Estimation and classification of popping expansion capacity in popcorn breeding programs using NIR spectroscopy

Gustavo Hugo Ferreira de Oliveira[1], Seth C. Murray[2], Luis Carlos Cunha Júnior[3], Kássio Michell Gomes de Lima[4], Camilo de Lelis Medeiros de Morais[5], Gustavo Henrique de Almeida Teixeira[6], Gustavo Vitti Môro[6]

[1]Universidade Federal de Sergipe (UFS), Núcleo de Graduação de Agronomia. Nossa Senhora da Glória/Sergipe, Brazil. Rodovia Engenheiro Jorge Neto, km 3, Silos, CEP: 49.680-000. Nossa Senhora da Glória, Sergipe, Brazil

[2]Texas A&M University, Department of Soil and Crop Science, 2474 TAMU MS, College Station, 77843. College Station, Texas, USA

[3]Universidade Federal de Goiás (UFG), Escola de Agronomia (EA), Setor de Horticultura. Rodovia Goiânia Nova Veneza, km 0, Campus Samambaia. Caixa Postal 131, Goiânia – GO, Brazil. CEP: 74.690-900

[4]Universidade Federal do Rio Grande do Norte (UFRN), Instituto de Química, Química Biológica e Quimiometria, Avenida Senador Salgado Filho, n° 3000, Bairro de Lagoa Nova, CEP: 59.078-970, Natal, Rio Grande do Norte, Brazil

[5]University of Central Lancashire, School of Pharmacy and Biomedical Sciences, Preston, Lancashire, PR1 2HE, United Kingdom

[6]Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias (FCAV), Campus de Jaboticabal. Via de Acesso Prof. Paulo Donato Castellane s/n, CEP: 14884-900 – Brasil, Jaboticabal, São Paulo, Brazil

*****Corresponding author: gv.moro@unesp.br

Abstract

One of the most important quality traits in popcorn breeding programs is the popping expansion (PE) capacity of the kernel, which is the ratio of the volume of the popcorn to the weight of the kernel. In this study, we evaluated whether near infrared spectroscopy (NIR spectroscopy) could be used as a tool in popcorn breeding programs to routinely predict and/or discriminate popcorn genotypes on the basis of their PE. Three generations (F_1, F_2, and $F_{2:3}$) were developed in three planting seasons by manual cross-pollination and self-pollination. A total of 376 ears from the $F_{2:3}$ generation were selected, shelled, and subjected to phenotypic analysis. Genetic variability was observed in the F_2 and $F_{2:3}$ generations, and their average PE value was 31.5 ± 6.7 $mL.g^{-1}$. PE prediction models using partial least square (PLS) regression were developed, and the root mean square error of calibration (RMSEC) was 6.08 $mL.g^{-1}$, while the coefficient of determination (R_C^2) was 0.26. The model developed by principal component analysis with quadratic discriminant analysis (PCA-QDA) was the best for discriminating the kernels with low PE (≤ 30 $mL.g^{-1}$) from those with high PE (> 30 $mL.g^{-1}$) with an accuracy of 78%, sensitivity of 81.2%, and specificity of 72.2%. Although NIR spectroscopy appears to be a promising non-destructive method for assessing the PE of intact popcorn kernels for narrow breeding populations, greater variability and larger sample sizes would help improve the robustness of the predictive and classificatory models.

Keywords: *Zea mays* L., selection methods, multivariate analysis, discrimination, prediction

1. Introduction

One of the most important quality traits in popcorn breeding programs is the popping expansion (PE) capacity of the kernel, which is defined as the ratio of the volume of the popcorn to the weight of the kernel (Guimarães et al., 2000). Owing to targeted breeding and selection, the PE capacity of popcorn has significantly increased over the last few decades. Recent reports have demonstrated that the PE values of popcorn have approximately doubled in comparison to those of the older American (25 $mL.g^{-1}$) (Galvão et al., 2015) and Brazilian popcorn varieties (15 $mL.g^{-1}$) (Zinsly and Machado, 1987), and the PE values of the current popcorn breeds are approximately 30 $mL.g^{-1}$ (Amaral Junior et al., 2012; Oliveira et al., 2019).

Although popcorn quality has noticeably improved in other parts of the world, Brazilian farmers still rely on imported seeds, especially from the United States of America (USA), where the PE capacity of popcorn is superior (Sawazaki, 2011). Additionally, since popcorn is bought by weight and sold by volume, the PE capacity is a vital criterion in determining the commercial value of popcorn.

One of the most difficult processes in popcorn breeding programs is phenotyping the high numbers of genotypes for assessing their PE capacity. In addition to the time required for selecting the genotypes with high PE, phenotyping destroys the kernels, as it requires heating the kernels in a microwave oven for rupturing the pericarp. This is a widely accepted procedure for determining the PE capacity of popcorn (Galvão et al., 2000). However, this method is destructive, requires large quantities of grains, and the crosses between the different genotypes in popcorn breeding programs produce a small number of kernels. Altogether, these methodological disadvantages challenge the progress of popcorn breeding programs towards maximizing genetic

gains. A non-destructive method would allow the assessment of superior genotypes without destroying the grains, and could also accelerate the breeding process.

Near infrared spectroscopy (NIR spectroscopy) could provide a better alternative for assessing the PE capacity of popcorn, as NIR spectroscopy is a non-destructive method for measuring the chemical constituents of biological materials (Pasquini et al., 2003). The PE capacity is related to the presence of a glassy (Quinn et al., 2005) or translucent endosperm with densely packed starch granules, which allow the kernels to expand (van der Sman and Bows, 2017). NIR spectroscopy has been used to predict the composition of maize kernel and has enabled the rapid selection of individual seeds with desirable traits, including the presence of starch, protein, oil, and phenolics (Baye et al., 2006; Meng et al., 2015). It also has been used to develop calibration models for the common varieties of corn, and for calibrating the quality traits of other species (Brito et al., 2013; Sinelli et al., 2010; Williams et al., 2009).

In order to select popcorn varieties with NIR spectroscopy on the basis of their phenotypic traits, and especially for enhanced PE capacity, samples of whole grains can be quickly screened, requiring no sample preparation, and the kernels can be preserved following measurement for further analyses and/or propagation. However, to the best of our knowledge, there are no NIR spectroscopy calibration and/or classification models that can be applied to correlate the traits of popcorn kernels in breeding programs. Therefore, the aim of this study was to evaluate whether NIR spectroscopy could be routinely used a tool in popcorn breeding programs for discriminating popcorn genotypes on the basis of their PE capacities.

2. Materials and Methods

2.1. Plant Materials

A total of 183 partial (S_3) inbred lines were obtained from nine different origins (commercial hybrids cultivars). The strains were separated based on the similarity of their agronomic characteristics, and nine populations were formed from the seed mix of the different strains in each group. Three generations (F_1, F_2, and $F_{2:3}$) were developed in three planting seasons by manual cross-pollination and self-pollination.

The first (F_1) and second (F_2) generations were developed at the experimental farm of Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias (FCAV), Campus de Jaboticabal, located in Jaboticabal, São Paulo, Brazil (latitude 21º 15' 17" S, longitude 48º 19' 20" W, and altitude of 605 m) during the season of 2012/2013. The F_1 generation and their reciprocal F_1' hybrids were obtained by complete diallel crosses between the populations thus formed. The F_2 generation was produced by sowing some seeds from the F_1 hybrids and some seeds from the F_1' reciprocals in the season of 2013/2014. The F_2 generation was generated by allowing self-fertilization of the hybrid combinations using the SIB (Self in Brothers) crossing method, following which the F_2 seeds thus generated from the 72 hybrid combinations were harvested and stored in a dry chamber.

The third ($F_{2:3}$) generation was developed in 2014 at the Texas A&M University Experimental Farm located in Weslaco, Texas, USA (latitude 26º 9' 33" N, longitude 97º 59' 15" W, and altitude of 24 m). Similar to the method employed for generating the F_2 plants, the $F_{2:3}$ plants were produced by self-fertilizing all the F_2 plants in the plot by manual pollination. All the ears were later identified, separately harvested, and dried in the shade. A total of 376 ears from the $F_{2:3}$ generation were selected, hand-shelled, and all the grains were considered for analysis. A total of 120 grains were randomly selected as samples from each ear for phenotyping and calibration by NIR spectroscopy.

2.2. Acquisition of NIR spectra

The NIR spectra of the intact popcorn kernels were obtained using a Thermo Scientific Antaris II FT-NIR Analyzer (Thermo Electron Co., USA). Prior to acquisition of the NIR spectra, the kernels were allow to equilibrate in the ambient humidity for two weeks in a controlled laboratory environment at ~25ºC and 12-13% relative humidity. All the measurements were made in a diffuse reflectance mode using a 225 mL rotating cup with a capacity to hold approximately 175 g of common maize over the integrating sphere module of the spectrometer. A foam support was placed in the cup to reduce its volume so as to accommodate 120 popcorn kernels per sample. A set of 331 samples was run once, with 64 scans for the samples of whole kernel. All the NIR spectra were computed at a resolution of 4 cm^{-1} across the spectral range of 4,000 - 10,000 cm^{-1} (1,000 to 2,500 nm) at ambient temperature (~25°C).

2.3. PE assessment: reference analysis

The samples were prepared according to the method described by Hoseney et al. (1983), but the time was adjusted according to the microwave model used and the number of kernels used for the study. Briefly, after scanning the samples of whole kernel by NIR spectroscopy, the 120 kernels in each sample were divided into three sub-samples of 40 kernels each. The sub-samples were weighed using a precision balance and popped in a paper bag in a microwave, using the maximum power setting of 1,350 w (60 Hz) for one minute and 30 seconds. The expansion volume of the samples was determined by calculating the mean ratio of the popcorn volume to the weight of each sub-sample ($mL.g^{-1}$). The volume of the popcorn was measured using a 1,000 mL graduated cylinder, and the cylinder was inverted once for each process to prevent packing. The variability in PE capacity among the samples was measured by analysis of

138 variance (ANOVA) using PROC MIXED in SAS software, (SAS, 2002). The analysis included
139 the F_2 population and $F_{2:3}$ generation as random effects in the model.
140
141 **2.4. Chemometrics**
142 All the spectral data was processed in a MATLAB® R2014b environment (Mathworks,
143 Natrick, USA) using the PLS Toolbox, version 7.9.3 (Eigenvector Research, Inc., Manson,
144 USA), and in-house algorithms. The data was pre-processed by the standard normal variate
145 (SNV) method, first by the Savitzky-Golay derivative (window of 7 points, 2^{nd} order polynomial
146 function), followed by vector normalization prior to chemometric analyses.
147 The prediction models for discriminating popcorn kernels based on the PE capacity were
148 developed using partial least squares (PLS), interval partial least squares (iPLS), and support
149 vector regression (SVR), where 70% of the samples were selected for calibration and full cross-
150 validation, and the remaining 30% was set aside for external validation. The samples were split
151 into calibration and validation sets using the Kennard-Stone sample selection algorithm
152 (Kennard and Stone, 1969).
153 For classification of the kernels on the basis of the PE values, the samples were divided
154 into training (70%), validation (15%), and test (15%) sets using the Kennard-Stone sample
155 selection algorithm (Kennard and Stone, 1969). The training set comprised 83 samples of class 1
156 kernels (PE $\leq$ 30 mL.g^{-1}) and 148 samples of class 2 kernels (PE > 30 mL.g^{-1}), and both the
157 validation and test sets had 18 samples of class 1 kernels and 32 samples of class 2 kernels each.
158 For classification, principal component analysis (PCA) with linear discriminant analysis
159 (PCA-LDA) and PCA with quadratic discriminant analysis (PCA-QDA) were performed. The
160 PCA-LDA and PCA-QDA algorithms are based on data reduction using PCA (Bro and Smilde,

2014) followed by discrimination of the PCA scores using LDA and QDA, respectively (Dixon and Brereton, 2009). The LDA (L_{ik}) and QDA (Q_{ik}) classification scores can be calculated by the following equations described by Costa et al., (2017):

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \sum_{\text{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) - 2\log_e \pi_k \quad (1)$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \sum_{k}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e |\sum_k| - 2\log_e \pi_k \quad (2)$$

Where, x_i is the vector containing the classification variables for sample i (*e.g.*, PCA scores for A components); $\bar{x}_k$ is the mean vector of class k; Σ_{pooled} is the pooled covariance matrix; Σ_k is the variance-covariance matrix of class k; and π_k is the prior probability of class k. The values of Σ_{pooled}, Σ_k, and π_k are calculated by the following equations described by Costa et al., (2017):

$$\sum_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \quad (3)$$

$$\sum_{\text{pooled}} = \frac{1}{n} \sum_{k=1}^{K} n_k \sum_k \quad (4)$$

$$\pi_k = \frac{n_k}{n} \quad (5)$$

Where n_k is the number of samples of class k; n is the total number of samples in the training set; and K is the number of classes.

The main difference between LDA and QDA is that LDA uses a pooled covariance matrix to calculate the discriminant function between the classes, whereas QDA uses the variance-covariance matrices of each class separately (Dixon and Brereton, 2009). Therefore, PCA-QDA usually achieves a better performance than PCA-LDA when analyzing complex datasets where the variance structures between the classes are very different.

In this study, classification was also achieved by variable selection using the genetic algorithm (GA), where both the LDA and QDA algorithms were combined with GA. These algorithms combine GA with LDA (GA-LDA) and GA with QDA (GA-QDA). In both the GA-

LDA and GA-QDA algorithms, GA is initially applied to reduce the pre-processed spectral data into a few number of variables based on an evolutionary process (McCall, 2005), following which LDA or QDA is applied to these selected variables using Eq. 1 or 2, respectively. The selected variables are of the same scale as the original spectral data and are selected according to the lowest risk of misclassification (*G*). The value of *G* is calculated from the validation set according to the following equation described by Carvalho et al. (2018):

$$G = \frac{1}{N_v}\sum_{n=1}^{N_v} g_n \quad (6)$$

Where, N_v is the number of validation samples, and g_n is defined as:

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \quad (7)$$

Where, the numerator is the squared Mahalanobis distance between sample x_n (of class index $I(n)$) and the mean $m_{I(n)}$ of its true class; and the denominator represents the squared Mahalanobis distance between sample x_n and the mean $m_{I(m)}$ of the closest unselected class. The GA was performed for 100 generations, with 200 chromosomes each. The cross-over and mutation probabilities were set at 60% and 1%, respectively. The algorithm was repeated three times and the best result was chosen.

The classification was finally performed using soft independent modeling of class analogy (SIMCA) and PLS regression with discriminant analysis (PLS-DA). SIMCA models are based on PCA models in which each class corresponds to a training set (Sabin et al., 2004). The use of PLS-DA maximizes the separation of pre-defined classes as it explains the variability within a general dataset (Wong et al., 2013).

2.4.1. Statistical evaluation

The algorithms used in this study were statistically evaluated by the accuracy, sensitivity, and specificity, which were calculated for each model. Sensitivity is defined as the proportion of positive samples correctly classified, and specificity is defined as the proportion of negative samples correctly classified. These figures of merit were calculated according to the following equations described by Baia et al. (2016) and Carvalho et al. (2016):

$$\text{Accuracy}\ (\%) = 100 - \left(\frac{1}{N}\sum_{h=1}^{H} y_h^*\right) \times 100 \quad (6)$$

$$\text{Sensitivity}\ (\%) = \frac{\text{TP}}{\text{TP+FN}} \times 100 \quad (7)$$

$$\text{Specificity}\ (\%) = \frac{\text{TN}}{\text{TN+FP}} \times 100 \quad (8)$$

Where, N is the total number of samples; H is the total number of classes; y_h^* is the number of samples incorrectly classified in class h; TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives.

3. Results

3.1. Genetic analysis

Genetic variability was observed in the F_2 and $F_{2:3}$ generations (Table 1). The coefficient of environmental variation (CVe) reflected the precision of the phenotyping measurement performed in this study, and the confidence of the analyses agreed with those of other popcorn phenotyping studies (Guimarães et al., 2018; Miotto et al., 2016).

3.2. NIR spectral features

The raw NIR spectra of all the popcorn kernels and the average SNV pre-processed spectra did not show obvious differences between the two classes (Figure 1), even when the first derivative of Savitzky-Golay was tested (data not shown). The NIR spectra exhibited two main absorption bands at 1116 nm and 1300 nm. The first is likely to be related to the second overtone

due to the C_{ar}-H group (Wust and Rudzik, 1996), attributed to the stretching vibration of the C-H and CH_3 groups associated with lignin, but a causal relationship could not be determined (Alves et al., 2010; Wust and Rudzik, 1996). The second absorption band could be attributed to the first overtone due to the C-H stretching vibration associated with hemicellulose (Schwanninger et al., 2011). Both bands indicated the hard popcorn kernel endosperm (Quinn et al., 2005; van der Sman and Bows, 2017). Five minor absorption bands were also observed and were related to the presence of cellulose (1854 nm and 2026 nm), O-H stretching vibration, and the second overtone of the C-O group (Fujimoto et al., 2007, Fujimoto et al., 2008; Osborne and Fearn, 1998; Siesler et al., 2002). The presence of lignin was again indicated by the absorption band at 2200 nm, and the C-H and C=O stretching vibrations (Workman and Weyer, 2007). Finally, the last absorption band at 2404 nm could be related to the presence of carbohydrates (starch), indicated by the C-H and C-C stretching vibrations (Schimleck and Evans, 2004).

3.3. PE capacity

The methodology described by Hoseney et al. (1983) was sufficient for determining the PE capacity for all the populations of popcorn. The average PE was 31.5 ± 6.7 $mL.g^{-1}$ with a standard error (SD) of 6.7 $mL.g^{-1}$. However, values as high as 48.9 $mL.g^{-1}$ and as low as 11.5 $mL.g^{-1}$ were also observed.

3.4. Chemometrics: PE prediction models

SVR and iPLS regression methods were applied to compare the predictive performance of the best regression model using PLSR-1D (Table 2). An iPLS model was built using 1300 wavelengths selected by cross-validating using the 1^{st} derived spectra (Savitzky-Golay, window

of 7 points, 2nd order polynomial function). The calibration and validation performances were estimated by the values of RMSEC and RMSEP, and the value of RMSEC was 6.03 $mL.g^{-1}$ (R_c^2 = 0.27) and that of RMSEP was 5.64 $mL.g^{-1}$ (R_p^2 = 0.06). An SVR model was constructed using a radial basis function kernel with 200 support vectors determined by cross-validation (cost = 100, epsilon = 1, and gamma = 10) with the 1st derived spectra, where the value of RMSEC for estimating the calibration performance was 7.05 $mL.g^{-1}$ (R_c^2 = 0.09), and the value of RMSEP for estimating the validation performance was 5.53 $mL.g^{-1}$ (R_p^2 = 0.001).

3.5. Chemometrics: PE classification models

The PE capacity represents the quality of the popcorn and the acceptable PE values for commercial purposes is approximately 30 $mL.g^{-1}$ in Brazil (Pina Matta and Viana, 2001). Therefore, the NIR spectra of all the popcorn kernels were separated into two classes (PE $\leq$ 30 and PE > 30 $mL.g^{-1}$). Genotypes that have PE values > 30 $mL.g^{-1}$ should be used in popcorn breeding programs.

The first attempt to classify popcorn kernels was carried out using the raw NIR spectra for developing the PCA-LDA and PCA-QDA models using four principal components (PCs) that accounted for a cumulative variance of 99.88% (Figure 2). The GA-LDA and GA-QDA models were developed by selecting the important variables. Specifically, the algorithm selected the spectral variables at 1300, 1316, 1564, 1916, 1944, 2156, 2220, and 2238 nm for GA-LDA and those at 1286, 1340, 1375, 1496, 1582, and 1672 nm for GA-QDA. However, it was not possible to obtain a good discrimination between the popcorn classes on basis of the resulting scores (Figure 2).

SNV was subsequently applied to pre-process the NIR spectra. The PCA-LDA and PCA-QDA models were developed using 4 PCs accounting for 98.69% of the cumulative variance (Figure 3). The GA-LDA and GA-QDA models used different variables, namely the spectral variables at 1280, 1211, 1230, and 2025 nm and the spectral variables at 1099, 1185, 1355, and 2022 nm, respectively. A better discrimination was subsequently obtained between the PE classes when compared to that achieved with the raw NIR spectra.

For the non-pre-processed data, SIMCA (3 PCs for class 1 and 4 PCs for class 2, 99% explained variance, determined by cross-validation with venetian blinds using 10 data splits) and PLS-DA (9 LVs determined by cross-validation with venetian blinds using 10 data splits, ~100% explained variance) were performed for comparing the performance of the classification. The SNV data were also analyzed by SIMCA (5 PCs for class 1 and 4 PCs for class 2, 99% explained variance, determined by cross-validation with venetian blinds using 10 data splits) and PLS-DA (6 LVs determined by cross-validation with venetian blinds with 10 data splits, 99% explained variance). For both the raw and pre-processed data, the performance of SIMCA and PLS-DA were inferior to that of PCA-QDA.

The classification rates were determined for all the models (Table 3). The sensitivity values of the SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA models were 68.7%, 65.6%, 68.7%, 81.2%, 65.6%, and 65.6%, respectively, when SNV pre-processed NIR spectra were used. Furthermore, the values of accuracy and specificity indicated that PE could be better classified by the PCA-QDA model than by the other discriminant models (Table 3).

4. Discussion

The genetic variability observed in the F_2 and $F_{2:3}$ populations should allow enhancement of the genetic gains in popcorn quality through selection. The PE capacity is the best example of such a trait in popcorn that can be enhanced through selection. The higher the precision in PE phenotyping the greater is the chance of obtaining the best popcorn genotype. In addition, most of the studies on popcorn genetic parameters have demonstrated that the genetic basis of PE is controlled by additive allele effects (Cabral et al., 2015), which further facilitates the enhancement of popcorn quality in breeding programs in terms of genetic gains.

The average PE value observed in this study reflected a modest but significant genetic variability. The F_2 and $F_{2:3}$ populations had the potential to produce genotypes with the desired PE values of approximately 40 $mL.g^{-1}$ (Pina Matta and Vianna, 2001), which is much higher than 30 $mL.g^{-1}$ found in commercial Brazilian popcorn populations (Amaral Júnior, et al. 2013).

The focus of this study was to evaluate whether NIR spectroscopy could be used as a tool for discriminating popcorn genotypes based on their PE capacity, and the results indicated that NIR spectroscopy could indeed discriminate popcorn genotypes on the basis of this quality parameter. The main NIR spectra absorption bands were associated with the presence of lignin (1116 nm and 2200 nm), hemicellulose (1300 nm), cellulose (1854 nm and 2026 nm), and starch (2404 nm), which possibly reflected the presence of a glassy (Quinn et al., 2005) and translucent endosperm with densely packed starch granules that allow the kernels to expand explosively (van der Sman and Bows, 2017). Maize kernels, including the hard, intermediate, and soft kernel types contain both glassy and floury endosperm in different ratios (Williams et al., 2009). In the hard kernel type, which is found in popcorn, the endosperm is primarily of a glassy nature. In soft kernels, the endosperm is mostly floury (Sweley et al., 2013), and the NIR spectra obtained

318 in this study was largely able to highlight the differences related to the presence of a hard kernel
319 in popcorn.

320 The accuracy of the PLS models for predicting the PE capacity of intact popcorn kernels
321 was reflected in the RMSEC (6.08 $mL.g^{-1}$), which represented 19.3% of the average PE.
322 However, an R_C^2 value of only 0.26 indicated a low proportion of explained variance in PE in the
323 calibration set. PLS models did not show good results due to the complexity of the data itself.
324 Even the testing results for variable selection by iPLS and non-linear regression by SVR were
325 inferior to that of PLS. This means that the spectral profile is affected by factors other than the
326 PE capacity of the kernel, and other experimental parameters such as moisture should be
327 measured for model correction. Baye et al. (2006) also reported low R^2 values while predicting
328 protein (0.16) and starch (0.23) in single maize kernels using NIR spectra and PLS models that
329 were intended to be used by geneticists and breeders for screening large numbers of samples.
330 NIR spectroscopy can still be used a tool for non-destructively reducing the number of samples
331 to be tested, however, some tolerance for false positives and negatives is necessary. On the
332 whole, the PLSR model had a low value of RMSEC, but the low values of residual predictive
333 deviation (RPD) and R_c^2 indicate that this model cannot be blindly used in popcorn breeding
334 programs for predicting PE values as inaccurate results are likely.

335 As the predictive model was not adequate, we tested the use of NIR spectroscopy for
336 classifying (discriminating) popcorn with different genetic compositions into different classes,
337 with the aim of rapid non-destructive phenotype selection, as this would probably be of greater
338 use to a popcorn breeding program. As the desirable value of PE for commercial purposes in
339 Brazil is approximately 30 $mL.g^{-1}$ (Amaral Júnior et al., 2013), the genotypes that have PE
340 values above 30 $mL.g^{-1}$ should be used in popcorn breeding programs for developing new

cultivars. However, in order to discriminate popcorn kernels into the two PE (≤ 30 and > 30 $mL.g^{-1}$) classes, no single NIR spectral feature could be employed for classification, making it necessary to apply computational analysis such as SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA.

The classification accuracy ranged from 62% (SIMCA) to 76% (GA-QDA) when the raw NIR spectra was used. However, a clear separation between popcorn kernel classes without overlap was not observed among the samples. This result might be related to the similarities between these popcorn classes and kernels of other popcorn populations (Sobierajski, 2012). The accuracy improved when the NIR spectra was pre-processed with SNV and the best result (78%) was obtained with PCA-QDA. Even though GA aided in selecting several important variables and reduced the problems of collinearity, this technique was not superior to PCA-QDA.

Overall, the discrimination between the two popcorn classes based on PE was more successful when PCA-QDA was applied, demonstrating that NIR spectroscopy together with powerful chemometric approaches has the potential to detect and identify popcorn genotypes that have PE values below and above the ideal target limit (30 $mL.g^{-1}$) in a breeding program. Recently, hyperspectral (NIR spectroscopy) imaging has been used to classify maize kernels on the basis of hardness (Williams and Kucheryavskiy, 2016), highlighting the possibilities of using NIR spectroscopy as a tool for discriminating popcorn kernels in breeding programs in the future.

5. Conclusion

NIR spectroscopy can be used as a tool for discriminating intact popcorn kernels based on their PE capacity. The quantitative PLS models developed herein should not be used in

popcorn breeding programs as inaccurate PE prediction values are expected. Instead, the PCA-QDA model can be applied for discriminating intact popcorn kernels with low PE (≤ 30 mL.g^{-1}) from those with high PE (> 30 mL.g^{-1}). Although NIR spectroscopy proved to be a promising non-destructive method for assessing the PE capacity of intact popcorn kernels, it is necessary to include more sources of variability and increase the sample size for improving the robustness of the predictive and classificatory models.

Acknowledgment

This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

References

Alves, A., Santos, A., Rozenberg, P., Paques, L.E., Charpentier, J.P., Schwanninger, M., Rodrigues, J., 2010. A common near infrared-based partial least squares regression model for the prediction of wood density of *Pinus pinaster* and *Larix* × *eurolepis*. Wood Science and Technology 46, 157-175.

Amaral Júnior, A.T.D., Gonçalves, L.S.A., Júnior, F., de Paiva, S., Candido, L.S., Vittorazzi, C., Scapim, C.A., 2013. UENF 14: a new popcorn cultivar. Crop Breeding and Applied Biotechnology 13, 218-220.

Amodio, M.L., Ceglie, F., Chaudhry, M.M., Piazzolla, F., Colelli, G., 2017. Potential of NIR spectroscopy for predicting internal quality and discriminating among strawberry fruits from different production systems. Postharvest Biology and Technology 125, 112–121.

Baia, T.C., Gama, R.A., de Lima, L.A.S., Lima, K.M.G., 2016. FTIR microspectroscopy coupled with variable selection methods for the identification of flunitrazepam in necrophagous flies. Analytical Methods 8, 968-972.

Baye, T.M., Pearson, T.C., Settles, A.M., 2006. Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. Journal of Cereal Science 43, 236-243.

Brito, A.L.B., Brito, L.R., Honorato, F.A., Pontes, M.J.C., Pontes, L.F.B.L., 2013. Classification of cereal bars using near infrared spectroscopy and linear discriminant analysis. Food Research International 51, 924-928.

Bro, R., Smilde, A.K., 2014. Principal component analysis. Analytical Methods 6, 2812-2831.

Carvalho, L.C., Morais, C.L.M., Lima, K.M.G., Cunha Júnior, L.C., Nascimento, P.A.M., Faria, J.B., Teixeira, G.H.A., 2016. Determination of the geographical origin and ethanol content of Brazilian sugarcane spirit using near-infrared spectroscopy coupled with discriminant analysis. Analytical Methods 8, 5658-5666.

Carvalho, L.C., Morais, C.L.M., Lima, K.M.G., Leite, G.W., Oliveira, G.S., Casagrande, I.P., Teixeira, G.H.A., 2018. Using intact nuts and near infrared spectroscopy to classify macadamia cultivars. Food analytical methods, 11, 1857–1866.

Costa, F.S., Silva, P.P., Morais, C.L.M., Theodoro, R.C., Arantes, T.D., Lima, K.M.G., 2017. Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish *Cryptococcus neoformans* and *Cryptococcus gattii* pathogenic fungi. Analytical Methods 9, 3968-3976.

Dixon, S.J., Brereton, R.G., 2009. Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis,

quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. Chemometrics and Intelligent Laboratory Systems 95, 1-17.

Fujimoto, T., Kurata, Y., Matsumoto, K., Tsuchikawa, S., 2008. Application of near infrared spectroscopy for estimating wood mechanical properties of small clear and full length lumber specimens. Journal of Near Infrared Spectroscopy 16, 529-537.

Fujimoto, T., Yamamoto, H., Tsuchikawa, S., 2007. Estimation of wood stiffness and strength properties of hybrid larch by near-infrared spectroscopy. Applied Spectroscopy 61, 882-888.

Galvão, J.C.C., Sawazaki, E., Miranda, G.V., 2000. Comportamento de híbridos de milho-pipoca em Coimbra, Minas Gerais, Brasil. Ceres 47, 201-2018.

Guimarães, A.G., Amaral Júnior, A.T., Lima, V.J.D., Leite, J.T., Scapim, C.A., Vivas, M., 2018. Genetic gains and selection advances of the UENF-14 popcorn population. Revista Caatinga 31, 271-278.

Hoseney, R.C., Zeleznak, K., Abdelrahman, A., 1983. Mechanism of popcorn popping. Journal of Cereal Science 1, 43-52.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

McCall, J., 2005. Genetic algorithms for modelling and optimisation. Journal of Computational and Applied Mathematics 184, 205-222.

Meng, Q., Murray, S.C., Mahan, A., Collison, A., Yang, L., Awika, J., 2015. Rapid estimation of phenolic content in colored Maize by near-infrared reflectance spectroscopy and its use in breeding. Crop Science, 55(5), 2234-2243.

Miotto, A.A., Pinto, R.J.B., Scapim, C.A., Matias Junior, J.L., Coan, M.M.D., Silva, H.A.D., 2016. Comparison of three tester parents in evaluating popcorn families derived from IAC-125. Revista Ciência Agronômica 47, 564-571.

Nicolai, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyna, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology 46, 99–118.

Oliveira, G.H.F.D., Amaral, C.B.D., Revolti, L.T.M., Buzinaro, R., Moro, G.V., 2019. Genetic variability in popcorn synthetic population. Acta Scientiarum 41, e39497.

Osborne, B.G., Fearn, T., 1998. Near infrared spectroscopy in food analysis. Longman Scientific & Technical, Harlow, Essex.

Pasquini, C., 2003. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemistry Society 14, 198-219.

Pina Matta, F., Viana, J.M.S., 2011. Testes de capacidade de expansão em programas de melhoramento de milho-pipoca. Scientia Agricola 58, 847-851.

Quinn, P.V., Hong, D.C., Both, J.A., 2005. Increasing the size of a piece of popcorn. Physica A: Statistical Mechanics and its Applications 353, 637-648.

Sabin, J.G., Ferrão, M.F., Furtado, J.C., 2004. Análise multivariada aplicada na identificação de fármacos antidepressivos.Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. Revista Brasileira de Ciências Farmacêuticas 40, 387-396.

SAS., 2002. SAS User's guide: statistics, eighth ed. SAS Institute Inc. Cary.

Sawazaki, E., 2011. Nova geração de híbrido de milho para pipoca. Disponível em: https://www.agrolink.com.br/noticias/nova-geracao-de-hibrido-de-milho-para-pipoca_131978.html. Accessed 14 February 2019.

Schimleck, L.R., Evans, R., 2004. Estimation of *Pinus radiata* D. Don tracheid morphological characteristics by near infrared spectroscopy. Holzforschung 5, 66-73.

Schwanninger, M., Rodrigues, J.C., Fackler, K., 2011. A review of band assignments in near infrared spectra of wood and wood components. Journal of Near Infrared Spectroscopy 19, 287–308.

Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M., 2002. Near-infrared spectroscopy. Principles, instruments, applications, first Ed. Wiley-VCH Verlag GmbH, Weinheim.

Cabral, P.D.S., Teixeira do Amaral, A., Pio Viana, A., Duarte Vieira, H., Jesus Freitas, I.L., Vittorazzi, C., Vivas, M., 2015. Combining ability between tropical and temperate popcorn lines for seed quality and agronomic traits. Australian Journal of Crop Science 9, 256.

Sinelli, N., Cerretani, L., Di Egidio, V., Bendini, A., Casiraghi, E., 2010. Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity. Food Research International 43, 369-375.

Sobierajski, G.R., 2012. Desenvolvimento e uso de marcadores SSR e DArT para estudos de diversidade genética em macadâmia (*Macadamia integrifolia*). PhD Thesis. Universidade São Paulo, Piracicaba.

Sweley, J. C., Rose, D. J., & Jackson, D. S. (2013). Quality traits and popping performance considerations for popcorn (*Zea mays* Everta). Food reviews international 29, 157-177.

van der Sman, R.G.M., Bows, J.R., 2017. Critical factors in microwave expansion of starchy snacks. Journal of Food Engineering 211, 69-84.

Williams, P., Geladi, P., Fox, G., Manley, M., 2009. Maize kernel hardness classification by near infrared (NIR) hyperspectral imaging and multivariate data analysis. Analytica Chimica Acta 653, 121-130.

Williams, P.J., Kucheryavskiy, S., 2016. Classification of maize kernels using NIR hyperspectral imaging. Food Chemistry 209, 131–138.

Wong, K.H., Razmovski-Naumovski, V., Li, K.M., Li, G.Q., Chan, K., 2013. Differentiation of *Pueraria lobata* and *Pueraria thomsonii* using partial least square discriminant analysis (PLS-DA). Journal of Pharmaceutical and Biomedical Analysis 84, 5-13.

Workman, J., Weyer, L., 2007. Practical guide to interpretive near-infrared spectroscopy, first ed. CRC Press, Boca Raton.

Wust, E., Rudzik, L., 1996. NIR-Spektroskopische analytik. In: Gunzler, A.M.B.H., Borsdorf, R., Danzer, K., Fresenius, W., Galensa, R., Huber, W., Luderwald, I., Schwedt, G., Tolg, G., Wisser, H. (Eds.), Infrarotspektroskopie. Highlight aus dem Analytiker-Taschenbuch. Springer, Berlin, pp. 217-232.

Zinsly, J.R., Machado, J.A., 1987. Milho-pipoca. In: Paterniani, E., Viegas, G.P. (Eds.). Melhoramento e produção de milho. Fundação Cargill, Piracicaba, pp.411-450.

492
493

494 **Tables**

495

496 **Table 1**. Deviance analysis by the likelihood ratio test (LRT) among 62 F_2 populations evaluated

497 for popping expansion (PE, $mL.g^{-1}$).

SV	DF (χ^2)	PE	
		σ^2	LRT^2
Random effect			
F_2 population	1	7.75^{**}	412.85
$F_{2:3}$ generation	1	30.90^{**}	19.62
Residual		39.62	
Fixed effect	DF	F-value	P > F
Rep	2	1.60^{ns}	0.20
CVe(%)			11.13

498 ** and ns, significant at 0.01, no-significant, respectively; SV: source of variation; DF **(χ^2):**

499 degrees of freedom of the chi-square analysis; Rep: Repetitions; PE: Popping expansion ($mL.g^{-1}$);

500 CVe(%):Environment coefficient of variation.

501

Table 2. Partial least squares regression (PLSR) for popcorn kernel popping expansion (PE, $mL.g^{-1}$).

PLSR						
	LV	R_c2	RMSEC	RMSECV	R_p2	RMSEP
PLSR-nil	3	0.18	6.37	6.51	0.05	5.56
SNV	4	0.19	6.35	6.62	0.08	5.49
PLSR-1D	3	0.26	6.08	6.50	0.10	5.38
PLSR-SVN+1D	2	0.21	6.29	6.60	0.06	5.54

LV = latent variable, RMSEC = root mean square error of calibration, RMSECV = root mean square error of cross-validation; RMSEP = root mean square error of prediction, nil = raw NIR spectra, SNV = standard normal variate, 1D = 1^{st} derivative Savitzky-Golay (window of 7 points, 2^{nd} order polynomial function).

Table 3. Classification of popcorn kernels based on popping expansion (PE, $mL.g^{-1}$) and NIR spectroscopy using SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA.

Pre-processing	Class	Correct Classification (%)			Figure of Merit (%)		
		Training	Validation	Test	Accuracy	Sensitivity	Specificity
SIMCA							
None	> 30	78.3	61.1	66.7	62.0	59.4	66.7
	< 30	64.9	65.6	59.4			
SNV	> 30	74.7	55.6	61.1	66.0	68.7	61.1
	< 30	66.2	71.9	68.7			
PLS-DA							
None	> 30	73.5	55.6	72.2	74.0	75.0	72.2
	< 30	77.7	84.4	75.0			
SNV	> 30	63.9	61.1	66.7	66.0	65.6	66.7
	< 30	68.9	78.1	65.6			
PCA-LDA							
None	> 30	57.8	44.4	66.7	70.0	71.9	66.7
	< 30	54.7	43.7	71.9			
SNV	> 30	63.9	61.1	66.7	68.0	68.7	66.7
	< 30	62.2	78.1	68.7			
PCA-QDA							
None	> 30	55.4	44.4	71.9	70.0	84.4	44.4
	< 30	67.6	59.4	84.4			
SNV	> 30	61.4	55.6	72.2	**78.0**	**81.2**	**72.2**
	< 30	71.6	75.0	81.2			
GA-LDA							
None	> 30	80.7	66.7	55.6	74.0	84.4	55.6
	< 30	77.0	84.4	84.4			
SNV	> 30	65.1	66.7	77.8	70.0	65.6	77.8
	< 30	60.8	78.1	65.6			
GA-QDA							
None	> 30	80.7	77.8	66.7	76.0	81.2	66.7
	< 30	66.9	78.1	81.2			
SNV	> 30	72.2	72.2	72.2	68.0	65.6	72.2
	< 30	78.1	78.1	65.6			

Soft independent modeling of class analogy (SIMCA), parcial least square regression with discriminant analysis (PLS-DA), principal component analysis linear with discriminant analysis (PCA-LDA), principal component analysis with quadratic discriminant analysis (PCA-QDA), genetic algorithm with linear discriminant analysis (GA-LDA), and genetic algorithm with quadratic discriminant analysis (GA-QDA). SNV = standard normal variate

Figures

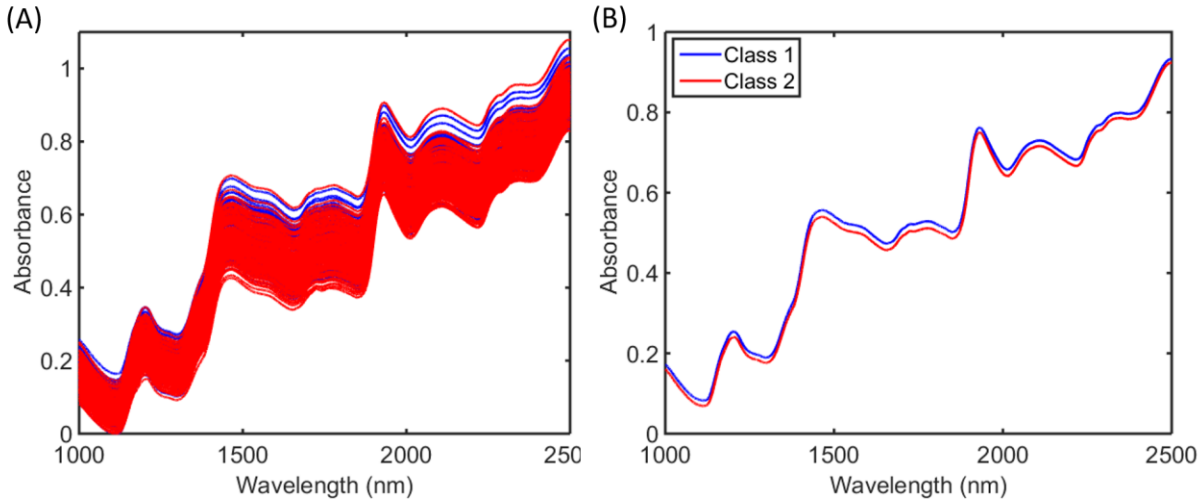


Figure 1. Near infrared spectra ($-\log_{10}R$) of all popcorn kernel samples (A), average spectra for each original class of popcorn sample (B). Class 1 ($PE \leq 30$ $mL.g^{-1}$) and Class 2 ($PE > 30$ $mL.g^{-1}$).

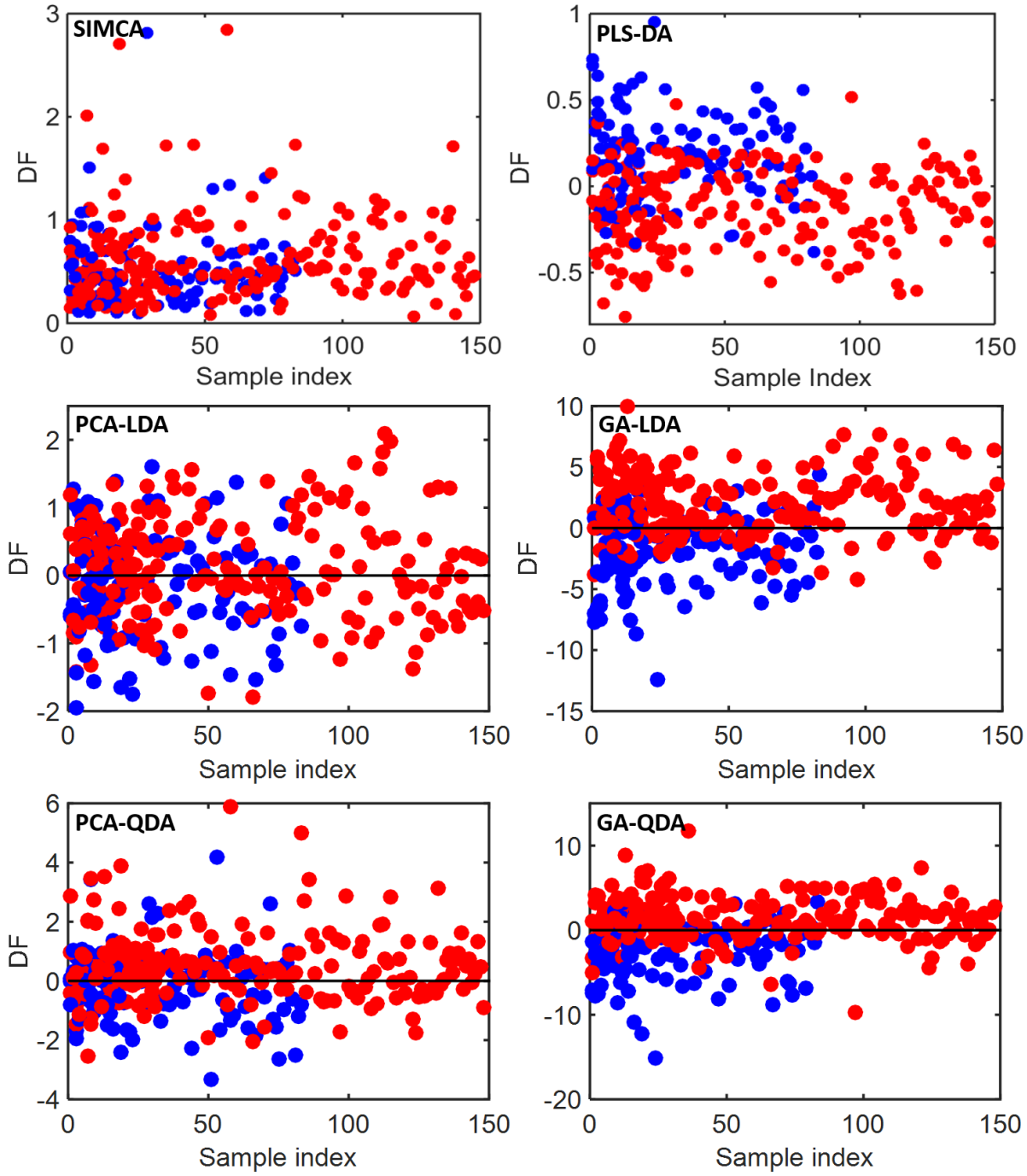
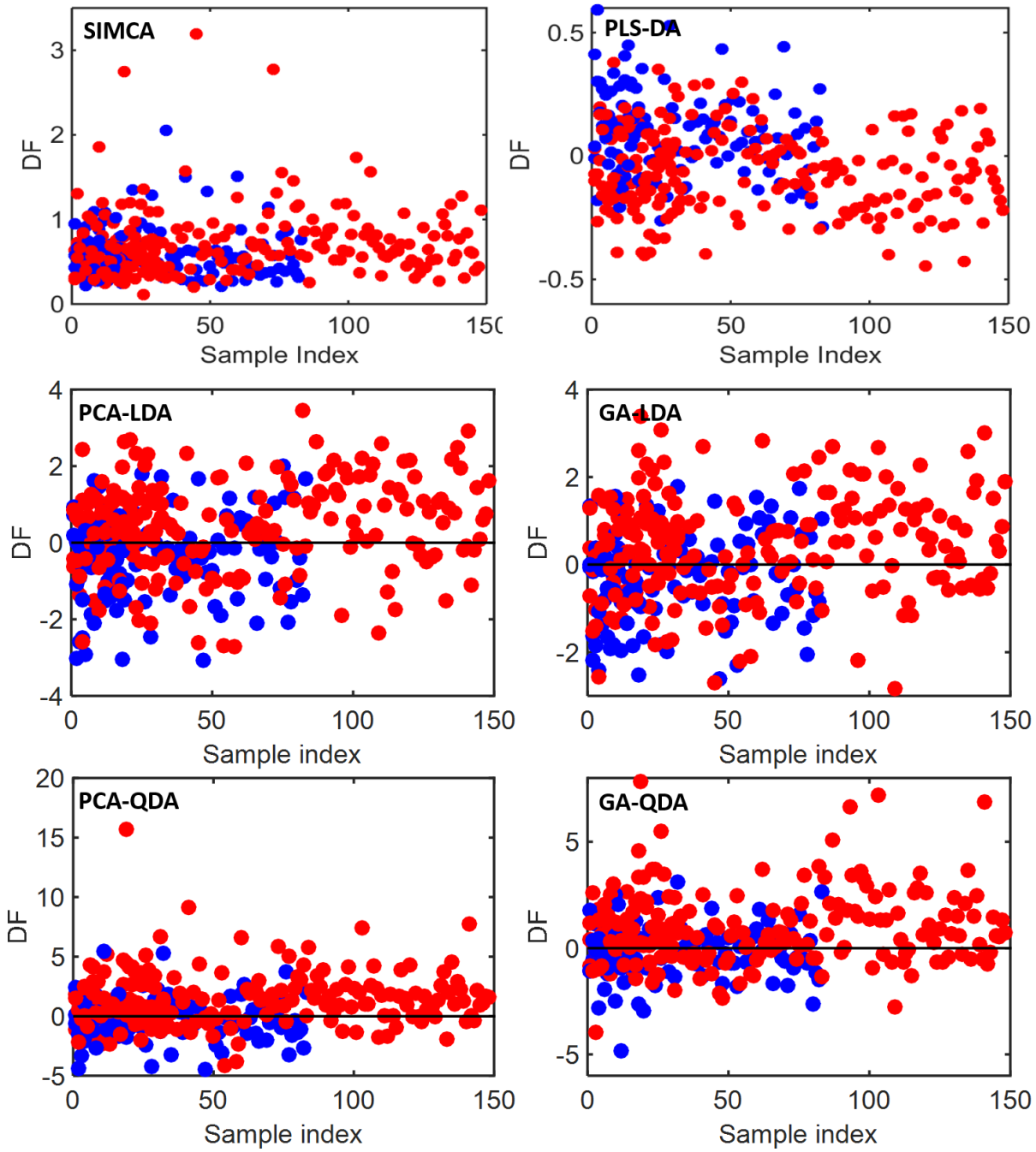


Figure 2. Discriminant function (DF) *versus* samples calculated by using SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA models from two classes of popping expansion (PE). Red dots, class 1 (PE ≤ 30 mL.g^{-1}) and blue dots, class 2 (PE > 30 mL.g^{-1}) without NIR spectra pre-processing.

536



537

538 **Figure 3.** Discriminant function (DF) *versus* samples calculated by using SIMCA, PLS-DA,
539 PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA models from two classes of popping expansion
540 (PE). Red dots, class 1 (PE $\leq$ 30 mL.g^{-1}) and blue dots, class 2 (PE > 30 mL.g^{-1}) with NIR
541 spectra pre-processed with SNV.