Classification of Individual Castor Seeds Using Digital Imaging and Multivariate Analysis

Welma T. S. Vilar,* Rayanne M. Aranha,^a Everaldo P. Medeiros^b and Márcio J. C. Pontes*,^a

^aLaboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Departamento de Química, Universidade Federal da Paraíba, 58051-970 João Pessoa-PB, Brazil
^bEmbrapa, Centro Nacional de Pesquisa de Algodão, 58428-095 Campina Grande-PB, Brazil

Este artigo apresenta um método baseado em imagem digital e análise multivariada para classificação de sementes de mamona com respeito ao tipo de cultivar. Para este propósito, dois grupos de sementes comumente empregados nas plantações brasileiras foram avaliados: cultivares BRS Nordestina e BRS Paraguacu (grupo I), cultivar BRS Energia e o genótipo CNPA 2009-7 (grupo II). Imagens destes dois grupos foram registradas usando uma webcam e a distribuição de frequência de índices de cores nos canais vermelho-verde-azul, matiz, saturação, intensidade e tons de cinza foi obtida. A análise discriminante pelos mínimos quadrados parciais (PLS-DA) e análise discriminante linear foram aplicadas separadamente para cada grupo de semente. Os melhores resultados foram obtidos usando o modelo PLS-DA, o qual classificou corretamente 97,5% e 98,8% das amostras de predição para o grupo I e II, respectivamente. O método proposto é simples, rápido, não destrutivo e de baixo custo.

This paper presents a method based on digital imaging and multivariate analysis for the classification of castor seeds with respect to the cultivar type. For this purpose, two seed groups most commonly employed on Brazilian plantations were evaluated: BRS Nordestina and BRS Paraguacu cultivars (group I) and BRS Energia cultivar and CNPA 2009-7 genotype (group II). Images of these two different seed groups were recorded from a webcam and the frequency distribution of color indexes in the red-green-blue (RGB), hue (H), saturation (S), intensity (I), and grayscale channels were obtained. Pattern recognition methods based on partial least squares-discriminant analysis (PLS-DA) and linear discriminant analysis (LDA) were applied separately to each seed group. The best results were obtained by using the PLS-DA model, which correctly classified 97.5% and 98.8% of the prediction samples for groups I and II, respectively. The proposed method is simple, fast, non-destructive and non-expensive.

Keywords: digital imaging, castor seeds, partial least square-discriminant analysis, linear discriminant analysis, successive projections algorithm

Introduction

The castor plant (Ricinus communis L.) belongs to the Euphorbiáceas family, which includes a large number of plants from tropical regions.1,2 The oil is the main product extracted from castor seed, which has a high ricinoleic acid content with levels in the range of 78.3-90.0% (m/m).3,4 This makes the oil soluble in alcohols with low molecular weight.5-7

Ricinoleic acid has three functional groups: primary carboxylic, unsaturated in C9, and hydroxyl in C12.8,9 These are used with products such as lubricants, pharmaceutical and polymeric products, synthetic fibers, and biodegradable plastics.10,11

The high price of castor oil, insufficient raw materials in the international market and growing demand for biodegradable products obtained from ricinoleic acid, have encouraged public and private companies to develop castor seed cultivars with high oil content. They are looking for high quality castor seeds with respect to germination, vigor, seeds sanity, and seeds free from

*e-mail: marciocelholo@quimica.ufpb.br
contamination of pathogens (fungi, bacteria, nematodes and insects).\textsuperscript{12}

To obtain seeds with high genetic quality, it is necessary to implement breeding programs that seek superior characteristics and classes of uniform and distinct genotypes.\textsuperscript{13,14} Castor seeds have different morphological characteristics: size, shape, color, mass, oil content, among others. Because of these differences, it is not possible to use simple measures to identify these characteristics in order to classify cultivars or genotypes.\textsuperscript{15}

As example, Figure 1 shows two distinct groups of castor seeds separated by their morphological similarities. Group I consists of the BRS Nordestina and BRS Paraguaçu cultivars. They are seeds of larger size, dark in color and have a high oil content (48% m/m). Group II (consisting of the BRS Energia cultivar and CNPA 2009-7 genotype) has seeds with different morphological characteristics from those shown in group I. The seeds of these cultivars are the most commonly used in Brazilian planting. Despite the visual similarity of the seeds in each group, as seen in Figure 1, these cultivars have different phenotypic characteristics, which impact the handling of their planting and the market value of the seeds.

![Figure 1. Castor seeds from (a) BRS Nordestina and BRS Paraguaçu cultivars (group I) and (b) BRS Energia and CNPA 2009-7 genotype (group II).](image)

When the seeds have similar morphological profiles, the identification of cultivars can be carried out by planting the seed and waiting for the germination and development of the plant for at least 30 days, when it can be identified. As an alternative, the use of molecular markers based on DNA analysis\textsuperscript{16-18} can be employed, but these techniques are destructive, time-consuming and cannot be easily employed for routine identifications.

A previous work\textsuperscript{19} proposed the use of near-infrared (NIR) diffuse reflectance spectroscopy and chemometrics to classify 150 samples of castor seeds with respect to cultivar type (BRS Nordestina and BRS Paraguaçu). Two classification methods were compared, namely soft independent modeling of class analogies (SIMCA) and partial least square-discriminant analysis (PLS-DA). The better results were obtained by using PLS-DA, which correctly classified all test samples. That work, however, still entailed the use of expensive instruments.

An alternative to NIR spectroscopy is the use of methods based on digital imaging, which involves simpler and cheaper equipment.

Mondo et al.\textsuperscript{20} analyzed the efficiency of the seed vigor image system (SVIS) evaluating priming protocols for lettuce seed. SVIS was also employed by Gomes-Junior et al.\textsuperscript{21} in order to analyze sweet corn.

Frequency distribution of color indexes in the red, green and blue (RGB), the hue, saturation and intensity (HSI) and the grayscale channels associated with multivariate analysis have been explored in a number of works.\textsuperscript{22-24} Diniz et al.\textsuperscript{22} classified tea samples with respect to geographical origin by using digital imaging and linear discriminant analysis. Ahmed et al.\textsuperscript{23} used digital images to classify crops and weeds in the field by using classification models based on support vector machines (SVM). Recently, Milanez and Pontes\textsuperscript{24} classified vegetable oil samples according to type and conservation state by using images obtained from a webcam and the frequency distribution of color indexes in the RGB, HSI and grayscale channels. Linear discriminant analysis (LDA) was applied to histogram data in order to build classification models on the basis of a reduced subset of variables.

In recent years, little has been investigated about the use of digital images and multivariate analysis for the purpose of seed classification.\textsuperscript{25-27} Dana and Ivo\textsuperscript{25} identified flax cultivars according to their commercial similarity by using principal component analysis (PCA) and high content analysis (HCA). Medina et al.\textsuperscript{26} also applied PCA and HCA to identify quinoa seeds according to their geographical origin. Pourreza et al.\textsuperscript{27} classified nine varieties of Iranian wheat seeds using LDA coupled to a stepwise algorithm.

It is important to emphasize that works involving the use of digital images with pattern recognition methods for castor seed classification have not been found in the literature.

In the present paper, a methodology is proposed based on digital imaging data and supervised pattern recognition techniques for the classification of individual castor seeds with respect to four cultivar types: BRS Nordestina, BRS Paraguaçu, BRS Energia and CNPA 2009-7. For this purpose, the frequency distribution of color indexes in the red (R), green (G), blue (B), hue (H), saturation (S), intensity (I), and grayscale channels were obtained from digital images. Classification models based on PLS-DA and LDA were built and compared in terms of the correct classification rate (CCR) for the prediction set.
Experimental

Materials and methods

Samples

A total of 400 castor seed samples from different cultivars (group I: BRS Nordestina: 100 and BRS Paraguaçu: 100; group II: BRS Energia: 100 and CNPA 2009-7: 100) were collected from the Embrapa Algodão localized in Campina Grande, Paraíba, Brazil. In order to have a representative data set, samples were collected at different periods of the year. The samples were stored with a temperature controlled at 23 ± 1 °C and relative humidity of 65%.

Apparatus

Figure 2 shows the system used for the acquisition images of the castor seed samples. These comprise a (a) compartment, (b) fluorescent lamp, (c) webcam, (d) Teflon® support, and (e) notebook.

A 30 cm × 22 cm × 23 cm open-sided box was built in order to facilitate the control of the light striking the seeds, ensuring quality and uniformity for the image acquisition. The box was open at the top, but a sheet of office paper was placed over it to diffuse the light coming from a spiral, 6 W, 4000 K white colored fluorescent lamp which was placed 16.0 cm above the samples in the box. The inside of the back of the box was also lined with white office paper. The front of the box had a sliding door, to facilitate seed placement and removal. During the photo shooting, this door remained closed to prevent interference with the experimental lighting system.

Inside the box, a Teflon® cell (diameter and length of 4.7 cm and 10.5 cm, respectively) was placed to support the seed samples, together with a Microsoft webcam with HD resolution (1280 × 720). The webcam was positioned at a distance of 2.5 cm from the seed samples.

Digital image acquisition

Five sequential images for each castor seed sample were recorded and stored in JPEG format. The recorded images contain 24-bit (16.7 million colors) and a spatial resolution of 2880 × 1620 pixels. A region with an ellipse format at the center of each image was maintained constant throughout the analysis. Using only the selected region of the images, the frequency distributions (histogram) of color indexes according to each color channel were obtained for each of the five images. Then, the average of the five histograms was calculated, for use in all chemometric procedures. It is important to mention that, to obtain these histograms, three models for the color of a pixel were used in this study, namely: RGB standard, HSI and grayscale system. Each color component of the models is composed of 256 tones, varying from 0 to 255 for each channel.

Chemometric procedure and software

The data matrix is formed by samples located in rows, while the columns represent the color levels obtained for each color component.

The analytical information contained in the histograms was used separately for each seed group. In the first case, models were built in order to identify the BRS Nordestina and BRS Paraguaçu cultivars, which belonged to group I. After that, the discrimination between the BRS Energia cultivars and the CNPA 2009-7 genotype (belonging to group II) was performed.

Raw histograms and some pre-processing strategies such as auto-scaling and standard normal variate (SNV) were evaluated in terms of overall classification errors.

PCA was applied to the overall data set to observe the natural groupings of the castor seed samples, in an exploratory analysis.

The Kennard-Stone (KS) algorithm was employed to divide the data set into training (60%), and test (40%) subsets. This procedure was applied separately to each cultivar type. The training set was used to calibrate the PLS-DA and LDA/ successive projections algorithm (SPA) models, whereas test samples were only used in the final stage to evaluate the true predictive ability of the calibrated model.

The threshold values adopted for PLS-DA models were calculated on the basis of the Bayes theorem. As described by Ballabio, the threshold is selected at the point where the number of false positives and false negatives is minimized. Leave-one-out cross validation was employed to determine the optimal number of factors in the PLS-DA models.
The use of LDA for classification of the high-dimensional data usually requires appropriate variable selection procedures. In the present paper, the SPA\textsuperscript{32,33} based on the criterion recently proposed by Soares et al.\textsuperscript{34} is employed as variable selection tool for the LDA models.

All calculations were carried out by using MATLAB\textsuperscript{®} 2010a software.

**Results and Discussion**

**Histograms**

Figure 3 shows the average histograms for the castor seed samples acquired with color indexes in the gray scale, red, green, blue, hue, saturation and intensity channels. It is important to mention that each color component of the models is composed of 256 tones, varying from 0 to 255 for each channel. Variables that had a response equal to zero were removed from the data set.

As can be seen in Figure 3a and 3b, samples belonging to group I have a different histogram profile when compared with samples of group II. In fact, seeds of these two groups have a different physical appearance (Figure 2). In contrast, when the samples within the group are compared, it becomes difficult to distinguish them based on a visual inspection of histograms and seeds. For this reason, the multivariate methods were applied separately to each seed group.

**Principal component analysis**

In order to observe the existence of natural groupings of castor seeds types, an exploratory analysis of data was performed. For this purpose, PCA was applied separately for each histogram data set (group I and group II). Figure 4a presents score plots resulting from the application of PCA to histograms of BRS Nordestina and BRS Paraguacu castor seeds. As can be seen in Figure 4a, the BRS Nordestina class can be distinguished from the BRS Paraguacu class along PC1 (42%) and PC2 (12%) axes. Figure 4b presents the score plot resulting from the application of PCA to histograms of BRS Energia and CNPA 2009-7 cultivars. In this case, a trend of separation of these two cultivars is found along the PC1 axis. More specifically, CNPA 2009-7 samples presented more positive score values along the PC1 axis, when compared with BRS Energia samples. However, an overlapping of some samples is still found along the PC1 (28%) and PC2 (23%) axes.

**Classification models**

In order to verify the discriminatory capacity of the different channels, classification models based on PLS-DA and LDA with variable selection by SPA were developed using individual and combined channels for each seed group. The classification performances of these methods are evaluated in terms of CCR obtained in the training set. The best results were achieved with the raw histogram (without preprocessing). Table 1 summarizes the results, both for group I and group II.

For both seed groups, the best results were found using the PLS-DA models. More specifically for group I, the RGB-HSI channels were those which achieved the highest CCR when PLS-DA was applied. For LDA, however, the combination of all channels (RGB, HSI and grayscale) resulted in a smaller number of errors.

In the case of group II, the PLS-DA model correctly classified 108 out of the 120 training samples also using all the channels. By using the LDA/SPA method, the best
performance was achieved by using the red channel. Thus, classification models built with those channels that provided the best results for the training set were, then, applied to the prediction set.

All images of misclassified samples were evaluated, but it was not possible to identify the presence of abnormalities based on a visual inspection. This reinforces the need for the use of multivariate methods.

Figure 5 presents the average histograms with selected variables by SPA (for both seed groups). In the first group (Figure 5a), only three variables were selected. Such variables are located in the gray and hue channels. In fact, in these channels, a higher discrimination between the histograms of BRS Nordestina and BRS Paraguaçu classes is found (as shown in Figure 2). In the case of group II, seven variables were selected by the SPA algorithm for building the models. These variables are located along the red channel.

Figure 6 shows the plot of the Fisher discriminant scores obtained by using the variable selected by SPA. It is worth

Table 1. Correct classification rates obtained with LDA/SPA and PLS/DA models for the training set

<table>
<thead>
<tr>
<th>Channel</th>
<th>Group I LDA/SPA</th>
<th>Group I PLS/DA</th>
<th>Group II LDA/SPA</th>
<th>Group II PLS/DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gray-RGB-HSI</td>
<td>96.7 (3)</td>
<td>96.7 (4)</td>
<td>78.3 (6)</td>
<td>90.0 (5)</td>
</tr>
<tr>
<td>Gray</td>
<td>84.2 (6)</td>
<td>81.7 (3)</td>
<td>83.3 (5)</td>
<td>81.7 (3)</td>
</tr>
<tr>
<td>RGB</td>
<td>85.0 (4)</td>
<td>88.3 (9)</td>
<td>79.2 (7)</td>
<td>85.0 (9)</td>
</tr>
<tr>
<td>HSI</td>
<td>91.7 (3)</td>
<td>95.8 (11)</td>
<td>77.5 (5)</td>
<td>85.8 (4)</td>
</tr>
<tr>
<td>Gray-RGB</td>
<td>86.7 (4)</td>
<td>87.5 (8)</td>
<td>82.5 (5)</td>
<td>85.0 (3)</td>
</tr>
<tr>
<td>Gray-HSI</td>
<td>93.3 (3)</td>
<td>95.8 (5)</td>
<td>79.2 (4)</td>
<td>87.5 (5)</td>
</tr>
<tr>
<td>RGB-HSI</td>
<td>94.2 (3)</td>
<td>98.3 (5)</td>
<td>76.7 (6)</td>
<td>88.3 (5)</td>
</tr>
<tr>
<td>Red</td>
<td>82.5 (2)</td>
<td>85.0 (2)</td>
<td>85.8 (7)</td>
<td>86.7 (2)</td>
</tr>
<tr>
<td>Green</td>
<td>84.2 (6)</td>
<td>82.5 (4)</td>
<td>75.0 (4)</td>
<td>82.5 (5)</td>
</tr>
<tr>
<td>Blue</td>
<td>88.3 (8)</td>
<td>83.3 (2)</td>
<td>75.0 (5)</td>
<td>72.5 (5)</td>
</tr>
<tr>
<td>Hue</td>
<td>95.0 (2)</td>
<td>95.8 (17)</td>
<td>75.8 (5)</td>
<td>84.2 (13)</td>
</tr>
<tr>
<td>Saturation</td>
<td>94.2 (3)</td>
<td>95.8 (9)</td>
<td>80.3 (9)</td>
<td>80.0 (11)</td>
</tr>
<tr>
<td>Intensity</td>
<td>84.2 (4)</td>
<td>85.0 (4)</td>
<td>85.3 (8)</td>
<td>85.0 (2)</td>
</tr>
</tbody>
</table>

**Figure 4.** PC1 × PC2 score plots for the overall set of castor seed samples of (a) group I and (b) group II. ●: BRS Nordestina; ■: BRS Paraguaçu; ♦: BRS Energia and ▲: CNPA 2009-7.
mentioning that an LDA model generates a number of discriminant functions equal to the number of classes under consideration minus one. Thus, in the present study, only one discriminant function (DF1) was generated for the two classes of each seed group. As can be seen, the separation between the classes is more apparent when compared with the PC score plots as presented in Figures 4a and 4b.

Table 2 presents the detailed classification results obtained by the LDA/SPA and PLS-DA models in the prediction set. This table expresses both correct classifications (predicted class equal to correct class) and incorrect classifications (predicted class different from correct class).

As can be seen in Table 2, the LDA/SPA and PLS-DA models presented satisfactory classification performances when applied to prediction sample set. In both groups, the PLS-DA method achieved a correct classification rate slightly higher than the LDA/SPA method. In particular, for group II, only one sample belonging to class BRS Energia was classified as belonging to class CNPA 2009-7, when the PLS-DA method was employed. This outcome corresponds to a correct prediction rate of 98.8%.

**Conclusions**

This paper presented a method based on digital imaging and non-supervised and supervised pattern recognition techniques for the classification of individual castor seeds from BRS Nordestina, BRS Paraguaçu, BRS Energia cultivars and CNPA 2009-7 genotype.
Table 2. Prediction results obtained for the LDA/SPA and PLS/DA models in group I and group II. The correct classification rate obtained by each model is indicated in parenthesis. N indicates the number of prediction samples employed in this study.

<table>
<thead>
<tr>
<th>Group I</th>
<th>LDA/SPA (96.2%)</th>
<th>Predicted class</th>
<th>PLS/DA (97.5%)</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>True class</td>
<td>N</td>
<td>BRS Nordestina</td>
<td>BRS Paraguaçu</td>
<td>BRS Nordestina</td>
</tr>
<tr>
<td>BRS Nordestina</td>
<td>40</td>
<td>40</td>
<td>–</td>
<td>40</td>
</tr>
<tr>
<td>BRS Paraguaçu</td>
<td>40</td>
<td>3</td>
<td>37</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group II</th>
<th>LDA/SPA (95.0%)</th>
<th>Predicted class</th>
<th>PLS/DA (98.8%)</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>True class</td>
<td>N</td>
<td>BRS Energia</td>
<td>CNPA 2009-7</td>
<td>BRS Energia</td>
</tr>
<tr>
<td>BRS Energia</td>
<td>40</td>
<td>37</td>
<td>3</td>
<td>39</td>
</tr>
<tr>
<td>CNPA 2009-7</td>
<td>40</td>
<td>1</td>
<td>39</td>
<td>0</td>
</tr>
</tbody>
</table>

A PCA study applied to the histogram data revealed a trend of separation of two cultivars from each seed group evaluated. When supervised recognition methods were used, the better results were obtained from the PLS-DA model, which correctly classified 97.5% and 98.8% of prediction samples for groups I and II, respectively.

The proposed system is simple, non-destructive, fast and non-expensive and can be adopted both by inspection agencies to identify fraud and mixtures of these four types of castor seeds, as well as by companies that sell seeds and/or are involved in genetic improvement programs. It is important to mention that the classification procedure is data-driven. Thus, the results cannot be easily generalized to the analysis of other seed types that were not included in the study.

Acknowledgments

The authors acknowledge the support of CNPq (475087/2012-7, 304530/2012-3 and 151686/2012-2). The English version of this text has been revised by Sidney Pratt, BA, MAT (The Johns Hopkins University), RSAdip (TEFL).

Reference

14. Xu, W.; Cui, Q.; Li, F.; Liu, A.; Plos One 2013, 8, 1.
20. Mondo, V. H. V.; Gomes-Júnior, F. G.; Cicero, S. M.; Bennett, M. A.; McDonald, M. B.; Agrarian 2013, 6, 402.
32. Silva, A. C.; Pontes, L. F. B. L.; Pimentel; M. F.; Pontes, M. J. C.; Talanta 2012, 95, 129.

Submitted: June 16, 2014
Published online: September 26, 2014