

Central Lancashire Online Knowledge (CLoK)

Title	Estimation and classification of popping expansion capacity in popcorn breeding programs using NIR spectroscopy
Type	Article
URL	https://clock.uclan.ac.uk/30398/
DOI	https://doi.org/10.1016/j.jcs.2019.102861
Date	2020
Citation	Ferreira de Oliveira, Gustavo Hugo, Murray, Seth C., Cunha Júnior, Luis Carlos, Gomes de Lima, Kássio Michell, Medeiros-De-morais, Camilo De Ielis orcid iconORCID: 0000-0003-2573-787X, Henrique de Almeida Teixeira, Gustavo and Môro, Gustavo Vitti (2020) Estimation and classification of popping expansion capacity in popcorn breeding programs using NIR spectroscopy. <i>Journal of Cereal Science</i> , 91 (102861). ISSN 0733-5210
Creators	Ferreira de Oliveira, Gustavo Hugo, Murray, Seth C., Cunha Júnior, Luis Carlos, Gomes de Lima, Kássio Michell, Medeiros-De-morais, Camilo De Ielis, Henrique de Almeida Teixeira, Gustavo and Môro, Gustavo Vitti

It is advisable to refer to the publisher's version if you intend to cite from the work.
<https://doi.org/10.1016/j.jcs.2019.102861>

For information about Research at UCLan please go to <http://www.uclan.ac.uk/research/>

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the <http://clock.uclan.ac.uk/policies/>

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

3

4 Gustavo Hugo Ferreira de Oliveira¹, Seth C. Murray², Luis Carlos Cunha Júnior³, Kássio Michell
5 Gomes de Lima⁴, Camilo de Lelis Medeiros de Morais⁵, Gustavo Henrique de Almeida
6 Teixeira⁶, Gustavo Vitti Môro⁶

7

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

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

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

16 ⁴Universidade Federal do Rio Grande do Norte (UFRN), Instituto de Química, Química
17 Biológica e Quimiometria, Avenida Senador Salgado Filho, nº 3000, Bairro de Lagoa Nova,
18 CEP: 59.078-970, Natal, Rio Grande do Norte, Brazil

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

21 ⁶Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias
22 (FCAV), Campus de Jaboticabal. Via de Acesso Prof. Paulo Donato Castellane s/n, CEP: 14884-
23 900 – Brasil, Jaboticabal, São Paulo, Brazil

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

25

26 **Abstract**

27 One of the most important quality traits in popcorn breeding programs is the popping
28 expansion (PE) capacity of the kernel, which is the ratio of the volume of the popcorn to the
29 weight of the kernel. In this study, we evaluated whether near infrared spectroscopy (NIR
30 spectroscopy) could be used as a tool in popcorn breeding programs to routinely predict and/or
31 discriminate popcorn genotypes on the basis of their PE. Three generations (F₁, F₂, and F_{2:3})
32 were developed in three planting seasons by manual cross-pollination and self-pollination. A
33 total of 376 ears from the F_{2:3} generation were selected, shelled, and subjected to phenotypic
34 analysis. Genetic variability was observed in the F₂ and F_{2:3} generations, and their average PE
35 value was $31.5 \pm 6.7 \text{ mL.g}^{-1}$. PE prediction models using partial least square (PLS) regression
36 were developed, and the root mean square error of calibration (RMSEC) was 6.08 mL.g^{-1} , while
37 the coefficient of determination (R_c^2) was 0.26. The model developed by principal component
38 analysis with quadratic discriminant analysis (PCA-QDA) was the best for discriminating the
39 kernels with low PE ($\leq 30 \text{ mL.g}^{-1}$) from those with high PE ($> 30 \text{ mL.g}^{-1}$) with an accuracy of
40 78%, sensitivity of 81.2%, and specificity of 72.2%. Although NIR spectroscopy appears to be a
41 promising non-destructive method for assessing the PE of intact popcorn kernels for narrow
42 breeding populations, greater variability and larger sample sizes would help improve the
43 robustness of the predictive and classificatory models.

44

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

46

47 **1. Introduction**

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

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

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

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

71 Near infrared spectroscopy (NIR spectroscopy) could provide a better alternative for
72 assessing the PE capacity of popcorn, as NIR spectroscopy is a non-destructive method for
73 measuring the chemical constituents of biological materials (Pasquini et al., 2003). The PE
74 capacity is related to the presence of a glassy (Quinn et al., 2005) or translucent endosperm with
75 densely packed starch granules, which allow the kernels to expand (van der Sman and Bows,
76 2017). NIR spectroscopy has been used to predict the composition of maize kernel and has
77 enabled the rapid selection of individual seeds with desirable traits, including the presence of
78 starch, protein, oil, and phenolics (Baye et al., 2006; Meng et al., 2015). It also has been used to
79 develop calibration models for the common varieties of corn, and for calibrating the quality traits
80 of other species (Brito et al., 2013; Sinelli et al., 2010; Williams et al., 2009).

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

89

90 **2. Materials and Methods**

91 **2.1. Plant Materials**

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

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

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

115 **2.2. Acquisition of NIR spectra**

116 The NIR spectra of the intact popcorn kernels were obtained using a Thermo Scientific
117 Antaris II FT-NIR Analyzer (Thermo Electron Co., USA). Prior to acquisition of the NIR
118 spectra, the kernels were allow to equilibrate in the ambient humidity for two weeks in a
119 controlled laboratory environment at ~25°C and 12-13% relative humidity. All the measurements
120 were made in a diffuse reflectance mode using a 225 mL rotating cup with a capacity to hold
121 approximately 175 g of common maize over the integrating sphere module of the spectrometer.
122 A foam support was placed in the cup to reduce its volume so as to accommodate 120 popcorn
123 kernels per sample. A set of 331 samples was run once, with 64 scans for the samples of whole
124 kernel. All the NIR spectra were computed at a resolution of 4 cm⁻¹ across the spectral range of
125 4,000 - 10,000 cm⁻¹ (1,000 to 2,500 nm) at ambient temperature (~25°C).

126

127 **2.3. PE assessment: reference analysis**

128 The samples were prepared according to the method described by Hosney et al. (1983),
129 but the time was adjusted according to the microwave model used and the number of kernels
130 used for the study. Briefly, after scanning the samples of whole kernel by NIR spectroscopy, the
131 120 kernels in each sample were divided into three sub-samples of 40 kernels each. The sub-
132 samples were weighed using a precision balance and popped in a paper bag in a microwave,
133 using the maximum power setting of 1,350 w (60 Hz) for one minute and 30 seconds. The
134 expansion volume of the samples was determined by calculating the mean ratio of the popcorn
135 volume to the weight of each sub-sample (mL.g⁻¹). The volume of the popcorn was measured
136 using a 1,000 mL graduated cylinder, and the cylinder was inverted once for each process to
137 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 F₂ 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 Natick, 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, 2nd 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 ($PE \leq 30 \text{ mL.g}^{-1}$) and 148 samples of class 2 kernels ($PE > 30 \text{ mL.g}^{-1}$), 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,

161 2014) followed by discrimination of the PCA scores using LDA and QDA, respectively (Dixon
 162 and Brereton, 2009). The LDA (L_{ik}) and QDA (Q_{ik}) classification scores can be calculated by the
 163 following equations described by Costa et al., (2017):

$$164 \quad L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \Sigma_{\text{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) - 2 \log_e \pi_k \quad (1)$$

$$165 \quad Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e |\Sigma_k| - 2 \log_e \pi_k \quad (2)$$

166 Where, x_i is the vector containing the classification variables for sample i (e.g., PCA
 167 scores for A components); \bar{x}_k is the mean vector of class k ; Σ_{pooled} is the pooled covariance matrix;
 168 Σ_k is the variance-covariance matrix of class k ; and π_k is the prior probability of class k . The
 169 values of Σ_{pooled} , Σ_k , and π_k are calculated by the following equations described by Costa et al.,
 170 (2017):

$$171 \quad \Sigma_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k) (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \quad (3)$$

$$172 \quad \Sigma_{\text{pooled}} = \frac{1}{n} \sum_{k=1}^K n_k \Sigma_k \quad (4)$$

$$173 \quad \pi_k = \frac{n_k}{n} \quad (5)$$

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

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

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

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

$$190 \quad G = \frac{1}{N_v} \sum_{n=1}^{N_v} g_n \quad (6)$$

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

$$192 \quad g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m) \neq I(n)} r^2(x_n, m_{I(m)})} \quad (7)$$

193 Where, the numerator is the squared Mahalanobis distance between sample x_n (of class index
194 $I(n)$) and the mean $m_{I(n)}$ of its true class; and the denominator represents the squared Mahalanobis
195 distance between sample x_n and the mean $m_{I(m)}$ of the closest unselected class. The GA was
196 performed for 100 generations, with 200 chromosomes each. The cross-over and mutation
197 probabilities were set at 60% and 1%, respectively. The algorithm was repeated three times and
198 the best result was chosen.

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

204

205 **2.4.1. Statistical evaluation**

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

$$211 \quad \text{Accuracy (\%)} = 100 - \left(\frac{1}{N} \sum_{h=1}^H y_h^* \right) \times 100 \quad (6)$$

$$212 \quad \text{Sensitivity (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (7)$$

$$213 \quad \text{Specificity (\%)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (8)$$

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

217

218 **3. Results**

219 **3.1. Genetic analysis**

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

224 **3.2. NIR spectral features**

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

229 due to the C_{ar}-H group (Wust and Rudzik, 1996), attributed to the stretching vibration of the C-H
230 and CH₃ groups associated with lignin, but a causal relationship could not be determined (Alves
231 et al., 2010; Wust and Rudzik, 1996). The second absorption band could be attributed to the first
232 overtone due to the C-H stretching vibration associated with hemicellulose (Schwanninger et al.,
233 2011). Both bands indicated the hard popcorn kernel endosperm (Quinn et al., 2005; van der
234 Sman and Bows, 2017). Five minor absorption bands were also observed and were related to the
235 presence of cellulose (1854 nm and 2026 nm), O-H stretching vibration, and the second overtone
236 of the C-O group (Fujimoto et al., 2007, Fujimoto et al., 2008; Osborne and Fearn, 1998; Siesler
237 et al., 2002). The presence of lignin was again indicated by the absorption band at 2200 nm, and
238 the C-H and C=O stretching vibrations (Workman and Weyer, 2007). Finally, the last absorption
239 band at 2404 nm could be related to the presence of carbohydrates (starch), indicated by the C-H
240 and C-C stretching vibrations (Schimleck and Evans, 2004).

241

242 **3.3. PE capacity**

243 The methodology described by Hoseney et al. (1983) was sufficient for determining the
244 PE capacity for all the populations of popcorn. The average PE was $31.5 \pm 6.7 \text{ mL.g}^{-1}$ with a
245 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
246 mL.g^{-1} were also observed.

247

248 **3.4. Chemometrics: PE prediction models**

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

252 of 7 points, 2nd order polynomial function). The calibration and validation performances were
253 estimated by the values of RMSEC and RMSEP, and the value of RMSEC was 6.03 mL.g⁻¹ (R_c^2
254 = 0.27) and that of RMSEP was 5.64 mL.g⁻¹ ($R_p^2 = 0.06$). An SVR model was constructed using
255 a radial basis function kernel with 200 support vectors determined by cross-validation (cost =
256 100, epsilon = 1, and gamma = 10) with the 1st derived spectra, where the value of RMSEC for
257 estimating the calibration performance was 7.05 mL.g⁻¹ ($R_c^2 = 0.09$), and the value of RMSEP for
258 estimating the validation performance was 5.53 mL.g⁻¹ ($R_p^2 = 0.001$).

259

260 **3.5. Chemometrics: PE classification models**

261 The PE capacity represents the quality of the popcorn and the acceptable PE values for
262 commercial purposes is approximately 30 mL.g⁻¹ in Brazil (Pina Matta and Viana, 2001).
263 Therefore, the NIR spectra of all the popcorn kernels were separated into two classes (PE ≤ 30
264 and PE > 30 mL.g⁻¹). Genotypes that have PE values > 30 mL.g⁻¹ should be used in popcorn
265 breeding programs.

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

274 SNV was subsequently applied to pre-process the NIR spectra. The PCA-LDA and PCA-
275 QDA models were developed using 4 PCs accounting for 98.69% of the cumulative variance
276 (Figure 3). The GA-LDA and GA-QDA models used different variables, namely the spectral
277 variables at 1280, 1211, 1230, and 2025 nm and the spectral variables at 1099, 1185, 1355, and
278 2022 nm, respectively. A better discrimination was subsequently obtained between the PE
279 classes when compared to that achieved with the raw NIR spectra.

280 For the non-pre-processed data, SIMCA (3 PCs for class 1 and 4 PCs for class 2, 99%
281 explained variance, determined by cross-validation with venetian blinds using 10 data splits) and
282 PLS-DA (9 LVs determined by cross-validation with venetian blinds using 10 data splits, ~100%
283 explained variance) were performed for comparing the performance of the classification. The
284 SNV data were also analyzed by SIMCA (5 PCs for class 1 and 4 PCs for class 2, 99% explained
285 variance, determined by cross-validation with venetian blinds using 10 data splits) and PLS-DA
286 (6 LVs determined by cross-validation with venetian blinds with 10 data splits, 99% explained
287 variance). For both the raw and pre-processed data, the performance of SIMCA and PLS-DA
288 were inferior to that of PCA-QDA.

289 The classification rates were determined for all the models (Table 3). The sensitivity
290 values of the SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA models were
291 68.7%, 65.6%, 68.7%, 81.2%, 65.6%, and 65.6%, respectively, when SNV pre-processed NIR
292 spectra were used. Furthermore, the values of accuracy and specificity indicated that PE could be
293 better classified by the PCA-QDA model than by the other discriminant models (Table 3).

294

295 **4. Discussion**

296 The genetic variability observed in the F₂ and F_{2:3} populations should allow enhancement
297 of the genetic gains in popcorn quality through selection. The PE capacity is the best example of
298 such a trait in popcorn that can be enhanced through selection. The higher the precision in PE
299 phenotyping the greater is the chance of obtaining the best popcorn genotype. In addition, most
300 of the studies on popcorn genetic parameters have demonstrated that the genetic basis of PE is
301 controlled by additive allele effects (Cabral et al., 2015), which further facilitates the
302 enhancement of popcorn quality in breeding programs in terms of genetic gains.

303 The average PE value observed in this study reflected a modest but significant genetic
304 variability. The F₂ and F_{2:3} populations had the potential to produce genotypes with the desired
305 PE values of approximately 40 mL.g⁻¹ (Pina Matta and Vianna, 2001), which is much higher than
306 30 mL.g⁻¹ found in commercial Brazilian popcorn populations (Amaral Júnior, et al. 2013).

307 The focus of this study was to evaluate whether NIR spectroscopy could be used as a tool
308 for discriminating popcorn genotypes based on their PE capacity, and the results indicated that
309 NIR spectroscopy could indeed discriminate popcorn genotypes on the basis of this quality
310 parameter. The main NIR spectra absorption bands were associated with the presence of lignin
311 (1116 nm and 2200 nm), hemicellulose (1300 nm), cellulose (1854 nm and 2026 nm), and starch
312 (2404 nm), which possibly reflected the presence of a glassy (Quinn et al., 2005) and translucent
313 endosperm with densely packed starch granules that allow the kernels to expand explosively (van
314 der Sman and Bows, 2017). Maize kernels, including the hard, intermediate, and soft kernel
315 types contain both glassy and floury endosperm in different ratios (Williams et al., 2009). In the
316 hard kernel type, which is found in popcorn, the endosperm is primarily of a glassy nature. In
317 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 ($6.08 \text{ mL}\cdot\text{g}^{-1}$), 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 $30 \text{ mL}\cdot\text{g}^{-1}$ (Amaral Júnior et al., 2013), the genotypes that have PE
340 values above $30 \text{ mL}\cdot\text{g}^{-1}$ should be used in popcorn breeding programs for developing new

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

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

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

360

361 **5. Conclusion**

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

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

370

371 **Acknowledgment**

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

374

375 **References**

376

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

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

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

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

390 Baye, T.M., Pearson, T.C., Settles, A.M., 2006. Development of a calibration to predict maize
391 seed composition using single kernel near infrared spectroscopy. Journal of Cereal Science 43,
392 236-243.

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

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

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

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

404 Costa, F.S., Silva, P.P., Morais, C.L.M., Theodoro, R.C., Arantes, T.D., Lima, K.M.G., 2017.
405 Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish
406 *Cryptococcus neoformans* and *Cryptococcus gattii* pathogenic fungi. Analytical Methods 9,
407 3968-3976.

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

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

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

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

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

419 Guimarães, A.G., Amaral Júnior, A.T., Lima, V.J.D., Leite, J.T., Scapim, C.A., Vivas, M., 2018.
420 Genetic gains and selection advances of the UENF-14 popcorn population. *Revista Caatinga* 31,
421 271-278.

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

424 Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. *Technometrics* 11,
425 137-148.

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

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

431 Miotto, A.A., Pinto, R.J.B., Scapim, C.A., Matias Junior, J.L., Coan, M.M.D., Silva, H.A.D.,
432 2016. Comparison of three tester parents in evaluating popcorn families derived from IAC-125.
433 Revista Ciência Agronômica 47, 564-571.

434 Nicolai, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyna, J.,
435 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy:
436 A review. Postharvest Biology and Technology 46, 99–118.

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

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

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

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

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

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

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

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

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

456 Schwanninger, M., Rodrigues, J.C., Fackler, K., 2011. A review of band assignments in near
457 infrared spectra of wood and wood components. *Journal of Near Infrared Spectroscopy* 19, 287–
458 308.

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

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

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

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

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

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

474 Williams, P., Geladi, P., Fox, G., Manley, M., 2009. Maize kernel hardness classification by near
475 infrared (NIR) hyperspectral imaging and multivariate data analysis. *Analytica Chimica Acta*
476 653, 121-130.

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

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

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

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

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

490

491

492
493

494 **Tables**

495

496 **Table 1.** Deviance analysis by the likelihood ratio test (LRT) among 62 F₂ 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			
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

502
503

504 **Table 2.** Partial least squares regression (PLSR) for popcorn kernel popping expansion (PE,
505 mL.g⁻¹).

PLSR						
	LV	R _c 2	RMSEC	RMSECV	R _p 2	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

506 LV = latent variable, RMSEC = root mean square error of calibration, RMSECV = root mean
507 square error of cross-validation; RMSEP = root mean square error of prediction, nil = raw NIR
508 spectra, SNV = standard normal variate, 1D = 1st derivative Savitzky-Golay (window of 7 points,
509 2nd order polynomial function).

510

511

512 **Table 3.** Classification of popcorn kernels based on popping expansion (PE, mL.g⁻¹) and NIR

513 spectroscopy using SIMCA, PLS-DA, PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA.

SIMCA							
Pre-processing	Class	Correct Classification (%)			Figure of Merit (%)		
		Training	Validation	Test	Accuracy	Sensitivity	Specificity
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			

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

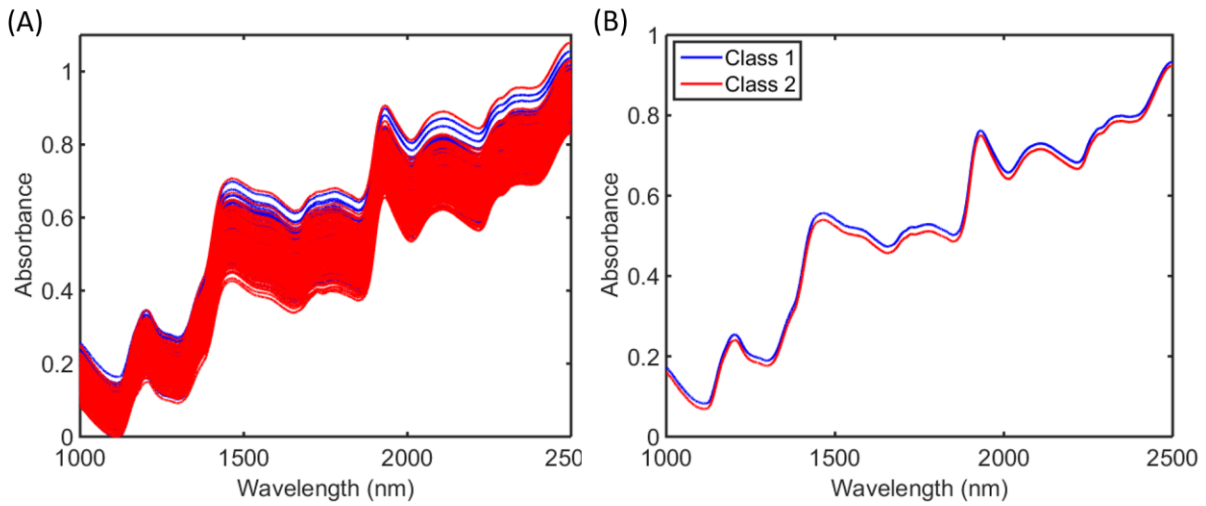
519

520

521
522

523 **Figures**

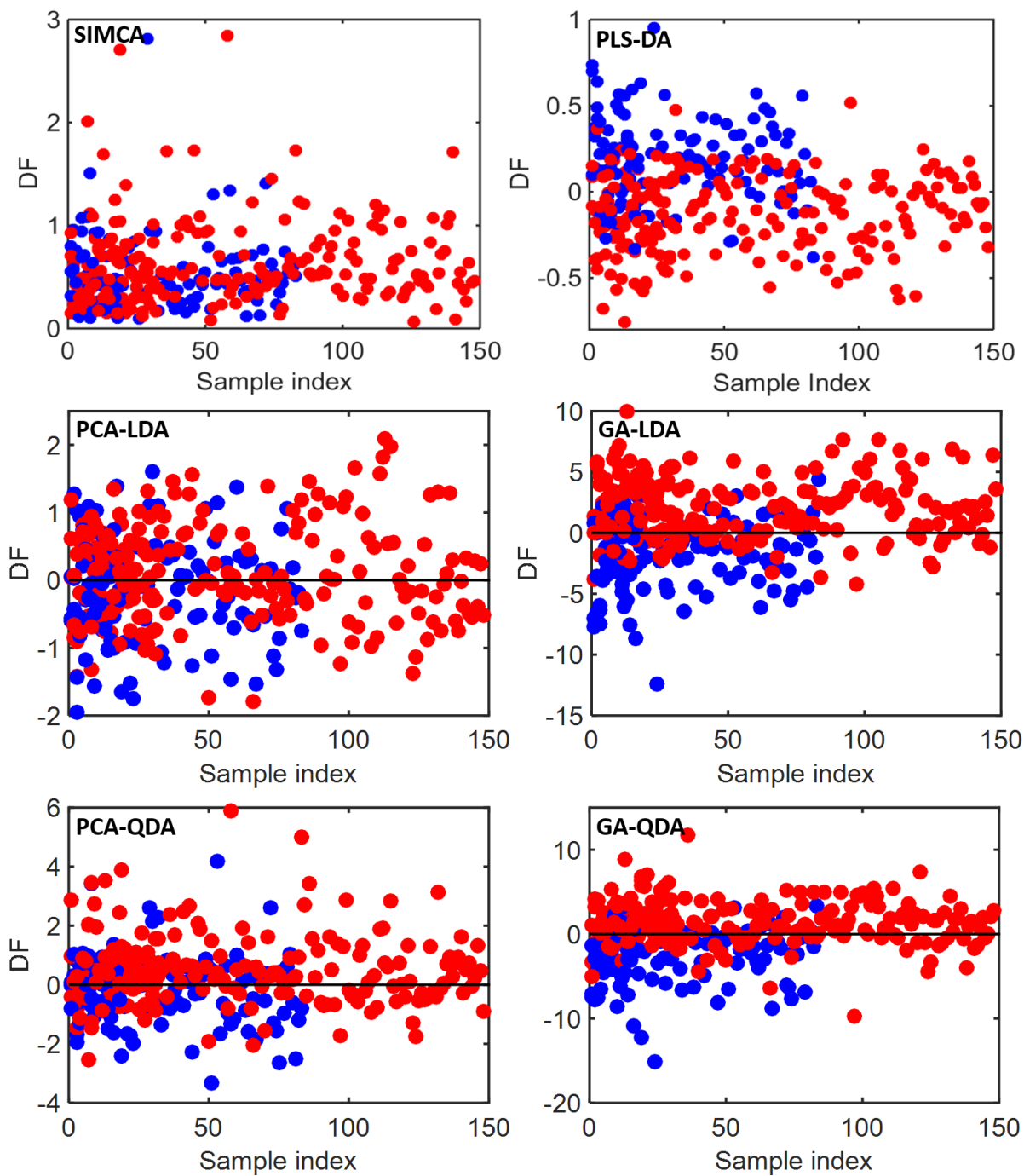
524



525

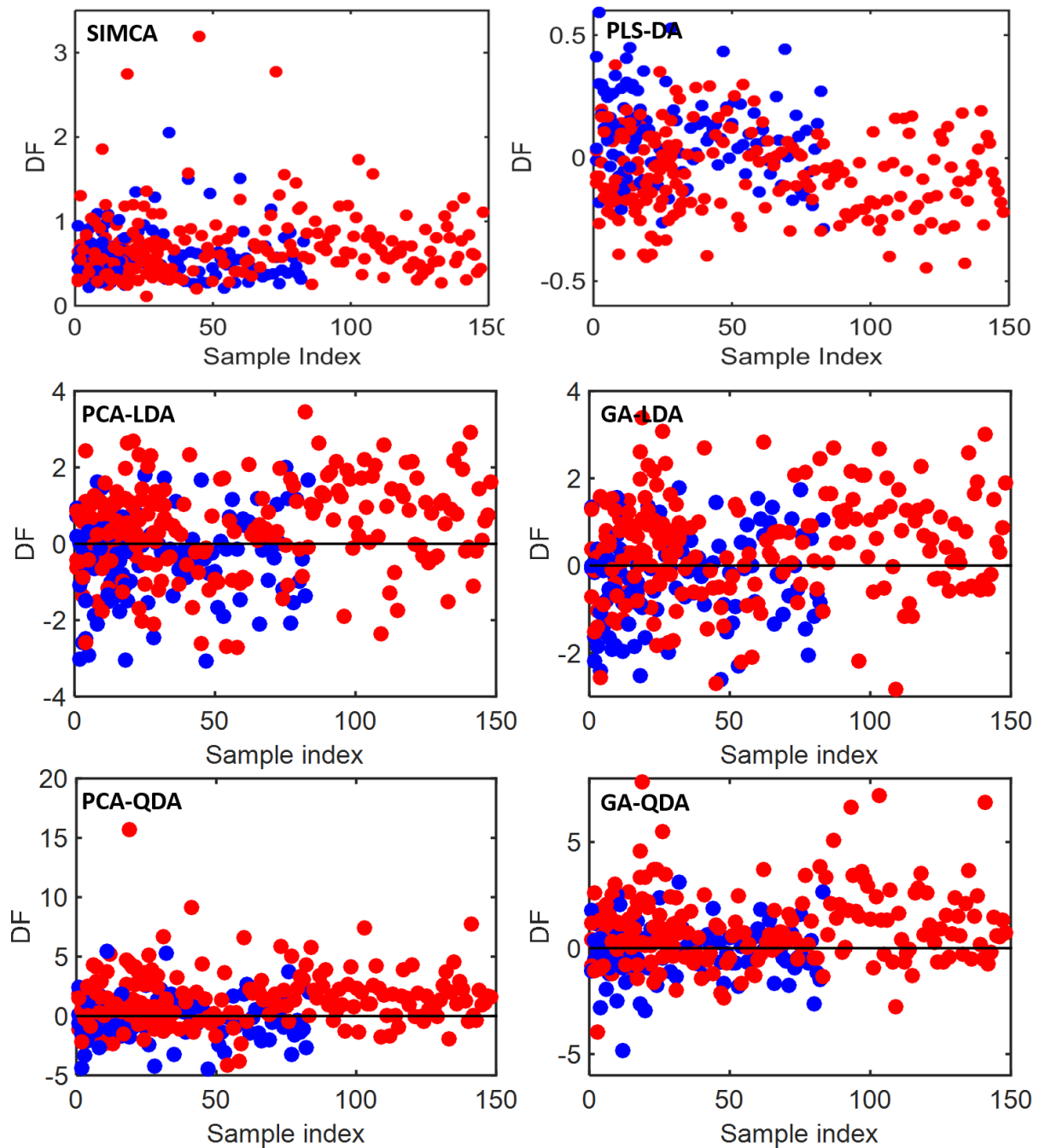
526 **Figure 1.** Near infrared spectra ($-\log_{10}R$) of all popcorn kernel samples (A), average spectra for
527 each original class of popcorn sample (B). Class 1 ($PE \leq 30 \text{ mL.g}^{-1}$) and Class 2 ($PE > 30 \text{ mL.g}^{-1}$).
528 ¹).

529



530
 531 **Figure 2.** Discriminant function (DF) *versus* samples calculated by using SIMCA, PLS-DA,
 532 PCA-LDA, PCA-QDA, GA-LDA, and GA-QDA models from two classes of popping expansion
 533 (PE). Red dots, class 1 ($PE \leq 30 \text{ mL.g}^{-1}$) and blue dots, class 2 ($PE > 30 \text{ mL.g}^{-1}$) without NIR
 534 spectra pre-processing.

535



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 ($PE \leq 30 \text{ mL.g}^{-1}$) and blue dots, class 2 ($PE > 30 \text{ mL.g}^{-1}$) with NIR
 541 spectra pre-processed with SNV.