Application of hyperspectral imaging to discriminate waxy corn seed vigour after aging

1Yuan, P., 2Pang, L., 1Wang, L. M. and 1*Yan, L.

1School of Technology, Beijing Forestry University, Beijing, 100083, China
2Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, Beijing, 100191, China

Abstract
A hyperspectral imaging system covering 400 - 1000 nm spectral range was applied for vigour detection of waxy maize seeds after artificial aging. After spectral pre-processing, the characteristic wavelength was selected by uninformative variable elimination (UVE), competitive adaptive reweighted sampling (CARS), and random frog (RF) methods. The moisture, starch, protein, and fat contents were measured for each grade of seed, and these values were correlated with the spectrum. Finally, the vitality detection model was established by least squares support vector machine (LS-SVM), extreme learning machine (ELM), and random forest (RF). The prediction sets exhibited high classification accuracy (> 99%) for 115 features. The model constructed from the bands significantly correlated with chemical composition (CC), and was better than the classic feature selection methods. The overall results indicated that hyperspectral imaging could be a potential technique to assess seed vigour.

Keywords
hyperspectral imaging, seed vigour, component detection, correlation analysis, discriminant model

Introduction

Waxy corn (Zea mays L. sinensis Kulesh) originated in China as a mutation of hard corn. This crop has high economic, nutritional, and processing values, and has become very popular in the corn industry. Waxy corn kernels have higher nutrient content than that in common corn, with 70 - 75% starch, more than 10% protein, and 4 - 5% fat. During storage, seeds undergo physiological and physiochemical changes including deterioration of seed chemistry (Kapoor et al., 2011; Eisvand et al., 2016). This natural aging is a long-term process, and aging study requires extensive sampling. As an alternative, artificially accelerated aging can be performed in the laboratory to simulate seed aging to facilitate and expedite the study (Men et al., 2017).

Like in traditional corn, the seed vigour of waxy corn is directly related to its seed germination performance and emergence. Conventional methods used to estimate seed viability include germination, tetrazolium (TZ), and electric conductivity tests (Kandpal et al., 2016). However, these traditional methods are time-consuming, laborious, and seed-damaging. Fourier near infrared (Altameme et al., 2015; Qiu et al., 2019), near-infrared (Rodríguez-Pulido et al., 2014; Jia et al., 2015), and Raman spectroscopy methods (Lee et al., 2017) are newer methods that can be applied to the study of seeds. However, these approaches typically do not obtain sufficient information by using point-based scanning techniques.

Hyperspectral imaging technology is a new non-destructive testing method that combines imaging and spectral data (Xia et al., 2019). In recent years, this approach for seed quality evaluation has received extensive attention and application in various fields, especially agricultural and forestry. This technique has been applied to seed type differentiation (Wang et al., 2016; Zhao et al., 2018), internal main component detection (Yang et al., 2018), origin (Gao et al., 2013), storage time assessment (Guo et al., 2017), seed infection detection (Kimuli et al., 2018; Polder et al., 2019), and seed vigour prediction (Ambroise et al., 2016; Wakhlo et al., 2018). Zhang et al. (2018) compared partial least squares discriminant analysis (PLS-DA) and support vector machines (SVM) models, finding that the PLS-DA model had high classification accuracy for whole wheat seeds (> 85.2%) and viable

© All Rights Reserved
seeds (> 89.5%) by using only 16 wavebands. Snider et al. (2016) studied the effects of cotton seed chemical composition on seedling vigour, concluding that seed quality and also oil and protein content positively correlated with seedling vigour. Cai et al. (2013) found a positive correlation of soluble protein with seed vigour. These applications of hyperspectral imaging technology successfully predicted components and vitality of different seeds, but further work is required to correlate spectral values and main components for vitality assessment.

For the identification of waxy corn seed vigour using hyperspectral imaging, the goal of the present work was to perform component-related feature band selection to improve the model discriminative ability. The specific steps of this approach were: (1) to obtain the hyperspectral information of different varieties of glutinous maize seeds subjected to different lengths of aging duration, and then measure seed vigour by standard germination experiments; (2) to analyse the correlation of main chemical components and spectra in seeds; (3) to perform spectral pre-processing and select characteristic bands to establish a classification model; and (4) to assess the ability to predict the level of vigour of different varieties of waxy corn seeds.

Materials and methods

Sample preparation


table 1. Major components (g/100 g) of eight groups of waxy corn seeds.

<table>
<thead>
<tr>
<th>Component</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moisture</td>
<td>9.06</td>
<td>8.29</td>
<td>7.93</td>
<td>7.81</td>
<td>9.7</td>
<td>8.68</td>
<td>8.1</td>
<td>7.77</td>
</tr>
<tr>
<td>Starch</td>
<td>44.3</td>
<td>26.3</td>
<td>45.6</td>
<td>29.2</td>
<td>65.6</td>
<td>36</td>
<td>34.7</td>
<td>43.3</td>
</tr>
<tr>
<td>Protein</td>
<td>10.8</td>
<td>9.7</td>
<td>9.5</td>
<td>9</td>
<td>11.1</td>
<td>10.3</td>
<td>10.6</td>
<td>10</td>
</tr>
<tr>
<td>Fat</td>
<td>4.1</td>
<td>3.7</td>
<td>4.5</td>
<td>3.6</td>
<td>4.2</td>
<td>4.3</td>
<td>4.6</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Hyperspectral image acquisition and correction

The hyperspectral image system included an SOC710 imaging spectrometer with a built-in C-type infrared correction lens (Surface Optics Corporation, USA), two 250 W halogen lamps (OSRAM GCA, Guangdong, China), tilted to 45° and placed symmetrically, a transport platform, and a computer terminal. This instrument has a spectral range of 400 - 1000 nm, a spectral resolution of 4.6875 nm (115 bands in total), and collects an image of 520 × 696 pixels.

For imaging, 288 samples were randomly selected from the remaining 200 g samples of each treatment class, and hyperspectral data were collected. The waxy corn seeds were set in a Petri dish (9 cm Ø), with 22 cm distance between the container and imaging spectrometer. After collecting hyperspectral images, the original images (I₀) were calibrated to remove noise. The black reference (I₀) was acquired when the camera lens was covered, and the white reference (I₀) was collected using a white Teflon board with approximately 100% reflectance. The corrected image (I) was calculated using Eq. 1:

\[ I = \frac{I_W - I_B}{I_W - I_B} \]  

(Eq. 1)
**Standard germination test**

Waxy corn seeds with hyperspectral information were subjected to standard germination test. Samples were placed in a separate Petri dishes with filter paper that was soaked with water to retain moisture. Following 7-d incubation at 25°C with 99% relative humidity and continuous light in the incubator, the germination was measured. According to the International Seed Testing Association standard (ISTA, 2015), if the germ of the seed is larger than 1 cm, it is considered a viable seed. The germination rates of the eight categories were 83.3, 95.8, 91.7, 54.2, 89.6, 93.8, 75.0, and 47.9% (corresponding to groups named A, B, C, D, E, F, G, and H), respectively. Short-term high temperature treatment can promote the improvement of seed vigour, consistent with Yan et al. (2014) study of oak (Quercus liaotungensis) seeds. The germination rate and average root length both initially increased and then decreased with the increase of aging duration.

**Spectral extraction and pre-processing**

The selection of an appropriate region of interest (ROI) is critical because it affects the extraction of the spectral data. Adaptive threshold segmentation is widely used due to its simplicity and efficiency, and was used for ROI selection in the present work (Karasulu and Korukoglu, 2010). Spectral extraction was done in four steps. The reflected image at 743.79 nm was chosen as this yielded the highest contrast between the seed and background (step I). Next, the binary image was established based on the grayscale image, and then, filtering was performed to enhance the binary image and eliminate noise (step II). The complete single area of each seed was set as the ROI (step III). Finally, the average spectrum of seeds was calculated from the average of the intensity values of all pixels in the ROI for each seed in 115 bands (step IV).

Multiplicative scatter correction (MSC) can effectively eliminate the effect of scattering, and enhance the spectral absorption information related to different components. Savitzky-Golay (SG) first derivative pre-processing is used to separate small absorption peaks, and improve spectral resolution. In the present work, MSC and SG first derivative methods were combined to remove the noise from the spectrum data and improve the prediction ability.

**Optimal wavelength selection**

Hyperspectral data are typically large, but contain a lot of redundant information, thus requiring extensive computing time and storage space for data processing. Uninformative variable elimination (UVE) (Cai et al., 2008), a commonly used method for variable selection, aims to select the best combination of important variables while removing useless information to improve verification performance. This method uses the partial least squares (PLS) regression coefficient as an indicator of wavelength importance.

Competitive adaptive reweighted sampling (CARS) (Li et al., 2009) is a variable selection method based on the principle of ‘survival of the fittest’ in Darwinian evolution theory. This method combines the exponentially decreasing function (EDF) and adaptive reweighted sampling technique (ARS) to select the variable point with the larger absolute value of the regression coefficient in the PLS model, and remove the variable point with the smaller weight value (Wang et al., 2017). Cross-validation can then be used to select the smallest root mean square error calculated from the cross-validation (RMSECV) subset in the PLS subset model. The variables contained in this subset are the optimal combination of variables.

Random frog (RF) (Bao et al., 2019) calculates the probability of being selected for each variable by simulating a Markov chain based on a steady state distribution. Whether the variable is selected or eliminated during each iteration is based on the size of the absolute value of each variable on the regression coefficient curve in the returned results.

**Recognition model development**

The least squares support vector machine (LS-SVM) starts from the loss function of machine learning, and uses the second norm in the objective function of the optimisation problem. Additionally, the inequality constraints in standard support vector machine (SVM) algorithms are replaced by equality constraints. In this way, the LS-SVM method changes the optimisation problem to a set of linear equations, which increases calculation efficiency over that of SVM, and improves the accuracy and precision of processing non-linear signals (Su et al., 2015). Radial basis function (RBF) was chosen as the kernel function of LS-SVM.

Extreme learning machine (ELM) is a type of machine learning system or method based on a feedforward neural network. In this kind of system,
the weights of hidden layer nodes are given randomly or artificially, no updating is needed, and the learning process only calculates output weights (Huang et al., 2006). For single hidden layer neural networks, ELM can randomly initialise input weights and biases to get the corresponding output weights. The connection weights between the hidden layer and the output layer do not need to be adjusted iteratively, but instead are determined a single time by solving the system of equations. ELM is widely used in computer vision, bioinformatics, and in regression problems in some earth and environmental sciences (Huang et al., 2015).

By integrating multiple trees through ensemble learning, a random forest (RF) is formed with a decision tree as the basic unit. The construction method of each decision tree is as follows: (1) \( N \) represents the number of samples, and \( M \) represents the number of features; (2) the input \( m \) (\( m \) is far less than \( M \)) is used to determine the result of the previous node of the decision tree; (3) for a sample with replacement in \( N \) samples, repeating \( N \) times is performed to form a training set, and prediction with unsampled ones; and (4) for each node, \( m \) features are randomly selected to calculate the best splitting method. In this way, a complete tree is constructed. RF can effectively run on large data sets, process input samples with high-dimensional features, and obtain good results for the default value problem (Breiman, 2001).

**Results and discussion**

**Spectral features**

Figure 1 shows the similarity of the average spectra and the difference in reflectance for all tested grades of seeds. All spectra were pre-processed by MSC and SG first derivative methods. This treatment removed scattering effects, and made it easier to distinguish multiple peaks and valleys. Crests and troughs of the short-wave near-infrared region are often caused by stretching and frequency doubling of the bending vibration of the X-H (N-H, C-H, O-H) bonds of proteins (Schrieve et al., 1991). It can be seen from the figure that after the same aging treatment, the average spectra of the seeds of Jingkenuo 2000 and Zhongpintiannuo F1 had obvious aggregation. Among them, the control group was particularly obvious, which was well distinguished from other grades. The vitality curves for all samples after pre-processing exhibited similar trends over the entire band, with some overlap in spectra. Therefore, chemometric methods were necessary for further classification and identification.

![Figure 1. Average spectra of eight groups of waxy corn seeds.](image)

**Reference analysis of chemical composition**

The main components of seeds will be affected by genetic factors, as well as by external environmental factors such as climate and soil conditions, harvest time, maturity, processing, packaging, transportation, and storage. The main components of seeds subjected to different storage conditions may be the same, but there may be variation in content, thus resulting in changes in the spectrum. The hyperspectral study of seed vigour requires good correlations between the spectrum and the main components of seeds.

The water, starch, protein, and fat contents varied in different types of waxy corn with different aging durations. Table 2 shows the contents of each major storage substance per 100 g of sample. With increased aging duration, the moisture content of seeds in the same category showed a decreasing trend. The amount of the primary substance in corn seeds, starch, also decreased suddenly with aging. Only small changes in protein and fat contents were observed, with the same decreasing trend. This may reflect changes in different enzyme activities during aging (Zhu et al., 2018; Wang and Ju, 2019). Enzymes and other substances could together increase the total amounts of starch, protein, and lipids in seeds.

A primary goal of the present work was to correlate the amounts of different components and the spectral data (average spectrum) of the eight groups.
of seeds. The Pearson correlation coefficient method was used, and the correlation coefficient (R) and the test p value were determined. R was used to assess the correlation between the spectrum and the composition of a component, and the p value was used to assess significance. Characteristic bands of $R > 0.8$ and $p < 0.05$ were selected for the four components of each category of the two categories of seeds, as shown in Table 2. The repeated parts of these bands were removed for subsequent modelling, as is done in conventional feature selection methods. Therefore, a total of 44 bands were extracted as feature bands.

Table 2. Extraction of bands significantly related to the main storage components.

<table>
<thead>
<tr>
<th>Variety</th>
<th>Component</th>
<th>Number of bands</th>
<th>Wavelength (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moisture</td>
<td>3</td>
<td>632.8, 727.8, 786.5</td>
</tr>
<tr>
<td>Jingkenuo 2000</td>
<td>Starch</td>
<td>7</td>
<td>765.1, 770.5, 775.8, 797.2, 802.6, 845.6, 965.2</td>
</tr>
<tr>
<td></td>
<td>Protein</td>
<td>4</td>
<td>400.6, 727.8, 786.5</td>
</tr>
<tr>
<td></td>
<td>Fat</td>
<td>8</td>
<td>482.3, 518.3, 717.2, 722.5, 754.5, 759.8, 840.2, 851.0</td>
</tr>
<tr>
<td>Zhongpintiannuo F1</td>
<td>Moisture</td>
<td>10</td>
<td>400.6, 441.3, 446.4, 461.8, 528.7, 591.2, 905.2, 970.7, 976.2</td>
</tr>
<tr>
<td></td>
<td>Starch</td>
<td>8</td>
<td>523.5, 669.6, 674.9, 680.2, 749.1, 782.6, 878.1, 954.3</td>
</tr>
<tr>
<td></td>
<td>Protein</td>
<td>2</td>
<td>431.1, 446.4, 451.5, 461.8, 690.7, 754.5, 781.2, 802.6, 851.0, 867.2, 926.9, 932.4, 937.9</td>
</tr>
<tr>
<td></td>
<td>Fat</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

Feature band extraction by different methods

A total of 115 features were included in the modelling based on the raw data, but there may be some redundant information in these features. To select valid information and improve operation speed, UVE, CARS, and RF methods were used to extract the characteristic bands from the original bands. When using the UVE method for selection, the optimal number of major factors was set to eight. The best results were obtained when the upper and lower thresholds were $+68$ and $-68$, respectively, and the minimum RMSECV reached the minimum value of 0.1770. This gave 43 characteristic bands. The CARS algorithm reduced the number of variables from 115 to 50. Five cross-validations were performed with 300 Monte Carlo samples. When RF was applied, the maximum number of latent variables for cross-validation when using the automatic scaling method was eight, and 32 characteristic bands were finally selected.

The bands significantly correlated with the chemical composition (CC) based on these three methods were compared. Figure 2 shows the number of characteristic bands selected by each method, and the specific location of each band.

Figure 2. Comparison of feature bands selected by different feature selection methods.

The bands selected by CARS had obvious two-end differentiation, with more features in the 400 - 600 and 800 - 1000 nm ranges, but sparse signal in
the 600 - 800 nm range. UVE and RF patterns were relatively uniformly distributed over the entire band. In the 500 - 700 nm range, CC had an interval close to 100 nm, and its characteristic band was more concentrated in the near-infrared spectral range.

**Discriminant model based on different feature sets**

To compare the different classification methods to detect and classify the corn seed spectral data and increase reliability, LS-SVM, ELM, and RF methods were employed to establish discriminant models for different grade of waxy corn seeds. For modelling, 288 seeds in each of eight groups were divided into three equal portions, and then randomly build a single calibration set and prediction set at a ratio of 2:1. Combining them, the sample sizes of the prediction set and calibration set were 768 and 1,536, respectively, during each modelling. This was done three times, and the average was considered the final classification result.

Table 3 summarises the results of discriminant models under different models and different feature sets. With 115 features in the full band, both the LS-SVM and RF calibration sets could achieve 100% accuracy, with higher prediction set accuracy for the LS-SVM model (about two percentage points). All three modelling methods gave similar results with fewer features, though was done as good as the full-band result. When compared with the results of the full-band prediction set, the feature model results were 0.69 - 7.59% less accurate. This result differs from that of Kandpal et al. (2016) who used PLS-DA to establish a melon seed vigour model, and found that the model with characteristic band selection by VIP and SR was similar to or better than the full band model.

<table>
<thead>
<tr>
<th>Discriminant model</th>
<th>Number of characteristic band (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>115 (Full)</td>
</tr>
<tr>
<td>LS-SVM Cal.¹</td>
<td>100</td>
</tr>
<tr>
<td>LS-SVM Pre.²</td>
<td>99.22</td>
</tr>
<tr>
<td>ELM Cal.³</td>
<td>91.71</td>
</tr>
<tr>
<td>ELM Pre.⁴</td>
<td>86.93</td>
</tr>
<tr>
<td>RF Cal.⁵</td>
<td>100</td>
</tr>
<tr>
<td>RF Pre.⁶</td>
<td>97.31</td>
</tr>
</tbody>
</table>

¹full band characteristics, ²bands correlated with the main chemical composition, ³calibration set, and ⁴prediction set.

UVE, CARS, and RF are classic and reliable feature selection methods. In the SVM model of wheat variety classification, RF feature selection allowed the use of 50 bands for results that were similar to those obtained using 256 bands (Bao et al., 2019). These methods all performed well in the present work. For the LS-SVM model, the accuracy of the training set was 100%, and the accuracy of the prediction set remained stable at greater than 96%. Modelling with ELM and RF was less successful, with poor performance based on RF characteristic bands. This may be due to the small number of bands, and the relatively discrete distribution. CC may be the optimal feature selection method, with recognition capabilities superior to the other three feature selection methods. The first three classic methods only used spectral data to select the most representative band. The CC method required only 44 bands to achieve close to the full-band effect. These 44 bands were significantly related to the main storage substances in waxy corn seeds, and could fully reflect the role of chemical components on the spectrum. The results obtained in the present work will facilitate the development and use of spectral methods to study seed composition and vigour.

**Conclusion**

Hyperspectral imaging technology allowed for the successful determination of the vigour of waxy maize seeds subjected to different aging times. Correlation analysis was performed between the spectral data and the levels of water, protein, starch, and fat; and 44 bands with significant correlation were identified. UVE, CARS and RF were also used for feature selection. Among the discriminant models
established by LS-SVM, ELM, and RF, the LS-SVM was the best, with good stability under different characteristic bands. In these models, the accuracy recognition rate of eight grades of seeds was over 96%. For feature selection, the CC selected band was more conducive for model establishment. The spectrum exhibited a strong correlation with the main chemical substances in seeds. The results of the present work showed that hyperspectral technology could effectively detect seed vigour for waxy corn seeds of different species and vigour. In future studies, waxy maize seeds of different varieties and vigour levels should be tested. More substances related to seed vigour should be measured, and the correlation analysis of substances related to vigour should be carried out to establish the basis for use of hyperspectral imaging technology to detect seed vigour.

Acknowledgement

The present work was financially supported by National Natural Science Foundation of China (grant no.: 31770769).

References


