeschini, M.H.; Milioli, A.S.; Rosa, A.C.; Dallacorte, L.V.; Panho, M.C.; Marchese, J.A.; Benin, G. Soybean genetic progress in South Brazil: Physiological, phenological and agronomic traits. *Euphytica* **2019**, *215*, 124. [CrossRef]
19. He, S.; Li, X.; Chen, M.; Xu, X.; Zhang, W.; Chi, H.; Shao, P.; Tang, F.; Gong, T.; Guo, M.; et al. Excellent Canopy Structure in Soybeans Can Improve Their Photosynthetic Performance and Increase Yield. *Agriculture* **2024**, *14*, 1783. [CrossRef]
20. Renato, C.C.; César, S.a.G.; Rotter, M.A.; Alexandre, G.; Crescêncio, S.C.P.; Maicon, N.; Aluizio, B. Association between unmanned aerial vehicle high-throughput canopy phenotyping and soybean yield. *Agron. J.* **2022**, *114*, 1581–1598.
21. Chen, L.; Nan, H.; Kong, L.; Yue, L.; Yang, H.; Zhao, Q.; Fang, C.; Li, H.; Cheng, Q.; Lu, S.; et al. Soybean AP1 homologs control flowering time and plant height. *J. Integr. Plant Biol.* **2020**, *62*, 1868–1879. [CrossRef] [PubMed]
22. Gao, C.; Yuan, J.; Lu, J.; Ye, W.; Zhi, J.; Li, Y.; Li, W.; Sun, M.; Kong, F.; Liu, B.; et al. COL3a simultaneously regulates flowering and branching to improve grain yield in soybean. *Plant Biotechnol. J.* **2024**, *23*, 201–203. [CrossRef] [PubMed]
23. Shan, F.; Sun, K.; Gong, S.; Wang, C.; Ma, C.; Zhang, R.; Yan, C. Effects of Shading on the Internode Critical for Soybean (Glycine Max) Lodging. *Agronomy* **2022**, *12*, 492. [CrossRef]
24. de Freitas Souza, M.; de Paiva Filho, S.V.; Rosa, V.C.S.; Braz, G.B.P.; Menezes, J.F.S.; Silva, D.V.; Reginaldo, L.T.R.T.; Ferreira, C.J.B.; da Silva Solino, A.J. Morphological and phenotypical traits and their relationship with soybean grain yield: A multivariate analysis approach. *Euphytica* **2024**, *220*, 26. [CrossRef]
25. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*, 52–65. [CrossRef] [PubMed]
26. Almeida, L.F.; Correndo, A.A.; Hefley, T.; Hintz, G.; Prasad, P.V.V.; Licht, M.; Casteel, S.; Singh, M.; Naeve, S.; Bais, J.; et al. Assessing the influence of environmental drivers on soybean seed yield and nitrogen fixation estimates and uncertainties in the United States. *Eur. J. Agron.* **2025**, *162*, 127428. [CrossRef]
27. Blancas, V.H.S.; Hernández, H.J.; Navarro, A.M.H.; Alvarado, J.M.Á.; Esparza, D.M.C.; Reséndiz, J.R. A Clustering and PL/SQL-Based Method for Assessing MLP-Kmeans Modeling. *Computers* **2024**, *13*, 149. [CrossRef]
28. Lin, C.; Du, R.; Guo, F. Implication of self-organizing map, stable isotopes combined with MixSIAR model for accurate nitrogen control in a well-protected reservoir. *Environ. Res.* **2024**, *248*, 118335. [CrossRef]
29. Zhang, D.; Mansaray, L.R.; Jin, H.; Sun, H.; Kuang, Z.; Huang, J. A universal estimation model of fractional vegetation cover for different crops based on time series digital photographs. *Comput. Electron. Agric.* **2018**, *151*, 93–103. [CrossRef]
30. Stepanov, A.; Dubrovin, K.; Sorokin, A.; Aseeva, T. Predicting Soybean Yield at the Regional Scale Using Remote Sensing and Climatic Data. *Remote Sens.* **2020**, *12*, 1936. [CrossRef]
31. Xie, J.; Zhou, Z.; Zhang, H.; Zhang, L.; Li, M. Combining Canopy Coverage and Plant Height from UAV-Based RGB Images to Estimate Spraying Volume on Potato. *Sustainability* **2022**, *14*, 6473. [CrossRef]
32. Stichter, E.C. Assessing the potential for soybean yield improvement through plant architectural modification. *Nat. Sci. Educ.* **2020**, *49*, e20002. [CrossRef]

33. Mauro, G.D.; Rotundo, J.L. Lodging dynamics and seed yield for two soybean genotypes with contrasting lodging-susceptibility. *Eur. J. Agron.* **2025**, *163*, 127445. [CrossRef]
34. Tobias, W.; Simon, M.L.; Longin, C.F.H. Genetic control of plant height in European winter wheat cultivars. *Theor. Appl. Genet.* **2015**, *128*, 865–874.
35. Jannink, J.L.; Jordan, N.R.; Orf, J.H. Index Selection for Weed Suppressive Ability in Soybean. *Crop Sci.* **2000**, *40*, 1087–1094. [CrossRef]
36. Wu, Z.; Xu, C.; Li, R.; Xu, Y.; Hua, J.; Sun, S.; Han, T.; Song, W.; Wu, C. Full-field straw mulching and fertilizer application improved the soybean seed yield through optimization of the root and canopy structure: A study case in Huang-Huai-Hai region. *Eur. J. Agron.* **2024**, *159*, 127280. [CrossRef]
37. Liu, T.; Zhu, S.; Yang, T.; Zhang, W.; Xu, Y.; Zhou, K.; Wu, W.; Zhao, Y.; Yao, Z.; Yang, G.; et al. Maize height estimation using combined unmanned aerial vehicle oblique photography and LIDAR canopy dynamic characteristics. *Comput. Electron. Agric.* **2024**, *218*, 108685. [CrossRef]
38. Hu, P.; Chapman, S.C.; Wang, X.; Potgieter, A.; Duan, T.; Jordan, D.; Guo, Y.; Zheng, B. Estimation of plant height using a high throughput phenotyping platform based on unmanned aerial vehicle and self-calibration: Example for sorghum breeding. *Eur. J. Agron.* **2018**, *95*, 24–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.