Vatter Thomas (Orcid ID: 0000-0001-7344-6351)
Gracia-Romero Adrian (Orcid ID: 0000-0001-8308-9693)

Preharvest phenotypic prediction of grain quality and yield of durum wheat using multispectral imaging

Thomas Vatter[1,2], Adrian Gracia-Romero[1,2], Shawn Carlisle Kefauver[1,2], María Teresa Nieto-Taladriz[3], Nieves Aparicio[4], José Luis Araus[1,2*]

1 Integrative Crop Ecophysiology Group, Plant Physiology Section, Faculty of Biology, University of Barcelona, Diagonal 643, 08028 Barcelona, Spain
2 AGROTECNIO (Center of Research in Agrotechnology), Av. Rovira Roure 191, 25198 Lleida, Spain
3 INIA-CSIC (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria), Ctra. de la Coruña Km. 7.5, 28040 Madrid, Spain
4 Technological and Agrarian Institute of Castilla y León (ITACyL), Agricultural Research, Ctra Burgos km 119, 47041 Valladolid, Spain

Running head

Asses grain quality and yield using spectral bands

Keywords: Multispectral imaging, quality trait prediction, yield prediction, protein content, test weight, vitreousness, grain yield, durum wheat, machine learning, unmanned aerial vehicle

This article has been accepted for publication and 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: 10.1111/tpj.15648

Corresponding author

Thomas Vatter

E-mail : tvatter@ub.edu

Address: Integrative Crop Ecophysiology Group, Plant Physiology Section, Faculty of Biology, University of Barcelona, Diagonal 643, 08028 Barcelona, Spain

Abstract

Durum wheat is an important cereal that is widely grown in the Mediterranean basin. In addition to high yield, grain quality traits are of high importance for farmers. The strong influence of climatic conditions makes the improvement of grain quality traits, like protein content, vitreousness, and test weight a challenging task. Evaluation of quality traits post-harvest is time and labor intensive and requires expensive equipment, such as near-infrared spectroscopy or hyper spectral imagers. Predicting not only yield but also important quality traits in the field before harvest is of high value for breeders aiming to optimize resource allocation. Implementation of efficient approaches for trait prediction, such as the use of high-resolution spectral data acquired by a multispectral camera mounted on unmanned aerial vehicles (UAVs), need to be explored. In this study, we have acquired multispectral image data with an 11-band multispectral camera mounted on a UAV and analyzed the data with machine learning (ML) models to predict grain yield and important quality traits in breeding micro-plots. Combining 11-band multispectral data for 34 cultivars and 16 environments allowed to develop ML models with good prediction capability. Applying the trained models to test sets explained a considerable degree of phenotypic variance with good accuracy showing r-squared values of 0.84, 0.69, 0.64, and 0.61 and normalized root mean squared errors of 0.17, 0.07, 0.14, and 0.03 for grain yield, protein content, vitreousness, and test weight, respectively.

Introduction

Durum wheat (*Triticum turgidum* L. subsp. *durum* (Desf) Husn.), a tetraploid cereal, is widely grown in the Mediterranean basin (Lidon et al. 2014). Although mainly associated with pasta

production, a high percentage of harvested grain is used for couscous and bulgur production as well (Beres et al. 2020). Increasing grain yield (GY) is the major focus of most breeding programs, while improving grain quality parameters is often a secondary objective (Michel et al. 2019). Although GY remains the major determinant for the economic benefit of producers, the grain quality requirements of the processing industry need to be met to obtain the optimal price (Longin et al. 2013; Zhang et al. 2021). Protein content (PC) is arguably the most important quality trait, as high PC positively influences baking and pasta quality (Nobile et al. 2005; Samaan et al. 2006; Xue et al. 2019). Test weight (TW), reflecting the bulk density of grain, and grain vitreousness (V), a visual marker for grain hardness, are two other important grain quality traits. Although studies have shown contradictory results, TW is still commonly used as an indicator of milling potential (Wang and Fu 2020). A high percentage of vitreous kernels will generally give high semolina yield (Dexter et al. 1988). Breeding for improved quality parameters is a major challenge because most quality parameters are greatly influenced by climatic parameters and management practices (Rharrabti et al. 2003). Heat, nitrogen availability and water stress have a major effect on durum wheat yield and grain quality, especially when occurring post flowering (Ben Mariem et al. 2020; Flagella et al. 2010). While heat and drought stress reduce GY, grain quality traits such as PC, TW, V, and flower yellowness have been observed to improve (Gagliardi et al. 2020; Li et al. 2013; Magallanes-López et al. 2017; Rharrabti et al. 2003). Increased nitrogen application and especially late-season nitrogen application have been observed to increase next to GY the grain quality parameters PC and TW (Blandino et al. 2015; Suprayogi et al. 2009). Therefore, the evaluation of genotypes in a wide range of environments is a prerequisite.

Use of near infrared spectroscopy (NIRS) at harvest or post-harvest has shown to be an accurate, rapid, and nondestructive technique to measure many quality parameters, including PC, T, and V (Dowell et al. 2006; Gorretta et al. 2006). In addition to this, reflectance, transmittance, and hyper spectral image analysis have been successfully applied to assess cereal grain quality post-harvest (Caporaso et al. 2018; Symons et al. 2003; Wiegmann et al. 2019; Xie et al. 2004). Prediction of cereal grain quality parameters on the basis of remote sensing data acquired in situ before harvest is much less studied. Hansen et al. (2002) examined canopy data derived with a hand-held spectroradiometer using eight medium broad bands to predict PC in one wheat and one barley cultivar. No relationship between the reflectance measures and PC was obtained in the case of barley, while for wheat a relationship was observed, although with low prediction accuracy (Hansen et al. 2002). This study reflects the inherent difficulty of predicting grain quality traits using few wavebands. Söderström et al. (2010) used a 45-band spectral sensor in combination with satellite imagery to predict protein content in barley, achieving reasonable PC prediction on a field level, although it was considerably less accurate when applied across multiple years and locations. Zhao et al. (2019), Tan et al. (2020), and Xu et al. (2020) predicted the PC of winter wheat using satellite data, with Zhao et al. (2019) also using hyperspectral data derived from an unmanned aerial vehicle (UAV). In all of these studies, PC was predicted with good to high accuracy. Furthermore, prediction models for PC based on the reflectance spectrum of wheat canopies taken at the ground level using a full-range portable high resolution spectroradiometer showed good prediction capability (Vergara Diaz et al. 2020a). However, hyperspectral cameras and high resolution spectroradiometers are expensive devices and data processing and further analysis is laborious. In a recent study by Zhou et al. (2021), a four-band spectral camera mounted on a UAV was used to predict GY and PC in

winter wheat. Models based on collected spectral data allowed an intermediate amount of phenotypic variance in PC to be explained, with the authors highlighting the need for further studies in a wide variety of environments. Moreover, the aim of this study was not crop phenotyping but instead to predict within-field variability in GY and PC using a single cultivar. Similar to this, Kang et al. (2021) showed that data obtained with a five-band spectral camera mounted on a UAV in combination with machine learning (ML) allowed prediction of PC in a rice cultivar with low error. As emphasized by Zhou et al. (2021), it is essential to verify obtained results in such studies by considering a wide range of environments. The goal of the current study is to evaluate the power of multispectral data for grain quality prediction based on a wider range of environments and cultivars. Moreover, to the best of our knowledge, no studies evaluating the use of spectral data obtained in the field for predicting TW and V have been published so far. Developing prediction models for these important grain quality parameters will be of high value for durum wheat breeders.

While data derived using hyperspectral sensors, either ground based or mounted on UAVs, allows for good prediction of PC, these sensors are costly, and are unaffordable in many cases. In contrast, most satellite data is freely available or of low cost, and despite considerably fewer wavelengths, allows good PC prediction when combined with ML models. However, spectral data obtained using satellites is greatly limited in resolution. Remote sensing using satellite-based imagery is only applicable in situations where larger sized plots are being assessed (Tattaris et al. 2016), therefore impeding its application in conventional breeding trials.

Considering these facts, the acquisition of spectral data via spectral cameras with limited waveband numbers and mounted on a UAV can be regarded as a powerful alternative, allowing the rapid accumulation of high-resolution spectral data at reasonable cost. If

combined with advanced ML methods like neural networks, this approach promises to be of high value for breeders aiming to predict quality parameters at the micro-plot level. In addition, our study may give clues for further implementation of phenotyping using future generations of satellites with higher resolutions. Alternatively, it may be possible to forecast grain quality traits at the field level using currently operational satellites, such as the Copernicus Sentinel-2 satellites, which have 13 spectral bands and a 10 m spatial resolution.

Therefore, the objective of this study was to evaluate the suitability of spectral data, derived using an 11-band multispectral camera mounted on a UAV, combined with machine learning to predict GY, PC, TW, and V in durum wheat in the anthesis stage. For this we made use of data based on a wide range of environments and cultivars, unprecedented in this extent in field-based quality trait prediction.

Results

Phenotypic data

A wide variation was observed across environments for all traits considered in the study. GY showed a range of 11.92 Mg ha^{-1}, PC of 11%, V of 87.5%, and TW of 19.95 kg hl^{-1}, respectively (Table 1). Training sets (i.e. used for model training) and test sets (i.e. used for assessing the prediction capability of the model) showed a similar distribution of data for all traits, with the distribution being highly left skewed in the case of the quality trait V (Supplementary Figure 1).

Table 1. Descriptive statistics for training and test set and heritability.

Trait	$Mean_{Train}$	$Mean_{Test}$	Min_{Train}	Min_{Test}	Max_{Train}	Max_{Test}	CV_{Train}[a]	CV_{Test}[a]	h^2[b]

Grain yield (Mg ha^{-1})	5.14	5.45	0.33	0.57	12.25	11.28	0.44	0.43	0.72
Protein content (%)	15.07	14.88	9.70	10	20.70	19	0.14	0.12	0.94
Vitreousness (%)	84.83	84.42	14.50	12.50	100	100	0.22	0.24	0.80
Test weight (kg hl^{-1})	79.84	78.92	70.42	67.35	87.30	86.30	0.04	0.04	0.96

[a]Coefficient of variation.
[b]Broad-sense heritability.

Broad-sense heritability (h^2) calculated across environments was high for all traits, with the highest h^2 being observed for TW (h^2 = 0.96) and lowest for GY (h^2 = 0.72) (Table 1).

Prediction of grain yield and quality traits

Overall, no strong differences in the prediction statistics for training and test set data were observed (Table 2). Regarding GY, application of the averaging neural network (avNNet) model to the test set enabled to explain a high percentage of phenotypic variance (R^2 = 0.84). The accuracy of the GY prediction was intermediate (normalized root-mean-square error; nRMSE = 0.17) (Table 2; Fig. 1a). In terms of the quality trait, PC, the phenotypic variance explained by the model (R^2) in the validation set was calculated as 69%. The accuracy of PC prediction was shown to be high (nRMSE = 0.07) (Table 2; Fig. 1b). For the quality trait V, the phenotypic variance explained by the avNNet model was calculated as 64%, with an intermediate prediction accuracy (nRMSE = 0.14) (Table 2; Fig. 1c). For the quality trait TW, the phenotypic variance explained was calculated at 61%, and very high accuracy of prediction was observed for TW (nRMSE = 0.03) (Table 2; Fig. 1d). Slopes observed for the regression lines based on the test data were 0.90, 0.70, 0.58, and 0.58 for the GY, PC, V, and TW traits, respectively. All slopes differed significantly ($p < 0.001$) from the 1:1 reference line.

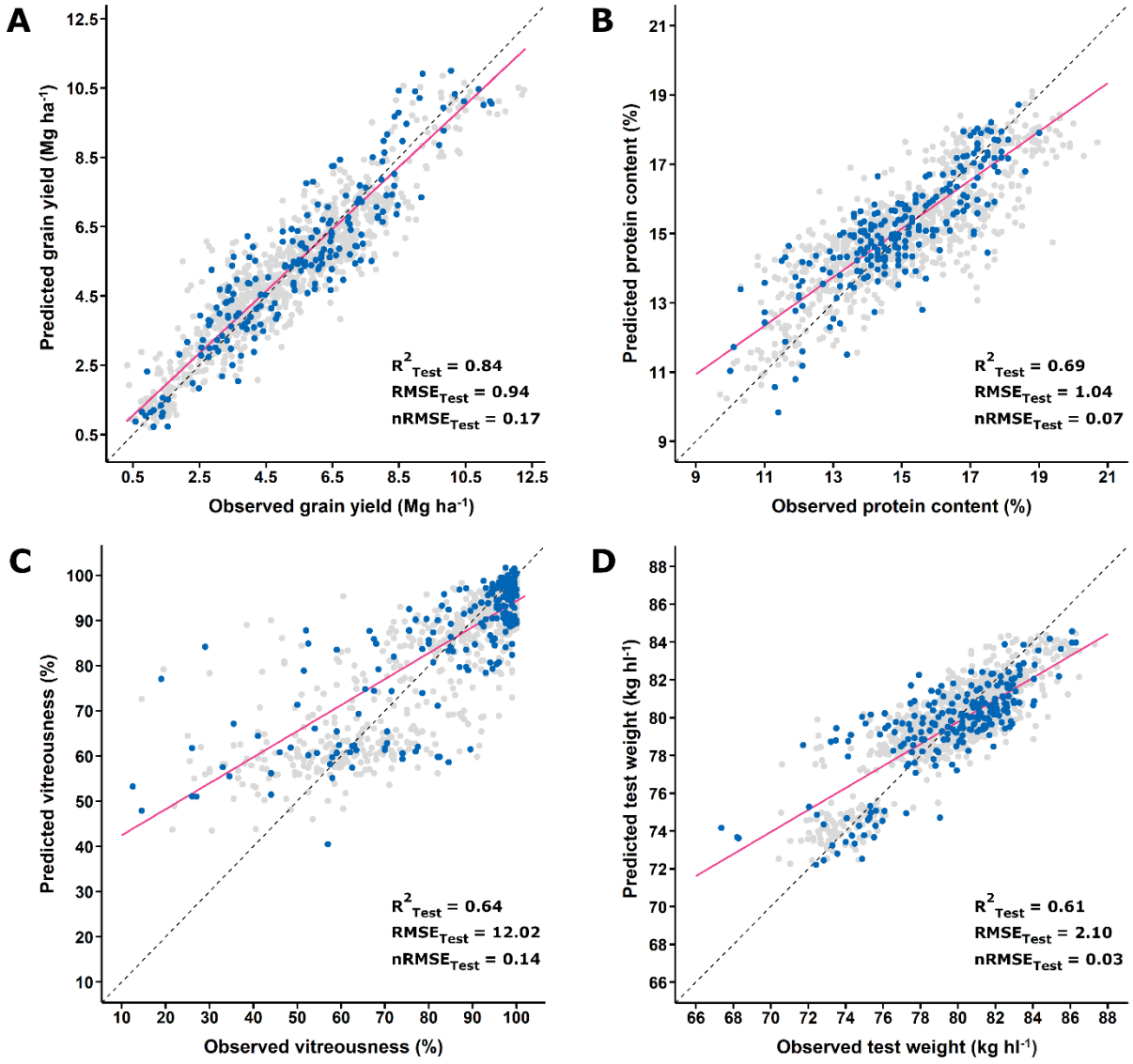


Figure 1: Scatterplots for observed and predicted data. (A) grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Blue dots depict data obtained by applying the trained averaging neural network (avNNet) model to the test set, while grey dots depict data observed in the training set. The red line shows a linear regression line based on the test set data, with a significance of $p < 0.001$ for all traits. As a reference, the black dashed line indicates a 1:1 relationship. The prediction statistics depicted in the plots refer to the test set.

Table 2. Prediction statistics for grain yield and quality traits for the training and test sets.

Trait	R^2_{Train}[a]	R^2_{Test}[a]	$RMSE_{Train}$[b]	$RMSE_{Test}$[b]	$nRMSE_{Train}$[c]	$nRMSE_{Test}$[c]
Grain yield (Mg ha^{-1})	0.85	0.84	0.88	0.94	0.17	0.17
Protein content (%)	0.67	0.69	1.21	1.04	0.08	0.07
Vitreousness (%)	0.72	0.64	9.86	12.02	0.12	0.14
Test weight (kg hl^{-1})	0.73	0.61	1.67	2.10	0.02	0.03

[a]Squared Pearson correlation coefficient.
[b]Root mean square error.
[c]Normalized root mean square error.

Residual plots of the test set predictions showed a balance in the avNNet model to over- or under-predict in the case of GY and PC in most ranges, while for V and TW over-prediction by the model was slightly more pronounced than under-prediction (Fig. 2).

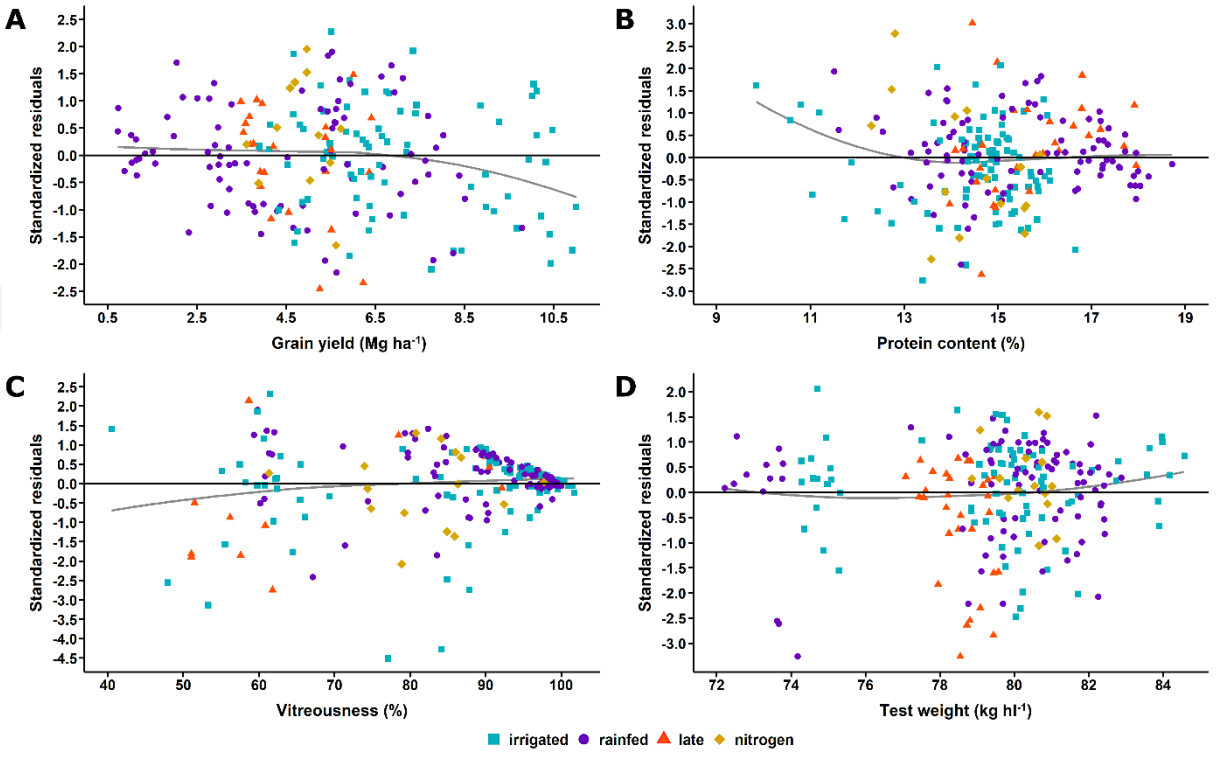


Figure 2: Standardized residual plots of averaging neural network (avNNet) models applied to the validation sets. The X-axis shows the predicted data for, **(A)** grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Colors refer to the four main treatments,

irrigated, rainfed, late, and nitrogen. The grey line shows the locally estimated scatterplot smoothing (LOESS) line across all main treatments.

The most extreme standardized residuals (< -4) were observed for the quality trait V, reflecting strong over-prediction of the observed low V percentages (Fig. 1c; Fig. 2c). No definite pattern was observed between standardized residuals and growth condition in any of the traits evaluated. However, a tendency towards over-prediction was observed for data originating from the late planting condition related to the quality trait TW (Fig. 2d). Notably, the residual plot of the quality trait TW indicated a separation into two clusters (Fig. 2d). The smaller cluster in the lower range of predicted TW values was linked to data acquired at a single timepoint in 2017 at the experimental station in Aranjuez. The three most extreme standardized residuals (< -2.5) in this cluster are linked to the cultivar Pedroso under rainfed conditions. Assigning each prediction in the test set to its respective cultivar and comparing it to the observed data for the respective cultivar only showed significant differences ($p < 0.01$; Tukey test) between TW data (Fig. 3). In particular, significant differences between observed and predicted cultivar TW data were only observed for the two cultivars Pedroso ($p = 5.9e\text{-}6$) and D Norman ($p = 0.027$) (Fig. 3d).

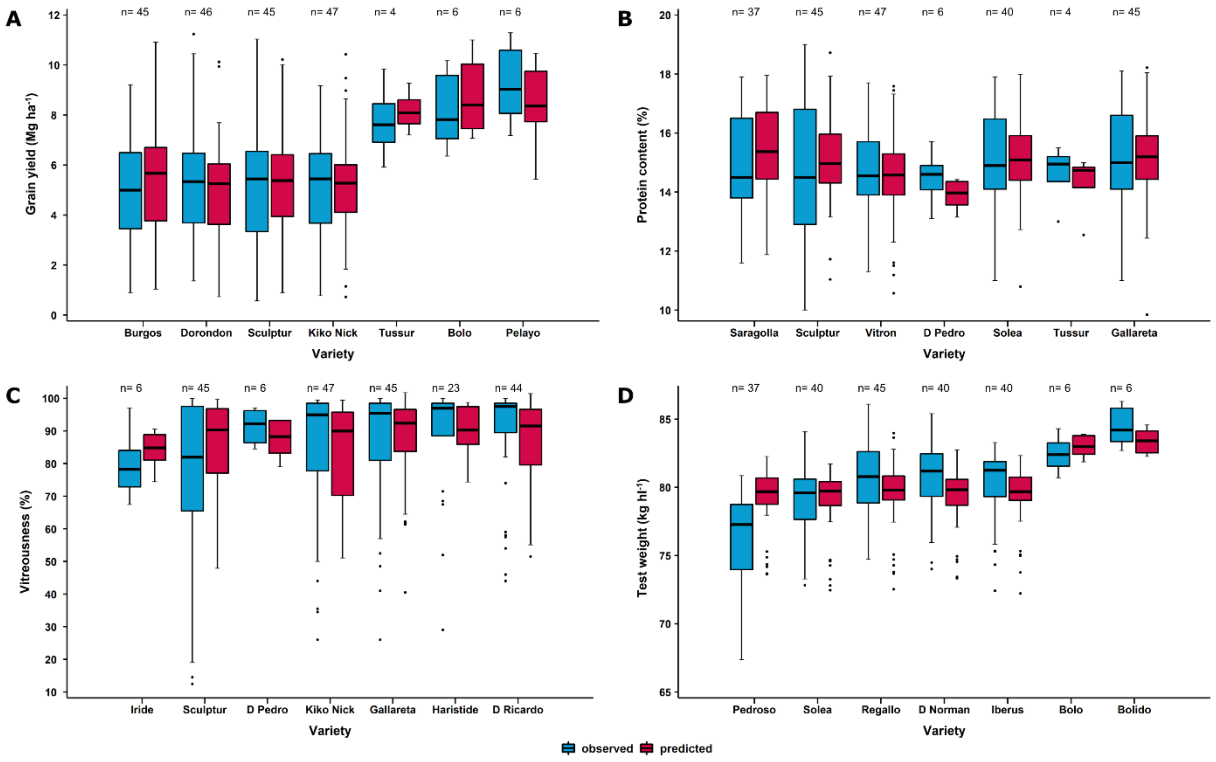


Figure 3: Boxplots for observed and predicted data of the validation sets merged by cultivar. (A) grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Blue shading indicates observed data, while red shading indicates predicted data obtained by applying the trained averaging neural network (avNNet) model to the test set. The number of data points in the validation set for each cultivar is indicated by the letter n. Cultivars comprising the test set vary because for each trait, seven cultivars were randomly sampled from the full data set.

Discussion

Until the current work, evaluations of the suitability of multispectral data to predict grain quality parameters prior to harvest had only been undertaken in studies that were greatly limited in the number of environments or cultivars considered. In addition, to our best knowledge, evaluation of spectral data to predict the key grain quality traits TW and V has not been pursued previously. Testing a high number of cultivars across a diverse set of

environments has enabled creation of a data set possessing high degrees of variation for all of the traits evaluated. As such, it provides the ideal basis for evaluating the potential of spectral data for in-field quality trait prediction.

High h^2 estimates for the investigated traits are of major importance when assessing the suitability of trait prediction models for their application in breeding programs. It is only when a sufficiently high h^2 is observed for the investigated quality traits that the selection of genotypes showing superior grain quality parameters will result in genetic gain, and thus contribute to developing improved cultivars in the long term. Therefore, for plant breeding, the development of prediction models is only worthwhile if the target trait to be predicted shows sufficient heritability. The h^2 estimates observed in this study have been shown to be high compared to those commonly cited in the literature, especially in the case of PC. In contrast, most studies report the h^2 of the PC in wheat as being in the range of 0.2 to 0.7, depending on the evaluated genotypes and environments (Giancaspro et al. 2019; Kramer 1979; Mahjourimajd et al. 2016; Suprayogi et al. 2009). Nevertheless, similar to this study, Thorwarth et al. (2018) reported an h^2 of PC in wheat of 0.91. The high h^2 estimates observed in our study can be explained by the high number of environments considered, resulting in a strongly reduced masking variance. Given the high h^2 estimates observed for all target traits, the prediction models developed in the framework of this study are of high value.

Prediction of yield during plant growth by use of a multi- or hyperspectral camera mounted on a UAV has been successfully applied in a range of major crops like wheat, maize, rice, rapeseed, potato, and soybean (Duan et al. 2019; Fu et al. 2020; Gong et al. 2018; Hassan et al. 2019; Li et al. 2020; Maimaitijiang et al. 2020; Maresma et al. 2016; Zhou et al. 2021). While predicting GY in durum wheat via a UAV-mounted multispectral camera is not a new

approach (Romero et al. 2019), previous studies have been limited in the number of genotypes or environments considered. With 32 cultivars, Hassan et al. (2019) evaluated the largest genotype set to date, but trials were restricted to two different treatments. By contrast, Li et al. (2020) evaluated 17 different treatments but focused on only six genotypes. Thus, the present study involving 34 cultivars and 16 environments represents, to our best knowledge, the most extensive evaluation of yield prediction so far using spectral data derived using a UAV.

While the focus of our study was mainly on prediction of quality traits, and GY was included in the study more as a well-characterized reference trait, the results obtained for GY are of high value for the breeding community. The prediction model for GY presented here was trained on the most diverse data set seen so far in the literature. Moreover, our work is also distinct in the way that spectral information has been used. Common to all of the research on yield prediction mentioned above is the application of vegetation indices (VIs) calculated with spectral band information, and VIs have become a standard approach when working with spectral information (Xue and Su 2017). Galvão et al. (2013) observed VIs to be less sensitive to changes in illumination and viewing geometry, which might contribute to their wider application. Furthermore, VIs representing a combination of two or more wavebands reduce the noise related to overall albedo variance that is inherent when using single wavebands (Ji et al. 2014; Liu and Huete 1995; Zhu et al. 2014). However, it has to be noted that the combination of single spectral bands to specific VIs comes at the cost of artificially narrowing down the information accessible by ML models. Therefore, in this study we did not combine the spectral bands into VIs but provided the single band information as input for the avNNet model. Thus, we followed a fully data driven approach, which gave the model full flexibility in its use of the bands for trait prediction. Although it is difficult to compare

studies of different sample size and with measurements performed at differing growth stages, the results of this study indicated that good GY prediction does not depend on the use of VIs (Table 2, Fig. 1). Furthermore, the results of this study highlight that good prediction can be achieved even with a fairly low number of spectral bands provided to the prediction model. To date, this approach of using individual wavebands to develop empirical (i.e. statistically based) prediction models has proven successful when using hundreds of wavebands acquired, for example, using high-resolution spectroradiometers (Vergara-Diaz et al. 2020a; Vergara Diaz et al. 2020b). However, focus should also be directed toward the selection of the ML method so that the one with the best prediction capability is determined for the specific trait and data set. Considering the different ML methods applied during the yield prediction studies, it is clear that there is no unique ML method that always outperforms the rest. In the present work an avNNet model was selected on the basis of its RMSE because it outperformed other commonly applied ML methods (Supplementary Table 1).

Despite the strong influence of environmental factors on grain quality traits, the ML models developed in this study explained a large amount of phenotypic variance at high accuracy in the PC, V, and TW (Table 2; Fig. 1). We thus demonstrated the suitability of spectral data derived using a limited band multispectral camera mounted on a UAV for the prediction of the V and TW quality traits, which previously have not been evaluated. The pronounced over-prediction of low V values (Fig. 1c, Fig. 2c) was most likely caused by the majority of observations being concentrated in the higher range of V (Supplementary Figure 1). Integrating additional genotypes showing low V would have provided the avNNet with more data in the lower ranges, most likely resulting in better prediction of V in this range. Furthermore, it has to be noted that the models were trained and tested on released

cultivars, which have all met the requirements of exceeding specific thresholds for the evaluated traits. The evaluation of the cultivars used in this study in diverse, and in part harsh environments, enabled introduction of variation into the data set. However, including early breeding lines in the study would likely have further increased variation in the target traits, specifically by increasing the observations in the lower range of measurements. Consequently, the developed models are likely to perform better when used to predict GY, PC, V, and TW in advanced durum wheat lines, comparable to those used in this study, rather than when applied to pre-breeding material. Nevertheless, the developed models for GY, PC, V, and TW prediction can be directly applied in advanced breeding trials.

In this study, prediction models were developed on spectral data collected during anthesis and thus cannot directly be applied to support breeders in their selection decisions in ongoing trials (e.g. guiding crosses). However, the developed models will allow reductions in costs in breeding programs by harvesting only favorable lines and limiting the need for extensive additional testing . For this reason, the developed models are valuable tools for durum wheat breeders aiming to optimize resource allocation. Moreover, the models may also be useful in farmer's fields for not only forecasting yield but also quality traits well before harvest and can help in making decisions about a late application of nitrogen fertilizer, which has been shown to improve yield and quality (Blandino et al. 2015). Furthermore, the models developed in this study enable the prediction of important quality traits at the plot level and could be used in breeding trials where there is limited seed available.

The results of this study have shown that model performance was comparable across more favorable and less favorable growing environments (Fig. 2). The distinct cluster with strong outliers observed in the residual plot for TW, linked to data obtained in Aranjuez, 2017, can

be explained by two officially declared meteorological heat waves that occurred after multispectral data was acquired. This heat stress resulted in formation of shriveled grains, greatly reducing TW. In addition, strong rain delayed the harvest in that year. Gan et al. (2000) observed a reduction in TW in spring wheat caused by a delayed harvest due to wet weather. In particular, the cultivar Pedroso was shown to respond to these extreme weather events in a strongly negative manner, therefore resulting in a considerable overestimation of its TW (Fig. 2d). Interestingly, the models for the remaining traits did not share this trend. This might be explained partly by the fact that the low TW values observed for Pedroso linked to this specific event are unique for this trait and were not part of the data set used for model training (Supplementary Figure 1). Thus, because the avNNet model was not provided with data in this low TW range during the training step it failed to correctly predict it in the test set. Regarding the other traits evaluated, this absence of a specific data range was not observed (Supplementary Figure 1). Overall, it is important to remember that the test sets used in this study were comprised only of seven randomly selected cultivars (Fig. 3). Therefore, while this facilitated appropriate evaluation of the prediction capability on unseen data, their limited size must be considered. Nevertheless, given the comparable prediction statistics between the training and test sets (Table 2), and the wide phenotypic variation observed in the data used in the present study, the prediction capabilities of models are not expected to differ substantially if applied to other durum wheat cultivars.

What is interesting from a breeding perspective is that when the trained models were applied to the independent test set, the predicted values for cultivars did not significantly differ from the values observed, except in two cases (Fig. 3). This points toward the suitability of the developed models to identify outperforming genotypes in durum wheat breeding populations.

In this study, multispectral data was derived from the plots without automatic removal of plot areas lacking vegetation cover. In some of the plots from trials conducted under harsh growing conditions, large areas with no vegetation cover existed within the plots. These plot areas were removed manually from the multispectral images before extracting the spectral information. This approach was chosen because the resolution of the multispectral camera did not allow for automatic selection with sufficient accuracy. The good prediction capability of the developed models shows that this limitation seems to have a minor influence on the outcome. Nevertheless, development of accurate automatic removal of plot areas lacking vegetation cover holds promise to further improve model prediction capability and the overall throughput of the process, and this will be evaluated in further studies.

Conclusion

Prediction models for GY, PC, V, and TW developed in the framework of this study are of high value for durum wheat breeders aiming to optimize resource allocation. The extent of environments and cultivars considered for the development of the prediction models from image data collected with a UAV-mounted multispectral camera is unprecedented in this field of study. The present work showcases the applicability of multispectral imaging to quality trait prediction in micro-plot breeding trials, an approach that was previously impeded by the low resolution of satellite data. This study also stresses the feasibility of developing strong prediction models based in the use of individual wavebands instead of VIs, even when a multispectral device with a limited number of bands was used. Adding additional genotypes to the training set promises to further improve the prediction capability of the developed models. Integrating an automatic removal of plot areas lacking

vegetation cover in the developed workflow is likely to further boost future quality trait prediction capabilities.

Material and methods

Plant material and field trials

This study is based on 34 post-Green Revolution durum wheat cultivars (*Triticum turgidum* L. subsp. *durum* (Desf) Husn.) widely grown in Spain during the past four decades (Supplementary Table 2). Data was obtained from field trials performed at three locations in Spain, differing in their average annual precipitation and temperature, in four consecutive years (2016 to 2019). In the case of the field trials performed at the experimental stations of the Spanish Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) of Coria del Rio (Cor), Sevilla (37°14´N, 06°03´W, 5 masl), trials were conducted without supplemental irrigation, but under good water conditions supported by a shallow water table. Meanwhile, at the experimental station of Colmenar de Oreja near Aranjuez (40°04´N, 3°31´W, 590 masl), also belonging to the INIA, the trials were conducted with and without supplemental irrigation, indicated in the paper as “irrigated” and “rainfed”, respectively. At the third location close to Valladolid (41°41´N, 04°42´W, 700 masl), which belongs to the Instituto de Tecnología Agraria de Castilla y León (ITACyL), trials were conducted with and without supplemental irrigation throughout the whole trial period, and in 2018 also under low nitrogen conditions, which is referred to as “nitrogen” in the paper. In addition, data obtained from late-planting trials performed at the experimental stations near Valladolid in 2017 and Aranjuez in 2018, indicated as “late” in the paper, were included in this study. Each location *x* growth condition *x* year combination was considered as one environment. Sixteen

environments, for which multispectral data obtained around anthesis was available, were used in this study (Supplementary Table 3). The minimum number of environments in which cultivars were replicated was two, with most cultivars replicated across 14 environments. At all locations field trials were performed using a randomized complete block design with three replications. Cultivars were sown in plots of 7 × 1.5 m^2 with a spacing distance of 0.2 m between rows and a planting density of 250 seeds per square meter.

Grain yield and quality

Plots were harvested using a combine harvester and grain yield (GY) was determined for each plot. A sample of 250 g of whole grain from each plot was cleaned and used for quality analysis. Test weight (TW) was determined according to AACC Method 55-10 (AACC 2011). The percentage of vitreous kernels (V) was visually determined on two lots of 100 seeds per plot. Protein content (PC) on a dry basis was determined using an Infratec 1226 Grain Analyzer (Foss Analytical, Hoganas, Sweden).

Multispectral data

Multispectral data was obtained using a Tetracam micro-MCA (Multiple Camera Array) 12 (Tetracam Inc., Chatsworth, CA, USA) camera mounted on an eight rotor Mikrokopter Oktokopter 6S12 XL UAV (HiSystems GmbH, Moomerland, Germany) around anthesis (Fig. 4). The multispectral camera consisted of twelve independent image sensors and optics, eleven facing downwards, each with user configurable filters of center wavelengths and full-width half-max bandwidths (450 ± 40, 550 ± 10, 570 ± 10, 670 ± 10, 700 ± 10, 720 ± 10, 780 ± 10, 840 ± 10, 860 ± 10, 900 ± 20, 950 ± 40 nm). In addition, an incident light sensor facing

upwards used micro-filters to provide an accurate band-by-band reflectance calibration in real-time.

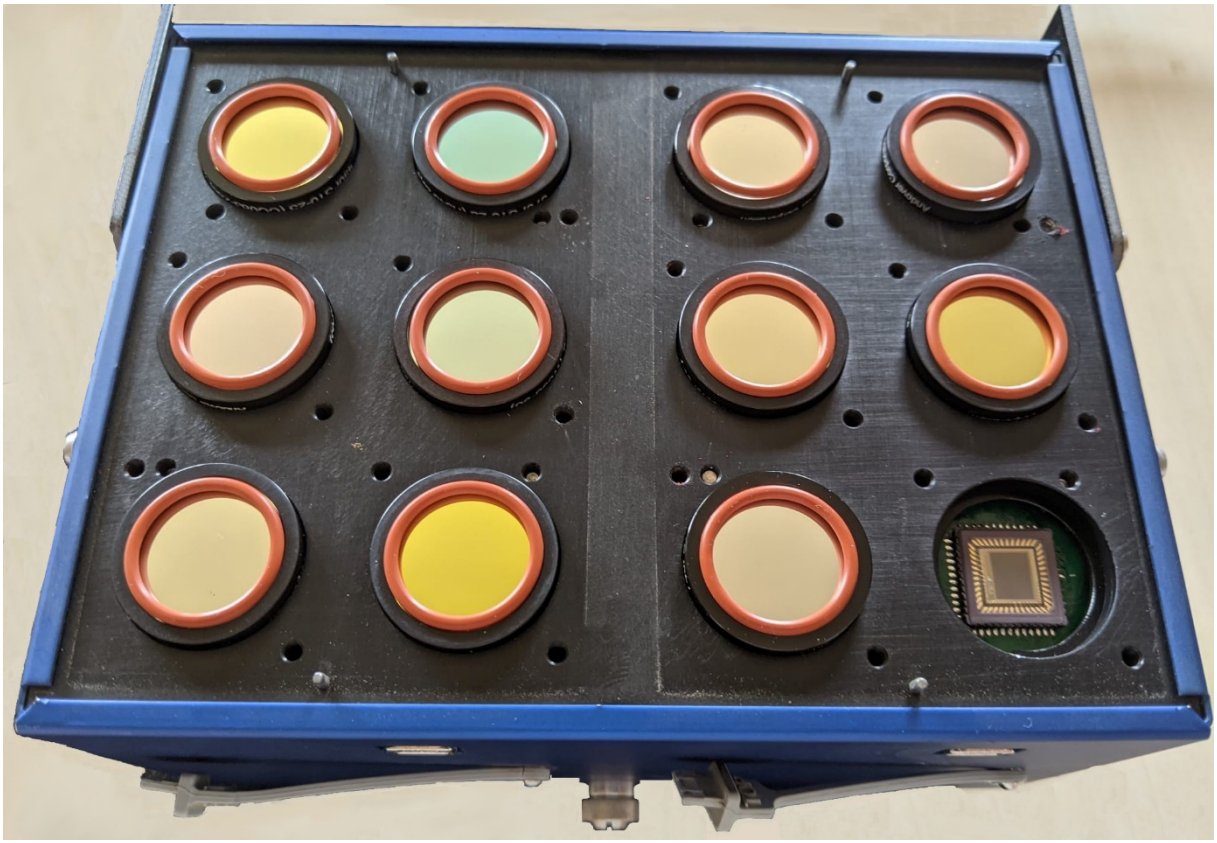


Figure 4: Opened Tetracam micro-MCA. From left to right and top to bottom, spectral sensors of the wavelengths 550 ± 10 nm, 670 ± 10 nm, 840 ± 10 nm, 900 ± 20 nm, 570 ± 10 nm, 700 ± 10 nm, 860 ± 10 nm, 950 ± 40 nm, 780 ± 10 nm, 450 ± 40 nm, 720 ± 10 nm, and incident light sensor (bottom right corner).

Images were taken at an altitude of 50 m every two seconds in order to ensure at least 80% forward and lateral overlap between images (Fig. 5). Subsequently, images were aligned and calibrated to reflectance using PixelWrench II version 1.2.2.2 (Tetracam, Chatsworth, CA, USA). Preprocessed images were then combined into a single orthomosaic using Agisoft

Metashape Professional software (Agisoft LLC, St. Petersburg, Russia, www.agisoft.com). High settings were used for alignment, dense cloud, and mesh formation using otherwise default parameters for the orthomosaic calculation. Images were aligned using the master channel centered at 840nm. Single plots were extracted from the orthomosaic and multispectral band data was obtained using the MosaicTool (Shawn C. Kefauver, https://integrativecropecophysiology.com/software-development/mosaictool/, https://gitlab.com/sckefauver/MosaicTool, University of Barcelona, Barcelona, Spain) integrated as a plugin for the open source image analysis platform FIJI (Fiji is Just ImageJ; http://fiji.sc/Fiji).

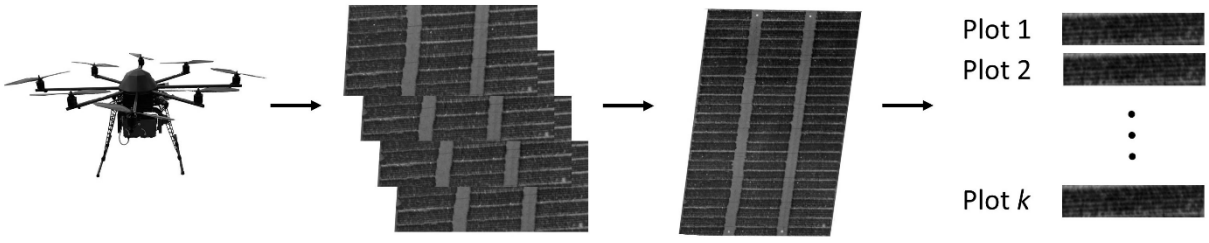


Figure 5: Workflow used for extracting spectral data from single plots. In the first step, multiple spectral images of the field are captured using an unmanned aerial vehicle (UAV); subsequently single images are preprocessed and combined into an orthomosaic; finally single plots are extracted from the orthomosaic.

Subsequently robust trimmed clustering was performed based on the spectral data to identify and remove outliers using the *tclust* package implemented in R software (RCore 2020). In total, 1079 unique data points remained and were used for further analysis.

Statistical analysis

Phenotypic correlation coefficients and broad sense heritability (h^2) for GY, TW and V were calculated on the raw data across environments. To estimate the variance components to be used for the calculation of broad sense heritability, a restricted maximum likelihood (REML) based model was applied. All model parameters, namely genotype (i.e. cultivar), environment, and their interaction were set as random. Broad sense heritability across environments was calculated as:

$$h^2 = \frac{V_G}{V_G + \frac{V_{GE}}{e} + \frac{V_R}{er}} \quad (1)$$

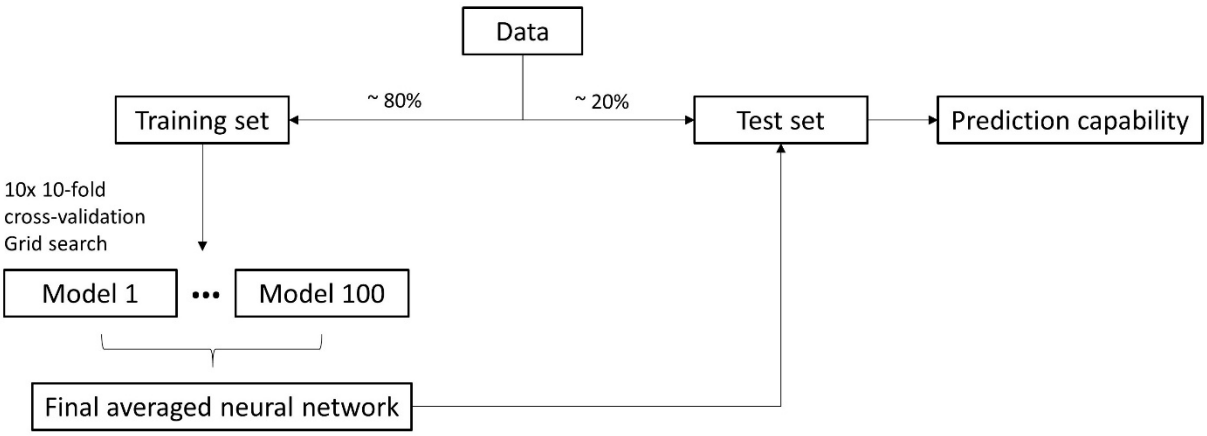
Where genotypic variance is coded by (V_G), genotype *x* environment variance is coded by (V_{GE}), and residual variance is coded by (V_R). The terms *e* and *r* indicate the number of environments and replicates, respectively.

Residuals shown in the residual plots were obtained by subtracting the predicted values, which were generated by applying the trained averaging neural network (avNNet) to the test set, from the respective observed values of the test set. Residuals were then standardized by applying the *scale* function implemented in R software (RCore 2020).

Pearson correlation coefficients between the data of each of the 11 wavebands and the investigated traits were calculated (Supplementary Figure 2).

Model training and quality trait prediction

Training of the averaging neural network (avNNet) and the subsequent prediction of evaluated quality traits using the trained model were performed using the *caret* package implemented in R software (Fig. 6; Core 2020). Input for training the avNNet model was the observed target trait for a genotype *x* environment *x* replication combination and the observed data for each of the 11 wavebands as predictor variables. Each trait was considered separately, in a first step seven cultivars, representing roughly 20% of data, were randomly sampled from the full data set to serve as holdout set, later referred to as test set. The data of the remaining 27 cultivars, roughly 80% of the full data set, was then used for training the avNNet. The model was trained using resampling in form of 10 times repeated 10-fold cross validation, with the final model selected being the one showing lowest root-mean-square error (RMSE). A grid search was used for model tuning to find the optimal value for number of nodes in the hidden layer and decay rate. Following Ripley (1996) each avNNet was based on a total of 100 neural networks with different random number seeds and averaged. The trained model was then applied to the test set to obtain predicted data for GY, PC, V and TW, to assess the prediction capability of the model.



The RMSE and normalized RMSE (nRMSE) were estimated to evaluate the accuracy of the model. The squared Pearson correlation coefficient (R^2) was calculated to obtain an estimate of the phenotypic variance explained by the model. RMSE and normalized RMSE were calculated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{n}} \quad (2)$$

$$nRMSE = \frac{RMSE}{\bar{y}} \quad (3)$$

Where y_i and $\hat{y}_i$ are the measured and the predicted traits, respectively, n is the total number of samples in the test set, and $\bar{y}$ is the mean of the measured trait in the test set.

Data availability statement

Data used in this study and models developed are available upon request from the first author, Thomas Vatter, E mail: tvatter@ub.edu

Acknowledgements

This study was supported in part by the Spanish project PID2019-106650RB-C21 from the Ministerio de Ciencia e Innovación. T. Vatter is recipient of a Juan de la Cierva-Formación, postdoctoral contract, from the Ministerio de Ciencia e Innovación, Spain. S. Kefauver is supported by the Ramon y Cajal RYC-2019-027818-I research fellowship from the Ministerio de Ciencia e Innovación, Spain. J.L. Araus also acknowledges the support of the Catalan

Institution for Research and Advanced Studies (ICREA, Generalitat de Catalunya, Spain), through the ICREA Academia Program.

Author contribution

Thomas Vatter: Data curation, investigation, data analysis, methodology, visualization, writing of original draft, manuscript review and editing.

Adrian Gracia-Romero: Conducted UAV flights for acquisition of multispectral data, image preprocessing, development of software for extraction of spectral data from plots, manuscript review.

Shawn Carlisle Kefauver: Conducted UAV flights for acquisition of multispectral data, image preprocessing, development of software for extraction of spectral data from plots, manuscript review.

José Luis Araus: Conceptualization, funding acquisition, project administration, writing of original draft, manuscript review.

María Teresa Nieto-Taladriz: Management of field trials, grain quality analyses, manuscript review.

Nieves Aparicio: Management of field trials, manuscript review.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supporting Information

Supplementary Figure 1: Density plot of grain yield and quality trait data

Supplementary Figure 2: Correlation matrix showing correlation coefficients between the data of each of the 11 wavebands and the investigated traits

Supplementary Table 1: Evaluation of additional models (i.e. others than avNNet) based on the root mean square error with the training set data as input

Supplementary Table 2: Set of modern semi-dwarf durum wheat cultivars tested in this study with year of release, country of origin and available information on provenance and/or pedigree

Supplementary Table 3: Environments used in this study

References

AACC, A.A.o.C.C. (2011). International Approved Methods of Analysis. Method 55-10.01: Test Weight per Bushel. 11th Edition

Beres, B.L., Rahmani, E., Clarke, J.M., Grassini, P., Pozniak, C.J., Geddes, C.M., Porker, K.D., May, W.E., & Ransom, J.K. (2020). A Systematic Review of Durum Wheat: Enhancing Production Systems by Exploring Genotype, Environment, and Management (G × E × M) Synergies. *Frontiers in Plant Science, 11*

Blanco, A., Mangini, G., Giancaspro, A., Giove, S., Colasuonno, P., Simeone, R., Signorile, A., De Vita, P., Mastrangelo, A.M., Cattivelli, L., & Gadaleta, A. (2012). Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars. *Molecular Breeding, 30*, 79-92

Blandino, M., Vaccino, P., & Reyneri, A. (2015). Late-Season Nitrogen Increases Improver Common and Durum Wheat Quality. *Agronomy Journal, 107*, 680-690

Ben Mariem, S., González-Torralba, J., Collar, C., Aranjuelo, I., & Morales, F. (2020). Durum Wheat Grain Yield and Quality under Low and High Nitrogen Conditions: Insights into Natural Variation in Low- and High-Yielding Genotypes. *Plants, 9*

Caporaso, N., Whitworth, M.B., & Fisk, I.D. (2018). Protein content prediction in single wheat kernels using hyperspectral imaging. *Food chemistry, 240*, 32-42

Core, T.R. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. In

Daaloul Bouacha, O., Nouaigui, S., & Rezgui, S. (2014). Effects of N and K fertilizers on durum wheat quality in different environments. *Journal of Cereal Science, 59*, 9-14

Dexter, J.E., Williams, P.C., Edwards, N.M., & Martin, D.G. (1988). The relationships between durum wheat vitreousness, kernel hardness and processing quality. *Journal of Cereal Science, 7*, 169-181

Dowell, F.E., Maghirang, E.B., Xie, F., Lookhart, G.L., Pierce, R.O., Seabourn, B.W., Bean, S.R., Wilson, J.D., & Chung, O.K. (2006). Predicting Wheat Quality Characteristics and Functionality Using Near-Infrared Spectroscopy. *Cereal Chemistry, 83*, 529-536

Duan, B., Fang, S., Zhu, R., Wu, X., Wang, S., Gong, Y., & Peng, Y. (2019). Remote Estimation of Rice Yield With Unmanned Aerial Vehicle (UAV) Data and Spectral Mixture Analysis. *Frontiers in Plant Science, 10*

Flagella, Z., Giuliani, M.M., Giuzio, L., Volpi, C., & Masci, S. (2010). Influence of water deficit on durum wheat storage protein composition and technological quality. *European Journal of Agronomy, 33*, 197-207

Fu, B.X., Wang, K., Dupuis, B., Taylor, D., & Nam, S. (2018). Kernel vitreousness and protein content: Relationship, interaction and synergistic effects on durum wheat quality. *Journal of Cereal Science, 79*, 210-217

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., Cao, Q., Tian, Y., Zhu, Y., Cao, W., & Liu, X. (2020). Wheat Growth Monitoring and Yield Estimation based on Multi-Rotor Unmanned Aerial Vehicle. *Remote Sensing, 12*, 508

Gagliardi, A., Carucci, F., Masci, S., Flagella, Z., Gatta, G., & Giuliani, M.M. (2020). Effects of Genotype, Growing Season and Nitrogen Level on Gluten Protein Assembly of Durum Wheat Grown under Mediterranean Conditions. *Agronomy, 10*

Galvão, L.S., Breunig, F.M., Santos, J.R.d., & Moura, Y.M.d. (2013). View-illumination effects on hyperspectral vegetation indices in the Amazonian tropical forest. *International Journal of Applied Earth Observation and Geoinformation, 21*, 291-300

Gan, Y.T., McCaig, T.N., Clarke, P., DePauw, R.M., Clarke, J.M., & McLeod, J.G. (2000). Test-weight and weathering of spring wheat. *Canadian Journal of Plant Science, 80*, 677-685

Giancaspro, A., Giove, S.L., Zacheo, S.A., Blanco, A., & Gadaleta, A. (2019). Genetic Variation for Protein Content and Yield-Related Traits in a Durum Population Derived From an Inter-Specific Cross Between Hexaploid and Tetraploid Wheat Cultivars. *Frontiers in Plant Science, 10*, 1509-1509

Gong, Y., Duan, B., Fang, S., Zhu, R., Wu, X., Ma, Y., & Peng, Y. (2018). Remote estimation of rapeseed yield with unmanned aerial vehicle (UAV) imaging and spectral mixture analysis. *Plant Methods, 14*, 70

Gorretta, N., Roger, J.M., Aubert, M., Bellon-Maurel, V., Campan, F., & Roumet, P. (2006). Determining Vitreousness of Durum Wheat Kernels Using near Infrared Hyperspectral Imaging. *Journal of Near Infrared Spectroscopy, 14*, 231-239

Gracia Romero, A., Kefauver, S., Fernandez-Gallego, J., Vergara Diaz, O., Nieto-Taladriz, M., & Araus, J. (2019). UAV and Ground Image-Based Phenotyping: A Proof of Concept with Durum Wheat. *Remote Sensing, 11*, 1244

Hansen, P.M., JØRgensen, J.R., & Thomsen, A. (2002). Predicting grain yield and protein content in winter wheat and spring barley using repeated canopy reflectance measurements and partial least squares regression. *The Journal of Agricultural Science, 139*, 307-318

Hassan, M.A., Yang, M., Rasheed, A., Yang, G., Reynolds, M., Xia, X., Xiao, Y., & He, Z. (2019). A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. *Plant Science, 282*, 95-103

Ji, L., Zhang, L., Rover, J., Wylie, B., & Chen, X. (2014). Geostatistical estimation of signal-to-noise ratios for spectral vegetation indices. *ISPRS Journal of Photogrammetry and Remote Sensing, 96*, 20–27

Kang, Y., Nam, J., Kim, Y., Lee, S., Seong, D., Jang, S., & Ryu, C. (2021). Assessment of Regression Models for Predicting Rice Yield and Protein Content Using Unmanned Aerial Vehicle-Based Multispectral Imagery. *Remote Sensing, 13*, 1508

Kramer, T. (1979). Environmental and genetic variation for protein content in winter wheat (Triticum aestivum L.). *Euphytica, 28*, 209-218

Laidig, F., Piepho, H.-P., Rentel, D., Drobek, T., Meyer, U., & Huesken, A. (2017). Breeding progress, environmental variation and correlation of winter wheat yield and quality traits in German official variety trials and on-farm during 1983–2014. *Theoretical and Applied Genetics, 130*, 223-245

Li, B., Xu, X., Zhang, L., Han, J., Bian, C., Li, G., Liu, J., & Jin, L. (2020). Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging. *ISPRS Journal of Photogrammetry and Remote Sensing, 162*, 161-172

Li, Y.-F., Wu, Y., Hernandez-Espinosa, N., & Peña, R.J. (2013). Heat and drought stress on durum wheat: Responses of genotypes, yield, and quality parameters. *Journal of Cereal Science, 57*, 398-404

Lidon, F., Almeida, A., Leitao, A., Silva, M., Pinheiro, N., Maçãs, B., & Costa, R. (2014). A synoptic overview of durum wheat production in the Mediterranean region and processing following the European Union requirements. *Emirates Journal of Food and Agriculture, 23*, 693

Liu, H.Q., & Huete, A. (1995). A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. *IEEE Transactions on Geoscience and Remote Sensing, 33*, 457-465

Longin, C.F.H., Sieber, A.-N., & Reif, J.C. (2013). Combining frost tolerance, high grain yield and good pasta quality in durum wheat. *Plant Breeding, 132*, 353-358

Magallanes-López, A.M., Ammar, K., Morales-Dorantes, A., González-Santoyo, H., Crossa, J., & Guzmán, C. (2017). Grain quality traits of commercial durum wheat varieties and their relationships with drought stress and glutenins composition. *Journal of Cereal Science, 75*, 1-9

Mahjourimajd, S., Taylor, J., Rengel, Z., Khabaz-Saberi, H., Kuchel, H., Okamoto, M., & Langridge, P. (2016). The Genetic Control of Grain Protein Content under Variable Nitrogen Supply in an Australian Wheat Mapping Population. *PLoS One, 11*, e0159371

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F.B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment, 237*, 111599

Maresma, Á., Ariza, M., Martínez, E., Lloveras, J., & Martínez-Casasnovas, J.A. (2016). Analysis of Vegetation Indices to Determine Nitrogen Application and Yield Prediction in Maize (Zea mays L.) from a Standard UAV Service. *Remote Sensing, 8*, 973

Michel, S., Löschenberger, F., Ametz, C., Pachler, B., Sparry, E., & Bürstmayr, H. (2019). Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. *Theoretical and Applied Genetics, 132*, 2767-2780

Mohammadi, M., Karimizadeh, R., Shafazadeh, M., & Sadeghzadeh, B. (2013). Statistical analysis of durum wheat yield under semi-warm dryland condition. *Australian Journal of Crop Science, 5*, 1292-1297

Nobile, M.A., Baiano, A., Conte, A., & Mocci, G. (2005). Influence of protein content on spaghetti cooking quality. *Journal of Cereal Science, 41*, 347-356

Oury, F.-X., F.-X., Bérard, P., Brancourt-Hulmel, M., Depatureaux, C., Doussinault, G., Galic, N., Giraud, A., Heumez, E., Lecomte, C., Pluchard, P., Rolland, B., Rousset, M., & Trottet, M. (2003). Yield and grain protein concentration in bread wheat : a review and a study of multi-annual data from a French breeding program. *Journal of Genetics and Breeding*, 59-68

RCore, T. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. In

Rharrabti, Y., Elhani, S., Martos, V., García, L., & Moral, D. (2000). Relationship between some quality traits and yield of durum wheat under southern Spain conditions. (pp. 529-531)

Rharrabti, Y., Villegas, D., Royo, C., Martos-Núñez, V., & García del Moral, L.F. (2003). Durum wheat quality in Mediterranean environments: II. Influence of climatic variables and relationships between quality parameters. *Field Crops Research, 80*, 133-140

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press

Rozbicki, J., Ceglińska, A., Gozdowski, D., Jakubczak, M., Cacak-Pietrzak, G., Mądry, W., Golba, J., Piechociński, M., Sobczyński, G., Studnicki, M., & Drzazga, T. (2015). Influence of the cultivar, environment and management on the grain yield and bread-making quality in winter wheat. *Journal of Cereal Science, 61*, 126-132

Samaan, J., El-Khayat, G.H., Manthey, F.A., Fuller, M.P., & Brennan, C.S. (2006). Durum wheat quality: II. The relationship of kernel physicochemical composition to semolina quality and end product utilisation. *International Journal of Food Science & Technology, 41*, 47-55

Schollenberger, J.H., & Kyle, C.F. (1927). Correlation of kernel texture, test weight per bushel, and protein content of hard red spring wheat *Journal of Agricultural Research, 35*, 1137-1151

Söderström, M., Börjesson, T., Pettersson, C.-G., Nissen, K., & Hagner, O. (2010). Prediction of protein content in malting barley using proximal and remote sensing. *Precision Agriculture, 11*, 587-599

Suprayogi, Y., Pozniak, C.J., Clarke, F.R., Clarke, J.M., Knox, R.E., & Singh, A.K. (2009). Identification and validation of quantitative trait loci for grain protein concentration in adapted Canadian durum wheat populations. *Theoretical and Applied Genetics, 119*, 437-448

Symons, S.J., Van Schepdael, L., & Dexter, J.E. (2003). Measurement of Hard Vitreous Kernels in Durum Wheat by Machine Vision. *Cereal Chemistry, 80*, 511-517

Tan, C., Zhou, X., Zhang, P., Wang, Z., Wang, D., Guo, W., & Yun, F. (2020). Predicting grain protein content of field-grown winter wheat with satellite images and partial least square algorithm. *PLoS One, 15*, e0228500-e0228500

Tattaris, M., Reynolds, M.P., & Chapman, S.C. (2016). A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. *Frontiers in Plant Science, 7*

Thorwarth, P., Piepho, H.P., Zhao, Y., Ebmeyer, E., Schacht, J., Schachschneider, R., Kazman, E., Reif, J.C., Würschum, T., & Longin, C.F.H. (2018). Higher grain yield and higher grain protein deviation underline the potential of hybrid wheat for a sustainable agriculture. *Plant Breeding, 137*, 326-337

Thungo, Z., Shimelis, H., Odindo, A., & Mashilo, J. (2020). Genotype-by-environment effects on grain quality among heat and drought tolerant bread wheat (Triticum aestivum L.) genotypes. *Journal of Plant Interactions, 15*, 83-92

Vergara Diaz, O., Kefauver, S., Araus, J., & Iker, A. (2020a). Development of novel technological approaches for a reliable crop characterization under changing environmental conditions. *NIR news, 31*, 14-19

Vergara-Diaz, O., Vatter, T., Kefauver, S.C., Obata, T., Fernie, A.R., & Araus, J.L. (2020b). Assessing durum wheat ear and leaf metabolomes in the field through hyperspectral data. *The Plant Journal, 102*, 615-630

Wang, K., & Fu, B.X. (2020). Inter-Relationships between Test Weight, Thousand Kernel Weight, Kernel Size Distribution and Their Effects on Durum Wheat Milling, Semolina Composition and Pasta Processing Quality. *Foods (Basel, Switzerland), 9*, 1308

Wiegmann, M., Backhaus, A., Seiffert, U., Thomas, W.T.B., Flavell, A.J., Pillen, K., & Maurer, A. (2019). Optimizing the procedure of grain nutrient predictions in barley via hyperspectral imaging. *PLOS ONE, 14*, e0224491

Xie, F., Pearson, T., Dowell, F.E., & Zhang, N. (2004). Detecting Vitreous Wheat Kernels Using Reflectance and Transmittance Image Analysis. *Cereal Chemistry, 81*, 594-597

Xu, S., Yu, J., Chen, Y., Tabori, M., Wang, X., McCallum, B., Fedak, G., Blackwell, B., Xue, A., Yang, Z., & Khanizadeh, S. (2018). Evaluation of Selected Advanced Spring Wheat Germplasm Lines In Eastern Canada. *Sustainable Agriculture Research, 7*, 63

Xu, X., Teng, C., Zhao, Y., Du, Y., Zhao, C., Yang, G., Jin, X., Song, X., Gu, X., Casa, R., Chen, L., & Li, Z. (2020). Prediction of Wheat Grain Protein by Coupling Multisource Remote Sensing Imagery and ECMWF Data. *Remote Sensing, 12*, 1349

Xue, C., Matros, A., Mock, H.-P., & Mühling, K.-H. (2019). Protein Composition and Baking Quality of Wheat Flour as Affected by Split Nitrogen Application. *Frontiers in Plant Science, 10*

Xue, J., & Su, B. (2017). Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors, 2017*, 1353691

Zhang, G., Chen, R.Y., Shao, M., Bai, G., & Seabourn, B.W. (2021). Genetic analysis of end-use quality traits in wheat. *Crop Science, n/a*, 1-15

Zhao, H., Song, X., Yang, G., Li, Z., Zhang, D., & Feng, H. (2019). Monitoring of Nitrogen and Grain Protein Content in Winter Wheat Based on Sentinel-2A Data. *Remote Sensing, 11*, 1724

Zhou, X., Kono, Y., Win, A., Matsui, T., & Tanaka, T.S.T. (2021). Predicting within-field variability in grain yield and protein content of winter wheat using UAV-based multispectral imagery and machine learning approaches. *Plant Production Science, 24*, 137-151

Zhu, G., Ju, W., Chen, J.M., & Liu, Y. (2014). A Novel Moisture Adjusted Vegetation Index (MAVI) to Reduce Background Reflectance and Topographical Effects on LAI Retrieval. *PLoS One, 9*, e102560

Figure/table legends

Table 1. Descriptive statistics for training and test set and heritability.

Figure 1: Scatterplots for observed and predicted data. (A) grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Blue dots depict data obtained by applying the trained averaging neural network (avNNet) model to the test set, while grey dots depict data observed in the training set. The red line shows a linear regression line based on the test set data, with a significance of $p < 0.001$ for all traits. As a reference, the black dashed line indicates a 1:1 relationship. The prediction statistics depicted in the plots refer to the test set.

Table 2. Prediction statistics for grain yield and quality traits for the training and test sets.

Figure 2: Standardized residual plots of averaging neural network (avNNet) models applied to the validation sets. The X-axis shows the predicted data for, **(A)** grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Colors refer to the four main treatments,

irrigated, rainfed, late, and nitrogen. The grey line shows the locally estimated scatterplot smoothing (LOESS) line across all main treatments.

Figure 3: Boxplots for observed and predicted data of the validation sets merged by cultivar. (A) grain yield, **(B)** protein content, **(C)** vitreousness, and **(D)** test weight. Blue shading indicates observed data, while red shading indicates predicted data obtained by applying the trained averaging neural network (avNNet) model to the test set. The number of data points in the validation set for each cultivar is indicated by the letter n. Cultivars comprising the test set vary because for each trait, seven cultivars were randomly sampled from the full data set.