1 Hyperspectral data mining to identify relevant canopy
2 spectral features for estimating durum wheat growth,
3 nitrogen status, and grain yield

4 K. R. Thorp^{a,*}, G. Wang^b, K. F. Bronson^a, M. Badaruddin^c, J. Mon^a

5 ^a*USDA-ARS, U.S. Arid Land Agricultural Research Center, 21881 N Cardon Ln,*
6 *Maricopa, Arizona, 85138*

7 ^b*Bridgestone Americas Agricultural Operations, 4140 West Harmon Road, Eloy,*
8 *Arizona, 85131*

9 ^c*Maricopa Agricultural Center, University of Arizona, 37860 West Smith-Enke Road,*
10 *Maricopa, Arizona, 85138*

11 **Abstract**

While hyperspectral sensors describe plant canopy reflectance in greater detail than multispectral sensors, they also suffer from issues with data redundancy and spectral autocorrelation. Data mining techniques that extract relevant spectral features from hyperspectral data will aid the development of novel sensors for plant trait estimation. The objectives of this research were to 1) compare broad-band reflectance, narrow-band reflectance, and spectral derivatives for estimation of durum wheat traits in the field and 2) develop a genetic algorithm to identify the most relevant spectral features for durum wheat trait estimation. Experiments at Maricopa, Arizona during the winters of 2010-2011 and 2011-2012 tested six durum wheat cultivars with six split-applied nitrogen (N) fertilization rates. Durum wheat traits, including leaf area index, canopy dry weight, and plant N content, were measured from destructive biomass samples on four occasions in each growing season. Grain

*Corresponding author

Preprint submitted to Elsevier

Email address: `kelly.thorp@ars.usda.gov` (K. R. Thorp)

January 17, 2017

yield and grain N content were also measured. Canopy spectral reflectance data in 701 narrow wavebands from 350 nm to 1050 nm were collected weekly using a field spectroradiometer. First- and second-order spectral derivatives were calculated using Savitzky-Golay filtering. The narrow-band data were also used to estimate reflectance in broad wavebands, as typically collected by two commercial multispectral instruments. Partial least squares regression (PLSR) compared the ability of each spectral data set to estimate each measured durum wheat trait. A genetic algorithm was developed to mine narrow-band canopy reflectance and spectral derivative data for spectral features that improved estimates of durum wheat traits. Multispectral data in 4 broad bands estimated leaf area index, canopy dry weight, and plant N content with root mean squared errors of cross validation (RMSECV) between 33.0% and 67.6%, while hyperspectral data in 701 narrow bands reduced RMSECV to values between 19.3% and 36.3%. Use of the genetic algorithm to identify less than 25 relevant spectral features further reduced RMSECV to values between 15.1% and 30.7%. Grain yield was optimally estimated from canopy spectral measurements between 110 and 130 days after planting with RMSECV less than 7.6% using the genetic algorithm approach. The timing corresponded to anthesis and early grain fill when presence of wheat heads likely affected canopy spectral reflectance. By using a genetic algorithm to mine hyperspectral reflectance and spectral derivative data, durum wheat traits were optimally estimated from a subset of relevant canopy spectral features.

12 *Keywords:* derivative spectra, genetic algorithm, partial least squares
13 regression, spectral reflectance, spectroradiometer

14 **1. Introduction**

15 Hyperspectral reflectance data in the visible (VIS) and near-infrared
16 (NIR) wavebands offer many opportunities to improve crop production sys-
17 tems for wheat (*Triticum aestivum* L.; *Triticum durum* Desf.) and other
18 agricultural commodities. For example, many spectral indices have been de-
19 veloped for detecting nitrogen (N) concentration in wheat canopies (Chen
20 et al., 2010; Feng et al., 2008; Fitzgerald et al., 2010; Mahajan et al., 2014;
21 Tilling et al., 2007). By monitoring plant N status during the growing season,
22 N fertilizer management can be tailored to the crop need, thus maintaining
23 N fertility for high photosynthetic yield while minimizing N loss to the envi-
24 ronment (Raun et al., 2008). Mid-season spectral reflectance data has also
25 been used to predict wheat yield and grain protein content (Li et al., 2012;
26 Serrano et al., 2000; Xue et al., 2007), both of which are economically im-
27 portant to wheat growers. Higher protein in wheat grain garners a premium
28 price for the grower, determines the end-product to be made from the grain,
29 and is dependent on water and N management (Ottman et al., 2000). Apari-
30 cio et al. (2002) developed spectral indices to estimate leaf area index (LAI)
31 and biomass for use as selection criteria in wheat breeding programs. Spec-
32 tral reflectance sensors and other sensing instrumentation are increasingly
33 important for field-based crop improvement and genetics research, as plant

34 scientists search for links between desirable plant traits and genes that con-
35 trol those traits (Araus and Cairns, 2014; Thorp et al., 2015; White et al.,
36 2012). Spectral reflectance data also integrate with crop growth and radia-
37 tive transfer models (Haboudane et al., 2004; Thorp et al., 2012), which can
38 assist retrievals of crop biophysical variables for a variety of remote sensing
39 applications. Rapid diagnostic tools are needed for crop monitoring and deci-
40 sion support in diverse agricultural applications, and hyperspectral data from
41 remote and proximal sensing provide ample information for development of
42 such tools.

43 Modern radiometric instruments provide hyperspectral reflectance data
44 in hundreds or thousands of narrow wavebands, thus a major challenge is
45 to intelligently subset the data by identifying spectral characteristics that
46 are meaningful for a given application. Many researchers have calculated
47 narrow-band spectral reflectance indices by ratioing reflectance data in key
48 wavebands or incorporating narrow-band spectral data into the widely known
49 Normalized Difference Vegetation Index (NDVI) equation (Chen et al., 2010;
50 Feng et al., 2008; Hansen and Schjoerring, 2003; Thorp et al., 2004). The
51 spectral index approach extends from an earlier era, when radiometric mea-
52 surements were available in only a few broad wavebands (Bannari et al.,
53 1995). With the advent of hyperspectral systems, there is opportunity to
54 both refine older spectral indices and develop novel data analysis techniques
55 that exploit the higher spectral resolution and nearly contiguous nature of hy-
56 perspectral data. For operational purposes, simpler radiometric instruments

could be developed based on the findings of hyperspectral data analyses. Alternatively, the analyses may reveal that narrow-band, contiguous reflectance data from a hyperspectral sensor is preferable or offers greater accuracy for a given sensing application.

Multicollinearity among neighboring wavebands is a persistent problem for hyperspectral data analysis. Statistical procedures such as principal component regression (PCR) and partial least squares regression (PLSR) reduce multicollinearity and dimensionality by decomposing the hyperspectral data into a subset of independent factors, with which crop biophysical traits can be regressed. Rather than using key wavebands for calculation of spectral indices, these methods incorporate full-spectrum reflectance data into statistical models for crop trait estimation. For example, Ecartot et al. (2013) developed PLSR models to estimate durum wheat leaf N content ($r^2 = 0.95$) and leaf mass per unit area ($r^2 = 0.94$) from spectral reflectance data measured between 400 and 2500 nm with a leaf clip. As demonstrated by Fu et al. (2014), PLSR analysis can incorporate not only spectral reflectance data but also spectral indices and other spectral metrics related to band depth and continuum-removed spectra. Hansen and Schjoerring (2003) compared estimates of wheat biophysical traits using 1) linear regression on narrow-band NDVI with optimal wavebands and 2) PLSR with all wavebands from 400 to 900 nm. The NDVI approach better estimated LAI and chlorophyll concentration, while the PLSR approach better estimated green biomass weight and leaf N concentration. In a similar comparison, Thorp et al. (2015) reported

that spectral reflectance data from 400 to 2400 nm, analyzed with PLSR, explained variability in cotton (*Gossypium barbadense* L.) leaf water content, specific leaf mass, leaf chlorophyll $a + b$ content, and LAI better than linear correlations with common vegetation indices. Of the myriad techniques for analysis of hyperspectral data, PLSR has become highly popular in recent years (Fu et al., 2014; Kaleita et al., 2006; Li et al., 2012; Thorp et al., 2011). However, PLSR usually emphasizes full-spectrum, contiguous data, and efforts to identify and subset relevant spectral features are often ignored.

Another option for analysis of hyperspectral reflectance data is the computation of spectral derivatives, which quantify slope, curvature and higher-order aspects of reflectance spectra. Peaks in derivative spectra can identify the “red edge” position between 680 and 750 nm in crop reflectance data, which results from the contrast of red light absorption by plant chlorophyll and NIR scattering by plant biomass (Demetriades Shah et al., 1990; Horler et al., 1983). Tsai and Philpot (1998) tested algorithms for reflectance data smoothing and derivative computation, including methods based on filter convolution (Savitzky and Golay, 1964) or finite divided difference approximation. Thorp et al. (2004) developed first- and second-derivative spectral indices by integrating derivative spectra within the range of derivative spectral peaks. Second derivative indices were particularly useful for estimating soybean (*Glycine max* (L.) Merr.) canopy cover ($r < 0.89$). Chen et al. (2010) developed the Double-peak Canopy Nitrogen Index (DCNI), which contrasted first derivative spectra at two locations near the red edge to esti-

103 mate plant N concentration in maize (*Zea mays* L.) and wheat ($r^2 = 0.64$).
104 In similar research, Feng et al. (2014) analyzed 20 spectral derivative features
105 near the red edge position and developed a novel index to estimate wheat leaf
106 N concentration ($r^2 < 0.85$). Spectral derivative analysis can reveal spectral
107 features that may not be apparent in reflectance data alone.

108 To overcome the multidimensional nature of hyperspectral data, genetic
109 algorithms have been developed to reduce dimensionality and mine the data
110 for spectral features that correlate to crop traits (Leardi, 2000). For exam-
111 ple, Yao and Tian (2003) combined a genetic algorithm with PCR to reduce
112 the dimensionality of hyperspectral images from 60 wavebands to less than
113 26 wavebands. The genetic algorithm removed wavebands that contributed
114 little to PCR models for maize leaf chlorophyll content, plant population,
115 and hybrid, which improved PCR model performance compared to models
116 based on the full spectrum, 60-band data. Similarly, Kaleita et al. (2006)
117 combined a genetic algorithm with PLSR to identify spectral features predic-
118 tive of tasseling and pollen shed in maize. Their algorithm 1) incorporated
119 operators for the maximum, minimum, and median reflectance values and
120 the slope and curvature of reflectance data over a range of wavebands and 2)
121 identified the spectral operators that were most influential for PLSR-based
122 estimation of maize canopy traits. Among the various hyperspectral data
123 analysis approaches, genetic algorithms uniquely offer the ability to mine hy-
124 perspectral data sets for spectral features relevant to a given sensing appli-
125 cation (Kaleita et al., 2006). More studies should incorporate this approach

126 to elucidate meaningful relationships between spectral reflectance data and
127 agricultural crop characteristics.

128 While there is a growing body of literature on myriad analysis techniques
129 for hyperspectral data, few studies comprehensively evaluate, compare, or
130 integrate multiple approaches, likely because the learning curve can be steep
131 and advanced computational skills are often necessary. Notable examples
132 that do contrast multiple techniques include Fu et al. (2014), Kaleita et al.
133 (2006), and Thorp et al. (2015). Literature is also now saturated with mul-
134 tiple examples of PLSR modeling to estimate a variety of crop traits from
135 spectral data. While PLSR is a valid and useful analysis approach, it is less
136 informative when treated solely as a black box statistical model. Hyperspec-
137 tral data analyses should always push toward better understanding of the
138 mechanisms for spectral reflectance from crop canopies and identification of
139 the important wavebands or spectral features that contribute to crop trait
140 estimates. Genetic algorithms, combined with PLSR, have potential for great
141 advancement toward the latter goals, but investigations using this approach
142 are far fewer than those using PLSR alone.

143 The overall objective of this study was to investigate canopy spectral
144 reflectance data and related data analysis techniques for estimating durum
145 wheat LAI, canopy dry weight, plant N content, grain yield, and grain N
146 content. Specific objectives were to 1) use PLSR to compare the ability of
147 different spectral data sets, including broad-band canopy reflectance data,
148 narrow-band canopy spectra, and derivative spectra, to explain variability in

149 durum wheat traits and 2) combine PLSR with a genetic algorithm to reduce
150 hyperspectral data dimensionality and identify relevant spectral features for
151 estimating durum wheat canopy traits, grain yield, and grain quality char-
152 acteristics.

153 **2. Materials and Methods**

154 *2.1. Field experiments*

155 Durum wheat experiments were conducted at the University of Arizona's
156 Maricopa Agricultural Center (MAC) near Maricopa, Arizona (33.068° N,
157 111.971° W, 360 m above sea level) over the winters of 2010-2011 and 2011-
158 2012 (Liang et al., 2014). The soil texture at the site was predominantly
159 sandy loam and sandy clay loam, as determined by textural analysis of soil
160 samples. A split-plot experimental design was used with four replications of
161 six durum wheat cultivars (Duraking, Topper, Kronos, Havasu, Orita, and
162 Ocotillo) as main-plot treatments and five split-applied N fertilizer rates as
163 sub-plot treatments. Based on 2010-2011 experimental results, a sixth N rate
164 was added in 2011-2012. Durum wheat was planted on 15 December 2010 and
165 9 December 2011 with a row spacing of 19.05 cm. Using a portable fertilizer
166 spreader, urea N fertilizer was split-applied according to the rate and timing
167 schedule reported in Table 1. Seasonal fertilization amounts ranged from 0
168 to 403 kg N ha⁻¹. A sudangrass (*Sorghum bicolor* (L.) Moench) cover crop
169 was grown in the summers before durum wheat planting to remove excess N
170 from the soil profile. In both seasons, the entire experimental area was flood
171 irrigated to avoid water deficits. Seasonal irrigation amounted to 840 mm in

172 2010-2011 and 710 mm in 2011-2012, applied during nine irrigation events
173 from early December to the end of April. Precipitation amounted to 29 and
174 41 mm in the 2010-2011 and 2011-2012 growing seasons, respectively. Further
175 details about the field investigation are provided by Liang et al. (2014).

176 *2.2. Biomass and yield measurements*

177 Durum wheat plants were destructively sampled from all experimental
178 plots on four dates in 2011 (January 18, February 24, March 22, and April 7)
179 and 2012 (January 10, February 16, March 13, and April 4). Plants in two
180 0.5 m row lengths within each plot were cut at the soil surface and bagged.
181 Within 24 hours, plants were dissected into component plant parts, including
182 leaves, stems, and spikes. The total leaf area of each sample was measured
183 on an area meter (model 3100, Li-Cor, Lincoln, Nebraska) and used to calcu-
184 late leaf area index (LAI). Samples were oven-dried at 65°C with ventilation
185 until constant weight was achieved. Canopy dry weight per hectare (“canopy
186 weight” hereafter) was calculated from oven-dried biomass weight measure-
187 ments. The dried biomass was then finely ground, and samples were prepared
188 for analysis of plant N content using a Carlo Erba elemental analyzer (model
189 NA1500 N/C, Carlo Erba Instruments, Milan, Italy). Mature durum wheat
190 was harvested with a plot combine on 2 June 2011 and 24 May 2012. Grain
191 samples were oven dried to estimate dry grain weight per hectare (“yield”
192 hereafter) for each plot, and grain N content was measured with the Carlo
193 Erba elemental analyzer.

194 Hierarchical linear mixed modeling was used to assess effects of cultivar

and N fertilizer rate on the biomass and yield measurements. Cultivar, N fertilizer rate, and their interaction were modeled as fixed effects. Block and its interaction with cultivar were modeled as random effects. Hierarchical tests required fitting random effects with 1) cultivar fixed effects alone, 2) N fertilizer rate fixed effects alone, 3) both cultivar and N fertilizer rate fixed effects, and 4) water and N fertilizer fixed effects and their interaction. Likelihood ratio tests were used to compare these hierarchical models, which showed whether the measurement was different among cultivar, N fertilizer rate, or their interaction. Linear mixed models were computed using the “lme4” package within the R Project for Statistical Computing software (<http://www.r-project.org>).

2.3. Radiometric measurements

Ground-based radiometric measurements were collected weekly over each experimental plot using a portable field spectroradiometer (GER 1500, Spectra Vista Corp., Poughkeepsie, New York). Radiometric information was reported in 512 narrow wavebands from 268 to 1095 nm with bandwidth ranging from 1.5 to 2.1 nm. The instrument was equipped with an 18° field-of-view fiber optic. A wand constructed from PVC tubing was used to position the fiber optic at a nadir view angle approximately 1.8 m above the soil surface. Spectral measurements typically occurred in the morning around the time of a 57° solar zenith angle, which insured consistent canopy bidirectional reflectance effects over the entire growing season. Three spectral measurements were collected over each plot, which was limited by the

size of the instrument's onboard memory and the need to measure more than 100 plots at optimum solar zenith angle. Frequent radiometric observations of a calibrated, 0.6 m², 99% Spectralon panel (Labsphere, Inc., North Sutton, New Hampshire) were used to characterize incoming solar irradiance throughout the data collection period. Canopy reflectance factors in each waveband were computed as the ratio of the canopy radiance over the corresponding time-interpolated value for Spectralon panel radiance. Using spline interpolation, reflectance factors were adjusted to integer wavelength values with spectral resolution of 1 nm. Due to instrument sensitivity issues at the limits of the detector, subsequent spectral analyses were based on 701 reflectance factors from 350 to 1050 nm. Radiometric measurements over each experimental plot were averaged to estimate the canopy spectral reflectance on each measurement date.

2.4. Broad-band calculations

To compare the narrow-band spectral data with information typically collected by broad-band, multispectral radiometers, the GER1500 data were averaged within the wavebands measured by two commercial, hand-held instruments (MSR5 and MSR87, Cropscan, Inc., Rochester, MN). Averaged GER1500 spectral reflectance data were used to mimic four wavebands from the Cropscan MSR5 instrument: 450 to 520 nm, 520 to 600 nm, 630 to 690 nm, and 760 to 900 nm. A fifth waveband measured by the Cropscan MSR5 was outside the spectral range of the GER1500 instrument. Similarly, GER1500 data was used to estimate eight wavebands from the Cropscan

MSR87 instrument. Band widths were each 9 nm with band centers at 460, 510, 560, 610, 660, 710, 760, and 810 nm. Cropsan instruments were not used to collect radiometric data in the field experiments, but data from the GER1500 instrument was used to estimate reflectance data typically measured by Cropsan.

2.5. Derivative calculations

Derivative spectra were calculated using Savitzky and Golay (1964) smoothing and filtering on the GER1500 spectral reflectance measurements. The method convolved shaped filters with size $2m + 1$ over the reflectance data to calculate spectral derivatives while simultaneously smoothing the noise introduced by derivative calculations. Savitzky and Golay (1964) provided tables of coefficients for various filter sizes, but Madden (1978) later developed polynomial equations that provide Savitzky and Golay (1964) coefficients for filters of various sizes and for up to fifth-order derivatives. In this study, the second-order polynomial equations of Madden (1978) were used to derive filter coefficients for calculation of first- and second-order spectral derivatives. A value of 3 was used for the m parameter in these equations, which provided a total filter size of 7 ($2m + 1$). Thus, spectral derivatives were calculated based on spectral reflectance data within a bandwidth of 7 nm. A Python script (<http://www.python.org>) was developed to conduct the spline interpolation (discussed above), calculate filter coefficients from the Madden (1978) equations, and obtain first- and second-order spectral derivatives by filter convolution.

264 *2.6. PLSR modeling*

265 PLSR was used to compare the ability of different spectral data sets to
266 estimate durum wheat traits: LAI, canopy weight, plant N content, grain
267 yield, and grain N content. Thorp et al. (2011) provided the details on the
268 PLSR methodology used in the present study. Briefly, if $\mathbf{Y}$ is an $n \times 1$ vector
269 of responses (plant measurements) and $\mathbf{X}$ is an n -observation by p -variable
270 matrix of predictors (a set of spectral metrics in p wavebands), PLSR aims
271 to decompose $\mathbf{X}$ into a set of A orthogonal scores such that the covariance
272 with corresponding $\mathbf{Y}$ scores is maximized. The X-weight and Y-loading
273 vectors that result from the decomposition are used to estimate the vector
274 of regression coefficients, β_{PLS} , such that

$$\mathbf{Y} = \mathbf{X}\beta_{PLS} + \boldsymbol{\epsilon} \quad (1)$$

275 where $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms.

276 The “pls” package (Mevik and Wehrens, 2007) within the R Project for
277 Statistical Computing (<http://www.r-project.org>) was used for PLSR in this
278 study. PLSR models were developed individually for each plant trait re-
279 gressed with different spectral data sets, including the two broad-band data
280 sets based on wavebands from Cropscan instruments, the narrow-band re-
281 flectance data measured with the GER1500 instrument, first-order derivative
282 spectra, second-order derivative spectra, and a combination of narrow-band
283 reflectance data with first- and second-order spectral derivatives. To choose
284 the appropriate number of factors for each model (A from above), leave-one-

285 out cross validation was used to test model predictions for independent data.
286 Results were reported for models with the number of factors that minimized
287 the root mean squared error of cross validation (RMSECV). Because PLSR
288 decomposed the spectral data sets to a set of explanatory factors, the PLSR
289 goodness-of-fit statistics (i.e., RMSECV) summarized the information con-
290 tent of different spectral data sets with regard to estimating plant traits, and
291 thus the predictive capability of each spectral data set could be compared.

292 *2.7. Genetic algorithm*

293 Following the example of Kaleita et al. (2006), a genetic algorithm was
294 developed to 1) subsample the narrow-band reflectance data and first- and
295 second-order derivative spectra, 2) identify up to 25 spectral features com-
296 puted as the mean of spectral data over a range of wavelengths, 3) assess
297 the goodness-of-fit of the spectral features to estimate plant traits via PLSR,
298 and 4) iterate the process to identify the optimum set of spectral features. A
299 genetic algorithm is a computational method designed to solve problems by
300 mimicking the process of natural selection for biological species evolution. In
301 this study, the genetic algorithm was programmed to establish a population
302 of 5000 individuals. Each individual was programmatically characterized by
303 a single chromosome with 25 genes. Each gene characterized a unique spec-
304 tral feature on the chromosome and was comprised of three genetic loci that
305 could take the value of several pre-defined alleles. At the first genetic locus,
306 an integer operator indicated whether the gene represented 1) canopy spectral
307 reflectance, 2) the first-order derivative of canopy spectral reflectance (i.e.,

308 slope), 3) the second-order derivative of canopy spectral reflectance (i.e., cur-
309 vature), or 4) no spectral data set. In the latter case, the gene was “turned
310 off” and had no effect on the individual’s goodness-of-fit. The second genetic
311 locus contained an integer value that represented the starting wavelength for
312 the spectral feature. It could range from 350 to 1050 nm. The third genetic
313 locus contained an integer value that defined the bandwidth of the spectral
314 feature. It could range from 1 to 200 nm with the caveat that no spectral
315 feature could extend beyond 1050 nm.

316 The fitness of each individual was calculated as the minimum RMSECV
317 from PLSR. Cross-validation was accomplished using random segment se-
318 lection, where the spectral data set was divided into 10 segments, selected
319 randomly. PLSR models were iteratively calibrated using 9 segments and
320 evaluated for the one left out. When the genetic algorithm finished iterating,
321 the final individuals in the population were reevaluated using leave-one-out
322 cross validation. This permitted better comparisons to PLSR models de-
323 scribed in the previous section. Because leave-one-out cross validation was
324 more computationally expensive, its use during the iteration of the genetic
325 algorithm could not be justified.

326 The survival rate of individuals in a generation was 50%. Thus, at each
327 new generation of the population, individuals were sorted by fitness score,
328 and the bottom half of individuals with poorest fitness were eliminated. New
329 individuals were created from the remaining individuals until the population
330 again contained 5000 individuals. The crossover rate was 90%. Thus, 10%

331 of new individuals were created via asexual reproduction and existed as the
332 identical twin or genetic duplicate of another individual. Otherwise, new in-
333 dividuals were created via a mating tournament between two subsets of eight
334 individuals, randomly selected from the population. Within each subset, the
335 fittest individual won the mating tournament 90% of the time. Otherwise, a
336 random individual was chosen for mating. Mating occurred via a two-point
337 crossover method, where the chromosome of the new individual was sectioned
338 in two random locations. Genes from one mate were copied to the two outer
339 sections, and genes from the other mate were copied to the middle section.

340 The genes of each new individual were subjected to a mutation process.
341 The mutation rate varied with each generation, generally with decreasing
342 mutation rates as the population matured. If the fitness score of the fittest
343 individual in the $(n + 1)$ th generation was identical to that for the n th gener-
344 ation, the mutation rate was reduced by 0.1%. If a new fittest individual was
345 found in the $(n + 1)$ th generation, the mutation rate was increased by 0.1%.
346 Limits on the mutation rate were between 0% and 100%, and the initial mu-
347 tation rate was 100%. According to the mutation rate, the genes of a new
348 individual's chromosome were modified on a gene-by-gene basis. If a gene
349 was selected for mutation, one of the its three loci was randomly selected for
350 modification. If the first locus was selected, the gene was either switched off
351 or switched to represent a different spectral data set type. If the second or
352 third locus was chosen, the starting waveband or bandwidth, respectively,
353 was adjusted by a value between ± 10 nm, while preserving the limits on

354 these values. Following mutation, the individual’s chromosome was checked
355 for genetic duplication, and any duplicates were replaced with a new gene,
356 chosen randomly until uniqueness was achieved. Duplicate genes on a chro-
357 mosome were undesirable, because it would lead to redundant information
358 in the PLSR model for that individual.

359 The genetic algorithm terminated when either 1) genetic diversity was
360 sufficiently reduced (i.e., more than half of the population was comprised of
361 identical twins) or 2) more than 100 generations had passed without finding
362 a new fittest individual. Typically, the algorithm iterated over 1100 to 1400
363 generations. Genes from the overall fittest individual were further analyzed
364 to identify the spectral features that led to optimum prediction of plant traits.

365 The genetic algorithm approach combined with PLSR was used to solve
366 for the optimum spectral features to estimate each of the durum wheat plant
367 traits individually: LAI, canopy weight, plant N content, grain yield, and
368 grain N content. In addition, the approach was used to evaluate different
369 spectral data sets: the narrow-band reflectance data, first-order derivative
370 spectra, second-order derivative spectra, and a combination of these three
371 data sets. The genetic algorithm was developed in the Python scripting
372 language, using the “rpy2” package to access R statistical software functions
373 for PLSR calculations.

3. Results and Discussion

3.1. Field data

Hyperspectral measurements of the wheat canopy followed typical patterns for spectral reflectance of vegetation in both the 2010-2011 (Fig. 1a) and 2011-2012 (Fig. 2a) growing seasons. Generally, scattering of near-infrared radiation led to greater variability in reflectance from 760 to 1050 nm as compared to the visible spectrum (400 to 700 nm) where chlorophyll absorbed radiation. The plotted reflectance data was collected over all cultivar and N fertilization treatments at 118 days after planting in the 2010-2011 growing season (12 April 2011) and at 111 days after planting in the 2011-2012 growing season (31 March 2012). Data collected on these dates best predicted durum wheat yield (discussed later). Plots of the first derivative of canopy spectral reflectance (Figs. 1b and 2b) demonstrated positive peaks at the red edge location (735 nm), highlighting the upward slope of reflectance between the red and near-infrared wavelengths. Additionally, negative first derivative peaks were associated with the decrease in reflectance data between 890 and 970 nm, due to the water absorption band centered at 970 nm. Plots of the second derivative of canopy spectral reflectance (Figs. 1c and 2c) demonstrated positive and negative peaks at approximately 720 and 750 nm, respectively, corresponding to wavelengths of maximum curvature in the reflectance data. Second derivative data were often noisy above 900 nm, due to magnification of noise in the original reflectance data through the differentiation process. These data remained in the analysis, because the

397 PLSR and genetic algorithm analysis methods could potentially reject noise.

398 In the 2010-2011 season, average measured LAI, canopy weight, and plant
399 N content on each measurement date ranged from 0.16 to 2.85 cm² cm⁻²,
400 0.11 to 7.33 Mg ha⁻¹, and 0.81% to 3.47%, respectively (Table 2). Average
401 measured yield was 4.75 Mg ha⁻¹ in 2010-2011, and average grain N concen-
402 tration was 2.23%. In the 2011-2012 season, average measured LAI, canopy
403 weight, and plant N content ranged from 0.17 to 3.80 cm² cm⁻², 0.10 to 9.75
404 Mg ha⁻¹, and 1.12% to 4.50%, respectively. Average measured yield was 5.45
405 Mg ha⁻¹ in 2011-2012, and average grain N concentration was 2.20%. Hi-
406 erarchical linear mixed modeling revealed differences in plant measurements
407 among cultivars on several measurement dates ($p < 0.05$, Table 2). Early
408 season plant N content was different among cultivars in both growing sea-
409 sons. All plant measurements were different among N fertilization rates on all
410 dates, except measurements collected at 34 DAP in 2010-2011. Most of the
411 variability in the measured plant data was due to N fertilization rates, where
412 higher N rates generally led to higher plant measurements. The interaction
413 of cultivar and N fertilization rate was significant for early season plant N
414 content in both seasons years ($p < 0.05$). However, interaction results were
415 inconsistent among growing seasons for all other measurements. The results
416 highlight differences in plant measurements among experimental treatments,
417 which makes the data set useful for analysis of proximal hyperspectral data
418 to estimate durum wheat traits.

3.2. Trait estimation

Minimum RMSECV from PLSR demonstrated the ability of different spectral data sets to estimate LAI, canopy weight, and plant N content on four dates during the 2010-2011 and 2011-2012 growing seasons (Table 3). Estimates were poorest ($33.0\% \leq \text{RMSECV} \leq 67.6\%$) when using spectral data from four broad wavebands that mimicked the information content of common multispectral radiometers. Better estimates ($21.1\% \leq \text{RMSECV} \leq 39.1\%$) were obtained when using a spectral data set based on eight 9-nm wavebands in the VIS-NIR spectrum. However, in both growing seasons, neither of these broad-band spectral data sets could estimate LAI, canopy weight, or plant N content better than the contiguous narrow-band data set collected with the VIS-NIR spectroradiometer ($19.3\% \leq \text{RMSECV} \leq 36.3\%$). Thus, in spite of multicollinearity issues with the hyperspectral reflectance data, the information content for estimating wheat canopy traits was superior to the content typically obtained from an 8-band multispectral radiometer.

For full-spectrum data with 701 wavebands, the PLSR analysis demonstrated little advantage to using spectral derivative data as compared to the narrow-band reflectance data set, because RMSECV was generally lower for the latter. Two exceptions were the RMSECV for LAI and plant N content in the 2010-2011 growing season when combining narrow-band spectral reflectance with the first and second derivative data sets, which were slightly lower as compared to narrow-band reflectance alone. However, the PLSR results generally showed that the narrow-band data alone provided the most

information content to estimate durum wheat LAI, canopy weight, and plant N content. There was little advantage to use of full-spectrum derivative data, an important result considering several past efforts to develop derivative methods for spectral data analysis (Demetriades Shah et al., 1990; Thorp et al., 2004; Tsai and Philpot, 1998).

By combining a genetic algorithm with PLSR to extract relevant features from the spectral data set, the RMSECVs for estimates of LAI, canopy weight, and plant N content were substantially reduced ($15.1\% \leq \text{RMSECV} \leq 30.7\%$) as compared to estimates from PLSR alone ($19.3\% \leq \text{RMSECV} \leq 67.6\%$). As compared to PLSR estimates based on four broad spectral bands, PLSR with genetic algorithm on narrow-band data could often reduce the RMSECV by more than half. Thus, the wavebands selected by the genetic algorithm were much better at estimating durum wheat traits as compared to the wavebands selected for inclusion on two commercial broad-band multispectral sensors. The genetic algorithm was also able to select relevant features within the derivative spectra, which led to slightly better estimation of the three canopy traits as compared to the features selected from the narrow-band reflectance data. With the exception of plant N content in 2012, the second derivative features selected by the genetic algorithm could estimate canopy traits better than the selected first derivative features and reflectance features. However, the best overall RMSECV for the three canopy traits ($15.1\% \leq \text{RMSECV} \leq 28.4\%$) was obtained when using PLSR with the genetic algorithm to select features from the combination of narrow-band

465 reflectance and derivative data. The genetic algorithm permitted rejection of
466 spectral information that contributed little to plant trait estimation, thereby
467 allowing the development of PLSR models that incorporated spectral features
468 at the wavelengths of greatest importance.

469 The PLSR modeling results demonstrated the optimum time for estimat-
470 ing durum wheat yield from in-season canopy spectral reflectance data (Fig.
471 3). For both growing seasons, yield estimates improved rapidly with each
472 data collection outing until 70 days after planting, roughly the third week of
473 February during the wheat tillering phase. Thus, canopy spectral reflectance
474 differences that ultimately led to yield differences were apparent in the early
475 season after the first of four split-applied N fertilizer treatments, which oc-
476 curred in mid-January in each season (Table 1). Up to 85 days after planting
477 in both growing seasons, the RMSECVs for yield estimates were not different
478 for PLSR models based on 4 broad wavebands versus 701 narrow wavebands.
479 Thereafter, the narrow-band data demonstrated a clear advantage for yield
480 estimation compared to broad-band data. In both growing seasons, PLSR
481 models developed from 701 narrow wavebands best estimated durum wheat
482 yield at 125 days after planting, corresponding to the period of late anthesis
483 and early grain filling in mid-April. Using PLSR with the genetic algorithm,
484 yield was estimated substantially better than PLSR alone on all spectral
485 data collection dates. Furthermore, RMSECV for yield estimation was im-
486 proved by allowing the algorithm to select features from the combination of
487 narrow-band reflectance data and first and second derivative spectra. Over-

all, yield estimation was optimized using spectra collected at 118 days after
planting in 2010-2011 ($\text{RMSECV} = 6.0\%$) and at 111 days after planting in
2011-2012 ($\text{RMSECV} = 7.0\%$). For best estimation of durum wheat yield,
canopy spectral reflectance data should be collected at flowering or shortly
thereafter. The presence of wheat heads at this time may alter the canopy
spectral reflectance response to provide a direct mechanism for wheat yield
prediction.

To estimate grain N content, there was no advantage to using 701 narrow
spectral wavebands versus 4 broad bands until 124 days after planting in
2010-2011, whereas the the advantage was clear at 103 days after planting in
2011-2012 (Fig. 4). Using PLSR with the genetic algorithm, the RMSECV
for grain N content estimation was lower than PLSR alone on all spectral data
collection dates. Furthermore, incorporation of first and second derivative
spectra into the algorithm provided additional improvements in RMSECV
in both growing seasons. In 2011-2012, estimation of grain N content was
optimized at 139 days after planting for several of the spectral data sets and
analysis techniques (Fig. 4b). Because this corresponded to the end of the
grain filling period, the mechanism was likely related to remobilization of
nutrients from the plant tissue to the grain. For N limited treatments, plant
tissue was depleted of N reserves more quickly than well-fertilized treatments
(Table 2). Although the optimum time for grain N content estimation oc-
curred later in the 2011-2012 season, the RMSECVs were less than 10% for
other data collection dates using PLSR with the genetic algorithm, indicating

511 potential to use canopy spectral reflectance measurements in key wavebands
512 to predict grain N content at mid-season. Additionally in the 2010-2011
513 season (Fig. 4a), optimum dates for grain N content estimation were found
514 at 100 or 120 days after planting. Because the final two N fertilizer appli-
515 cations occurred at 98 and 116 days after planting in 2010-2011 (Table 1),
516 there is ample opportunity to use canopy spectral reflectance data for guiding
517 mid-season N fertilizer applications to achieve end-of-season goals for grain
518 N content and related protein content. The methodologies described herein
519 can assist in identifying appropriate spectral features for grain N prediction
520 at mid-season.

521 3.3. Spectral feature selection

522 Using PLSR with the genetic algorithm, the spectral reflectance features
523 identified for LAI estimation were predominantly in the NIR region in both
524 growing seasons (Fig. 5). For the 2010-2011 data, only 2 of 13 reflectance
525 features were identified in the visible wavelengths: one from 350 to 415 nm
526 and another from 500 to 631 nm. The remaining 11 features were identified
527 in the near-infrared region between 710 and 1050 nm. For the 2011-2012
528 data, several features were selected between 534 and 657 nm, indicating vis-
529 ible green and red light, in addition to many features in the near-infrared
530 region. The contrast of visible and near-infrared radiation mimics the pur-
531 pose of many vegetation indices designed to estimate LAI. First derivative
532 spectra over broad visible light wavebands, many greater than 100 nm in
533 width, were useful for LAI estimation in both seasons, likely highlighting

534 the difference in slope between spectral reflectance of canopies with low and
535 high LAI at these wavelengths. Also, the genetic algorithm eliminated all
536 first derivative features between 869 and 917 nm in both years, highlighting
537 a lack of useful information at these wavelengths. First derivative features
538 resulting from the water absorption band at 970 nm were also important for
539 LAI estimation. Second derivative spectral features for LAI estimation were
540 scattered throughout the VIS/NIR spectrum from 400 to 1000 nm. Several
541 narrow second derivative features associated with curvature due to the red
542 edge from 698 to 753 nm were identified.

543 Spectral features for estimation of canopy weight (Fig. 6) were often
544 grouped at specific wavebands, highlighting a main advantage of the genetic
545 algorithm to identify spectral regions of importance. For canopy spectral
546 reflectance data, several wavebands between 459 and 601 nm were identified
547 for canopy weight estimation in both growing seasons. These wavelengths
548 are associated with the transition between visible blue and green light, where
549 chlorophyll absorption differentially affects healthy and stressed vegetation.
550 For 2010-2011 data, many reflectance features were identified between 770
551 and 912 nm, so NIR scattering aided canopy weight estimation in this season.
552 Reflectance features at the red edge from 672 to 817 nm were more prominent
553 for 2011-2012. First derivative features in the visible green region centered at
554 550 nm were identified for canopy weight estimation in both growing seasons,
555 indicating the importance of changes in slope of green light reflectance curves.
556 Also, the genetic algorithm identified several first derivative features between

557 739 and 856 nm, the transition from red edge to NIR. Optimum second
558 derivative features for canopy weight often covered broad wavebands, many
559 greater than 120 nm in width, which highlighted the importance of broad
560 changes in curvature from 400 to 900 nm.

561 The genetic algorithm selected relatively few features in the NIR region
562 (> 750 nm) for estimating plant N content, particularly in the 2010-2011
563 growing season (Fig. 7). Reflectance features were commonly selected be-
564 tween 583 and 722 nm in both growing seasons, which encompasses the pri-
565 mary wavelengths for visible red light absorption by chlorophyll. First and
566 second derivative features for plant N content were focused in the visible
567 wavelengths, also likely due to the effects of chlorophyll absorption on visible
568 light reflectance. However, first derivative features associated with the wa-
569 ter absorption band at 970 nm were consistently identified in both growing
570 seasons.

571 Spectral features for durum wheat yield estimation were analyzed at 124
572 days after planting in 2010-2011 and 130 days after planting in 2011-2012
573 (Fig. 8), which corresponded to the time of late flowering and early grain fill.
574 There were several similarities between the selected reflectance features in the
575 two growing seasons on these dates, including 1) narrow reflectance features
576 between 362 and 387 nm, 2) predominant selection of reflectance features in
577 the transition between blue and green light from 470 to 560 nm, and 3) several
578 NIR reflectance features grouped between 760 and 800 nm, 915 and 938 nm,
579 and 1018 and 1049 nm. Possible mechanisms to explain the yield prediction

580 capability from canopy spectral reflectance include changes in reflectivity due
581 to the presence of fully developed wheat heads with awns and also the effects
582 of remobilization of N and other nutrients from the vegetative components
583 to the grain. Spectral features identified for yield prediction at other growth
584 stages were less similar among the two growing seasons (not shown). Yield
585 can be affected by many dynamic processes that occur between the time of
586 spectral measurement and grain harvest, which complicates yield estimation
587 from spectral data. However, identification of consistent spectral features at
588 late flowering in two growing seasons (Fig. 8) combined with improved yield
589 estimation at that time (Fig. 3) could indicate a real mechanism for yield
590 estimation from spectral data. Future efforts should aim to better understand
591 the mechanism and fine-tune spectral techniques for wheat yield prediction
592 at flowering and early grain fill.

593 Spectral features for durum wheat grain N content were analyzed at 112
594 and 118 days after planting in the 2010-2011 and 2011-2012 growing seasons,
595 respectively (Fig. 9). These dates corresponded to the time of the final split
596 fertilizer application at the beginning of anthesis (Table 1). Thus, remote
597 sensing data collected at this time would provide the last opportunity for
598 assistance with N fertilizer management decisions. Many spectral reflectance
599 features were selected between 600 and 800 nm in both growing seasons,
600 which highlights the importance of the red edge region for estimation of grain
601 N content with mid-season canopy spectral reflectance. Similar to grain yield
602 estimation, a complicating factor for estimation of grain N content is that

many processes affect grain N outcomes following spectral measurements at mid-season. However, the results suggest that final grain N content can be estimated reasonably ($\text{RMSECV} < 10\%$, Fig. 4) by developing PLSR models from relevant spectral features at mid-season (Fig. 9). Future efforts should fine-tune spectral techniques for prediction of grain N content at mid-season and expand the analysis to include estimation of grain protein.

4. Conclusions

The study demonstrated clear advantages to using hyperspectral data from spectroradiometers to estimate wheat biophysical traits. The PLSR models based on full-spectrum, contiguous, narrow-band hyperspectral reflectance data provided better estimates of wheat traits than PLSR models using data from four or eight broad wavebands that mimicked two commercial multispectral radiometers. However, the hyperspectral data was also shown to contain excess data of little value for trait estimation, because a genetic algorithm was able to select less than 25 spectral features from 701 narrow wavebands, which led to substantial improvement in trait estimates. Thus, full-spectrum, contiguous, narrow-band reflectance data is not necessarily better than non-contiguous, broad-band data. Rather, the former can be used to determine the optimum composition of the latter for improved estimation of crop traits. Future efforts should focus on fine-tuning methodologies, via genetic algorithm or otherwise, to prune hyperspectral data sets of irrelevant information prior to PLSR model development.

625 Derivative spectra calculated from hyperspectral reflectance data was ad-
626 vantageous to trait estimation, but only when the genetic algorithm was used
627 to extract relevant spectral derivative features from the full-spectrum data.
628 Explicit comparisons of the entire spectral derivative data set with its corre-
629 sponding narrow-band reflectance data demonstrated little advantage to the
630 former. If derivative spectra is deemed necessary for a given sensing applica-
631 tion, a hyperspectral sensor must be deployed to produce the narrow-band,
632 contiguous reflectance data needed for derivative calculations. Otherwise,
633 data from hyperspectral instruments are mainly useful for exploratory anal-
634 yses with a goal to design simpler radiometers for crop reflectance sensing
635 applications in specific wavebands. As discussed herein, one approach uses
636 a genetic algorithm to improve crop trait estimation by extracting relevant
637 information from full-spectrum data. Future efforts can use this approach to
638 finalize sensor designs or select band-pass filters for improved estimation of
639 specific crop traits.

640 Durum wheat yield and grain N content were estimated from mid-season
641 canopy spectral reflectance data with RMSECV less than 9.4%. Future ef-
642 forts will apply the present findings to improve current proximal sensing
643 techniques for N fertilizer management in durum wheat. With further de-
644 velopment, these techniques will be useful for making mid-season N fertilizer
645 management decisions, optimizing durum wheat grain protein content for
646 maximum grower profit, and minimizing N losses to the environment.

5. Acknowledgments

The authors acknowledge the USDA-ARS-ALARC technicians (Suzette Maneely and Sharette Rockholt), Maricopa Agricultural Center technicians (Ken Randolph and Cory Runyon), University of Arizona graduate student (Ruth Asiimwe), and the day crews for their help with field and laboratory work. The authors also acknowledge the Arizona Grain Research and Promotion Council for contributing funds for this research.

6. References

- Aparicio, N., Villegas, D., Araus, J. L., Casadess, J., Royo, C., 2002. Relationship between growth traits and spectral vegetation indices in durum wheat. *Crop Science* 42 (5), 1547–1555.
- Araus, J. L., Cairns, J. E., 2014. Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science* 19 (1), 52–61.
- Bannari, A., Morin, D., Bonn, F., Huete, A. R., 1995. A review of vegetation indices. *Remote Sensing Reviews* 13 (1-2), 95–120.
- Chen, P., Haboudane, D., Tremblay, N., Wang, J., Vigneault, P., Li, B., 2010. New spectral indicator assessing the efficiency of crop nitrogen treatment in corn and wheat. *Remote Sensing of Environment* 114 (9), 1987–1997.
- Demetriades Shah, T. H., Steven, M. D., Clark, J. A., 1990. High resolution derivative spectra in remote sensing. *Remote Sensing of Environment* 33, 55–64.

-
- 668 Ecartot, M., Compan, F., Roumet, P., 2013. Assessing leaf nitrogen content
669 and leaf mass per unit area of wheat in the field throughout plant cycle
670 with a portable spectrometer. *Field Crops Research* 140, 44–50.
- 671 Feng, W., Guo, B. B., Wang, Z. J., He, L., Song, X., Wang, Y. H., Guo,
672 T. C., 2014. Measuring leaf nitrogen concentration in winter wheat using
673 double-peak spectral reflection remote sensing data. *Field Crops Research*
674 159, 43–52.
- 675 Feng, W., Yao, X., Zhu, Y., Tian, Y. C., Cao, W. X., 2008. Monitoring leaf
676 nitrogen status with hyperspectral reflectance in wheat. *European Journal*
677 *of Agronomy* 28 (3), 394–404.
- 678 Fitzgerald, G., Rodriguez, D., O’Leary, G., 2010. Measuring and predict-
679 ing canopy nitrogen nutrition in wheat using a spectral index-The canopy
680 chlorophyll content index (CCCI). *Field Crops Research* 116 (3), 318–324.
- 681 Fu, Y., Yang, G., Wang, J., Song, X., Feng, H., 2014. Winter wheat biomass
682 estimation based on spectral indices, band depth analysis and partial
683 least squares regression using hyperspectral measurements. *Computers and*
684 *Electronics in Agriculture* 100, 51–59.
- 685 Haboudane, D., Miller, J. R., Pattey, E., Zarco Tejada, P. J., Strachan, I. B.,
686 2004. Hyperspectral vegetation indices and novel algorithms for predicting
687 green LAI of crop canopies: Modeling and validation in the context of
688 precision agriculture. *Remote Sensing of Environment* 90 (3), 337–352.

-
- 689 Hansen, P. M., Schjoerring, J. K., 2003. Reflectance measurement of canopy
690 biomass and nitrogen status in wheat crops using normalized difference
691 vegetation indices and partial least squares regression. *Remote Sensing of*
692 *Environment* 86 (4), 542–553.
- 693 Horler, D. N. H., Dockray, M., Barber, J., 1983. The red edge of plant leaf
694 reflectance. *International Journal of Remote Sensing* 4 (2), 273–288.
- 695 Kaleita, A. L., Steward, B. L., Ewing, R. P., Ashlock, D. A., Westgate, M. E.,
696 Hatfield, J. L., Nov. 2006. Novel analysis of hyperspectral reflectance data
697 for detecting onset of pollen shed in maize. *Transactions of the ASABE*
698 49 (6), 1947–1954.
- 699 Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection
700 in spectral data sets. *Journal of Chemometrics* 14 (5-6), 643–655.
- 701 Li, C. J., Wang, J. H., Wang, Q., Wang, D. C., Song, X. Y., Wang, Y.,
702 Huang, W. J., 2012. Estimating wheat grain protein content using multi-
703 temporal remote sensing data based on partial least squares regression.
704 *Journal of Integrative Agriculture* 11 (9), 1445–1452.
- 705 Liang, Z., Bronson, K. F., Thorp, K. R., Mon, J., Badaruddin, M., Wang,
706 G., 2014. Cultivar and N fertilizer rate affect yield and N use efficiency in
707 irrigated durum wheat. *Crop Science* 54 (3), 1175–1183.

-
- 708 Madden, H. D., 1978. Comments on the Savitzky-Golay convolution method
709 for least-squares fit smoothing and differentiation of digital data. Analyti-
710 cal Chemistry 50 (9), 1383–1386.
- 711 Mahajan, G. R., Sahoo, R. N., Pandey, R. N., Gupta, V. K., Kumar, D.,
712 2014. Using hyperspectral remote sensing techniques to monitor nitrogen,
713 phosphorus, sulphur and potassium in wheat (*Triticum aestivum* L.). Pre-
714 cision Agriculture 15, 499–522.
- 715 Mevik, B. H., Wehrens, R., 2007. The pls package: principle component and
716 partial least squares regression in R. Journal of Statistical Software 18 (2),
717 1–24.
- 718 Ottman, M. J., Doerge, T. A., Martin, E. C., 2000. Durum grain quality as
719 affected by nitrogen fertilization near anthesis and irrigation during grain
720 fill. Agronomy Journal 92 (5), 1035–1041.
- 721 Raun, W. R., Solie, J. B., Taylor, R. K., Arnall, D. B., Mack, C. J., Edmonds,
722 D. E., 2008. Ramp calibration strip technology for determining midseason
723 nitrogen rates in corn and wheat. Agronomy Journal 100 (4), 1088–1093.
- 724 Savitzky, A., Golay, M. J. E., 1964. Smoothing and differentiation of data
725 by simplified least square procedures. Analytical Chemistry 36 (8), 1627–
726 1639.

-
- 727 Serrano, L., Filella, I., Peuelas, J., 2000. Remote sensing of biomass and yield
728 of winter wheat under different nitrogen supplies. *Crop Science* 40 (3), 723–
729 731.
- 730 Thorp, K. R., Dierig, D. A., French, A. N., Hunsaker, D. J., 2011. Analysis
731 of hyperspectral reflectance data for monitoring growth and development
732 of lesquerella. *Industrial Crops and Products* 33 (2), 524–531.
- 733 Thorp, K. R., Gore, M. A., Andrade Sanchez, P., Carmo Silva, A. E., Welch,
734 S. M., White, J. W., French, A. N., 2015. Proximal hyperspectral sensing
735 and data analysis approaches for field-based plant phenomics. *Computers
736 and Electronics in Agriculture* 118, 225–236.
- 737 Thorp, K. R., Tian, L., Yao, H., Tang, L., Jan. 2004. Narrow-band and
738 derivative-based vegetation indices for hyperspectral data. *Transactions of
739 the ASAE* 47 (1), 291–299.
- 740 Thorp, K. R., Wang, G., West, A. L., Moran, M. S., Bronson, K. F., White,
741 J. W., Mon, J., 2012. Estimating crop biophysical properties from remote
742 sensing data by inverting linked radiative transfer and ecophysiological
743 models. *Remote Sensing of Environment* 124, 224–233.
- 744 Tilling, A. K., O’Leary, G. J., Ferwerda, J. G., Jones, S. D., Fitzgerald,
745 G. J., Rodriguez, D., Belford, R., 2007. Remote sensing of nitrogen and
746 water stress in wheat. *Field Crops Research* 104 (1-3), 77–85.

-
- 747 Tsai, F., Philpot, W., 1998. Derivative analysis of hyperspectral data. Re-
748 mote Sensing of Environment 66 (1), 41–51.
- 749 White, J. W., Andrade Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt,
750 T. A., Conley, M. M., Feldmann, K. A., French, A. N., Heun, J. T.,
751 Hunsaker, D. J., Jenks, M. A., Kimball, B. A., Roth, R. L., Strand, R. J.,
752 Thorp, K. R., Wall, G. W., Wang, G., 2012. Field-based phenomics for
753 plant genetics research. Field Crops Research 133, 101–112.
- 754 Xue, L. H., Cao, W. X., Yang, L. Z., 2007. Predicting grain yield and pro-
755 tein content in winter wheat at different N supply levels using canopy
756 reflectance spectra. Pedosphere 17 (5), 646–653.
- 757 Yao, H., Tian, L., 2003. A genetic-algorithm-based selective principal com-
758 ponent analysis (GA-SPCA) method for high-dimensional data feature
759 extraction. IEEE Transactions on Geoscience and Remote Sensing 41 (6
760 PART I), 1469–1478.

Table 1: Split-applied nitrogen (N) fertilizer rates at different durum wheat growth stages in the 2010-2011 and 2011-2012 growing seasons at Maricopa, Arizona, USA.

Growth Stage	Application Date		N Rate (kg N ha ⁻¹)					
	Season 1	Season 2	0	72	124	186	269	403 ¹
Preplant	N/A	08 Dec 2011	0	0	0	0	0	90
Feekes 1-2	18 Jan 2011	11 Jan 2012	0	17	34	62	90	112
Feekes 5	09 Mar 2011	28 Feb 2012	0	11	22	34	45	56
Feekes 10	24 Mar 2011	13 Mar 2012	0	22	34	45	67	67
Feekes 10.5	11 Apr 2011	09 Apr 2012	0	22	34	45	67	78

¹ The 403 kg N ha⁻¹ rate was used in the 2011-2012 growing season only.

Table 2: Chi squared (χ^2) statistics and probability (p) values from hierarchical linear mixed modeling demonstrate the effects of cultivar, nitrogen (N) fertilization rate, and their interaction on leaf area index (LAI, $\text{cm}^2 \text{ cm}^{-2}$), canopy weight (CWT, Mg ha^{-1}), and plant N content (PNC, %) on four dates after planting (DAP) in 2010-2011 (top section) and 2011-2012 (bottom section) and final grain yield (YLD, Mg ha^{-1}) and grain N content (GNC, %) in each season. Significance codes are “****” ($p < 0.001$), “***” ($p < 0.01$), and “*” ($p < 0.05$).

				Cultivar			N Rate			Interaction		
	DAP	Mean	StDev	χ^2	p		χ^2	p		χ^2	p	
LAI	34	0.16	0.05	10.4	0.0644		4.7	0.3146		16.1	0.2449	
CWT	34	0.11	0.03	9.5	0.0907		4.8	0.3119		21.5	0.0639	
PNC	34	3.47	0.81	12.6	0.0277	*	2.3	0.6720		36.6	0.0005	***
LAI	71	1.30	0.64	6.7	0.2452		105.3	0.0000	***	18.2	0.5756	
CWT	71	1.21	0.55	1.5	0.9072		99.6	0.0000	***	21.1	0.3890	
PNC	71	2.47	0.56	10.3	0.0682		136.1	0.0000	***	25.1	0.1975	
LAI	97	2.85	1.50	8.8	0.1180		187.2	0.0000	***	24.1	0.2361	
CWT	97	4.38	1.62	5.6	0.3482		195.5	0.0000	***	33.1	0.0332	*
PNC	97	1.25	0.43	4.2	0.5197		12.5	0.0137	*	20.4	0.4345	
LAI	113	2.08	1.13	20.8	0.0009	***	218.2	0.0000	***	20.8	0.4068	
CWT	113	7.33	2.68	8.7	0.1208		220.4	0.0000	***	12.2	0.9095	
PNC	113	0.81	0.36	10.7	0.0580		40.8	0.0000	***	27.8	0.1146	
YLD	178	4.75	2.09	2.4	0.7949		268.5	0.0000	***	20.3	0.4407	
GNC	178	2.23	0.40	8.1	0.1488		106.0	0.0000	***	21.4	0.3762	
LAI	32	0.17	0.05	29.5	0.0000	***	40.9	0.0000	***	51.6	0.0014	**
CWT	32	0.10	0.03	26.9	0.0001	***	53.2	0.0000	***	48.6	0.0032	**
PNC	32	4.50	0.54	15.0	0.0103	*	135.9	0.0000	***	43.0	0.0139	*
LAI	69	2.25	1.00	12.1	0.0341	*	174.5	0.0000	***	32.5	0.1436	
CWT	69	1.58	0.62	5.4	0.3635		157.4	0.0000	***	24.1	0.5161	
PNC	69	2.92	0.70	39.9	0.0000	***	185.0	0.0000	***	38.2	0.0441	*
LAI	95	3.80	2.11	4.6	0.4720		183.8	0.0000	***	21.4	0.6729	
CWT	95	5.22	2.17	7.5	0.1848		147.5	0.0000	***	26.2	0.3951	
PNC	95	1.73	0.43	14.5	0.0129	*	117.2	0.0000	***	25.7	0.4231	
LAI	117	2.94	1.86	7.4	0.1900		226.6	0.0000	***	43.1	0.0138	*
CWT	117	9.75	3.67	7.7	0.1721		166.3	0.0000	***	24.1	0.5159	
PNC	117	1.12	0.41	5.1	0.4093		43.6	0.0000	***	26.4	0.3844	
YLD	167	5.45	2.44	28.7	0.0000	***	462.4	0.0000	***	35.9	0.2121	
GNC	167	2.20	0.30	28.4	0.0000	***	174.5	0.0000	***	63.0	0.0004	***

Table 3: Minimum root mean squared error of cross validation (RMSECV, %) from partial least squares regression (PLSR) and PLSR combined with a genetic algorithm (GA) used to estimate leaf area index (LAI), canopy weight (CWT) and plant nitrogen content (PNC) on four dates after planting in the 2010-2011 and 2011-2012 durum wheat growing seasons. The results demonstrate ability to estimate plant growth metrics from different spectral data sets, each containing n variables derived from narrow-band canopy spectral reflectance as measured with a field spectroradiometer.¹

Method	Data Set	n	2011 LAI	2011 CWT	2011 PNC	2012 LAI	2012 CWT	2012 PNC
PLSR	BBR	4	40.5	43.4	44.1	43.0	67.6	33.0
PLSR	BBR	8	37.8	26.0	28.5	39.1	33.1	21.1
PLSR	NBR	701	34.2	24.8	28.4	36.3	31.8	19.3
PLSR	FDR	701	35.2	25.2	28.6	38.9	32.7	19.4
PLSR	SDR	701	34.8	29.1	31.6	39.6	44.3	23.2
PLSR	NBR,FDR,SDR	2103	33.9	25.1	28.2	39.0	33.6	19.8
PLSR-GA	NBR	<25	29.5	21.3	23.0	30.7	27.6	16.3
PLSR-GA	FDR	<25	28.5	22.0	23.0	30.1	28.4	16.2
PLSR-GA	SDR	<25	26.3	20.8	21.4	30.0	27.6	16.4
PLSR-GA	NBR,FDR,SDR	<25	26.1	20.3	21.2	28.4	26.0	15.1

¹ broad-band reflectance, BBR; canopy weight, CWT; first derivative of narrow-band reflectance, FDR; genetic algorithm, GA; leaf area index, LAI; narrow-band reflectance, NBR; partial least squares regression, PLSR; plant nitrogen content, PNC; second derivative of narrow-band reflectance, SDR

Figure 1: Spectral measurements of durum wheat on 118 days after planting in the 2010-2011 growing season (April 12, 2011), including a) canopy spectral reflectance, b) the first derivative of canopy spectral reflectance (FDR), and c) the second derivative of canopy spectral reflectance (SDR).

Figure 2: Spectral measurements of durum wheat on 111 days after planting in the 2011-2012 growing season (March 31, 2012), including a) canopy spectral reflectance, b) the first derivative of canopy spectral reflectance (FDR), and c) the second derivative of canopy spectral reflectance (SDR).

Figure 3: Minimum root mean squared error of cross validation (RMSECV) from partial least squares regression (PLSR) used to estimate durum wheat grain yield from canopy reflectance measured on a) 19 dates after planting in 2010-2011 and b) 17 dates after planting in 2011-2012. Yield estimates were based on PLSR with four broad-band reflectance values (PLSR1), PLSR with 701 narrow-band reflectance values (PLSR2), and PLSR combined with a genetic algorithm to identify relevant spectral reflectance features from narrow-band reflectance data (PLSR-GA1) and narrow-band reflectance combined with first and second spectral derivative data (PLSR-GA2).

Figure 4: Minimum root mean squared error of cross validation (RMSECV) from partial least squares regression (PLSR) used to estimate durum wheat grain nitrogen (N) content from canopy reflectance measured on a) 19 dates after planting in 2010-2011 and b) 17 dates after planting in 2011-2012. Grain N content estimates were based on PLSR with four broad-band reflectance values (PLSR1), PLSR with 701 narrow-band reflectance values (PLSR2), and PLSR combined with a genetic algorithm to identify relevant spectral reflectance features from narrow-band reflectance data (PLSR-GA1) and narrow-band reflectance combined with first and second spectral derivative data (PLSR-GA2).

Figure 5: Wavebands of relevant spectral features identified using a genetic algorithm with partial least squares regression to estimate durum wheat leaf area index from canopy spectral reflectance (REF), the first derivative of canopy spectral reflectance (FDR), and the second derivative of canopy spectral reflectance (SDR) during the a) 2010-2011 and b) 2011-2012 growing seasons. A plot of the average canopy reflectance spectra (red dashed curved) is provided for reference.

Figure 6: Wavebands of relevant spectral features identified using a genetic algorithm with partial least squares regression to estimate durum wheat canopy weight from canopy spectral reflectance (REF), the first derivative of canopy spectral reflectance (FDR), and the second derivative of canopy spectral reflectance (SDR) during the a) 2010-2011 and b) 2011-2012 growing seasons. A plot of the average canopy reflectance spectra (red dashed curved) is provided for reference.

Figure 7: Wavebands of relevant spectral features identified using a genetic algorithm with partial least squares regression to estimate durum wheat plant nitrogen content from canopy spectral reflectance (REF), the first derivative of canopy spectral reflectance (FDR), and the second derivative of canopy spectral reflectance (SDR) during the a) 2010-2011 and b) 2011-2012 growing seasons. A plot of the average canopy reflectance spectra (red dashed curved) is provided for reference.

Figure 8: Wavebands of relevant spectral features identified using a genetic algorithm with partial least squares regression to estimate durum wheat yield from canopy spectral reflectance (REF), the first derivative of canopy spectral reflectance (FDR), and the second derivative of canopy spectral reflectance (SDR) collected during the a) 2010-2011 growing season at 124 days after planting and the b) 2011-2012 growing season at 130 days after planting. A plot of the average canopy reflectance spectra (red dashed curved) is provided for reference.

Figure 9: Wavebands of relevant spectral features identified using a genetic algorithm with partial least squares regression to estimate durum wheat grain nitrogen content from canopy spectral reflectance (REF), the first derivative of canopy spectral reflectance (FDR), and the second derivative of canopy spectral reflectance (SDR) collected during the a) 2010-2011 growing season at 112 days after planting and the b) 2011-2012 growing season at 118 days after planting. A plot of the average canopy reflectance spectra (red dashed curved) is provided for reference.