LWT - Food Science and Technology 80 (2017) 76–83

Contents lists available at ScienceDirect

LWT - Food Science and Technology

journal homepage: www.elsevier.com/locate/lwt

# Analysis of bee pollen constituents from different Brazilian regions: Quantification by NIR spectroscopy and PLS regression

Maria Cristina A. Costa [a, *], Marcelo A. Morgano [b], Marcia Miguel C. Ferreira [a], Raquel F. Milani [b]

[a] Theoretical and Applied Chemometrics Laboratory (LQTA), Institute of Chemistry, University of Campinas - Unicamp, P.O. Box 6154, 13084-971, Campinas, SP, Brazil

[b] Food Science and Quality Center (CCQA), Food Technology Institute (ITAL), Av. Brasil 2880, P.O. Box 139, 13070-178, Campinas, SP, Brazil

ARTICLE INFO

*Article history:*
Received 12 July 2016
Received in revised form
5 December 2016
Accepted 3 February 2017
Available online 7 February 2017

*Keywords:*
Chemometrics
Diffuse reflectance
Protein
Fructose
Ash

ABSTRACT

In the present work partial least square regression (PLS) models were built for quantification of the major components of 154 Brazilian bee pollen samples. Bee pollen has nutritive and therapeutic properties that make it attractive for human health. However, studies on the nutrient and bioactive compound composition of this product are needed, as well as the verification of the presence of contaminants that are harmful to health. The conventional analysis methods are costly and time-consuming, while near infrared spectroscopy (NIR) associated to PLS regression allows a fast and non-costly quantification of the bee pollen components without samples pre-treatment.

The calibration models exhibited the determination coefficients, $R^2 > 0.94$. The mean percent calibration error varied from 1.49 to 5.58%. For external validation, $R^2$ ranged from 0.89 to 0.98 among the six. The results indicated that some models are good for quantification, while others are qualified for screening calibration.



## 1. Introduction

Bee pollen has nutritive and therapeutic properties that make it attractive for human health (Serra Bonvehí, & Escolà Jordà, 1997). It is a known and natural therapeutical agent, for which the chemical composition depends strongly on the plant source and geographic origin, among other factors. It is used as apitherapeutic treatment, presenting pharmacological activity such as antifungal, antimicrobial, antiviral, anti-inflammatory, anticancer immunostimulating and local analgesic. In the composition of the bee pollen, there are about 250 substances including protein, free amino acids, lipids (triglycerides, phospholipids), vitamins, macro- and micronutrients, and flavonoids (Komosinska-Vassev, Olczyk, Kafmierczak, Mencner, & Olczyk, 2015; Ribeiro & Silva, 2007). Because of the high nutritional value, its consumption as nutritional supplement for the human diet has increased in the last years (Szczesna, 2007) and besides medicine and nutritional applications, it is also utilized as ingredient in cosmetics such as sunscreens, creams, masks, lipsticks, soaps, shampoos, etc (Ribeiro & Silva, 2007).

Pollen is collected by bees from plant anthers, mixed with secretion from salivary glands or nectar, and placed in specific baskets situated on the tibia of their hind legs, then transported to the hive, where it is packed in honeycomb cells, covered with honey and wax. This created substance, the bee bread, constitutes the basic protein source for the colony (Komosinska-Vassev et al., 2015; Ribeiro & Silva, 2007).

With its extensive territory, diversified flora, and favorable climate throughout the year, Brazil has a great potential for apiculture. However, more studies are needed on the nutrient and bioactive compound composition of this product to promote its greater production and commercialization in domestic and international markets, as verify the presence of contaminants that are harmful to health (Morgano et al., 2010). Minerals content in pollen are usually expressed as ash content (Kostic et al., 2015), or as the content of macro and microelements (Yang et al., 2013) which includes not only important minerals to good nutrition, but also toxic heavy metals (Kostic et al., 2015; Morgano et al., 2010) The ash content depends on the soil type, geographical origin, floral species

* Corresponding author.
*E-mail addresses:* cristina.costa@iqm.unicamp.br (M.C.A. Costa), morgano@ital.sp.gov.br (M.A. Morgano), marcia@iqm.unicamp.br (M.M.C. Ferreira), raquel.milani@ital.sp.gov.br (R.F. Milani).

http://dx.doi.org/10.1016/j.lwt.2017.02.003
0023-6438/

and the plant's capacity to accumulate minerals in pollen (Serra Bonvehi, Gonell Galindo, & Gomez Pajuelo, 1986). However, the presence of mineral impurities due to inefficient cleaning procedures, may increase the ash content, which makes this analysis an important quality index for pollen (Baldi, Grasso, Pereira & Fernández, 2004).

Martins, Morgano, Vicente, Baggio, and Rodriguez-Amaya (2011) in a previous work, carried out the quantification of the Brazilian bee pollen components (ash, lipid, protein, glucose, fructose and free acidity) for samples collected in twelve different locations (Table 1). They discussed the results according to the regulatory quality standards and nutritive value for the human diet. Among the 154 analyzed samples, only one (from Sergipe) presented the ash level slightly above the limit of 4% (m/m) established by Brazilian and Argentinean regulations. Nevertheless, Campos et al. (2008) suggested a maximum value of 6% m/m. Some other countries define ash content in the range from 2 to 6 g per 100 g of pollen as quality requirement (Kostic et al., 2015). The results for lipid, protein, fructose and glucose were within the requirement of the Brazilian regulations (Martins et al., 2011; MAPA, 2001). Bee pollen is naturally acidic, with pH ranging from 4 to 6. Brazilian regulations stipulate a maximum limit of 300 meq/kg (MAPA, 2001). Half of the analyzed samples surpassed this limit, but some authors and regulations do not recommend this analysis as a quality index of the product (Campos et al., 2008; Codigo Alimentario Argentino, 1998).

The constituents of the bee pollen samples were quantified through several wet chemistry techniques for fructose and glucose, potentiometric titration for free acidity, and others detailed in the Methodology section. However these methods are longstanding, sometimes requires expensive equipments and consume many toxic chemical reagents. In order to reduce or minimize such disadvantages, methods based on near infrared spectroscopy (NIR) have been widely used in the determination and quantification of constituents of food and agricultural samples (Bagchi, Sharma, & Chattopadhyay, 2016; Pedro & Ferreira, 2007; Teye et al., 2015; Viegas, Mata, Duarte, Kássio, & Lima, 2016). Among the techniques for employing the NIR spectral region, diffuse reflectance is one of the most used (Hooton, 1978). Over the last decades, near-infrared diffuse reflectance spectroscopy (NIR) has become a premier method for the rapid analysis of agricultural products (Calderon et al., 2007; Costa et al., 2016). NIR spectroscopy is fast and nondestructive; however, the overtone and combination bands seen in the near-IR spectra are typically very broad, leading to complex spectra. Multivariate calibration techniques such as partial least square regression (PLS) are often employed to extract the desired chemical information (Luo, Wu, Wang, Lin, & Li, 2013; Rambo, Amorim, & Ferreira, 2013).

The aim of this work was to propose a fast, clean and no-costly procedure for the quantification of the Brazilian bee pollen components in samples from twelve different regions, using NIR spectroscopy combined with PLS regression. These twelve different regions were selected because they are the major pollen productive regions in the country.

## 2. Methodology

### 2.1. Samples

A total of 154 samples of Brazilian bee pollen from eleven states and Federal District were studied (Table 1). The samples of dehydrated bee pollen ready to be marketed (dried, cleaned and packaged) were acquired directly from the apicultural producers. They were sent to the laboratory, usually in the same month of collection, mostly in plastic bags or some in glass containers, during the beekeeping period. The samples (200–330 g) were quartered and then ground in a refrigerated mill (M20, IKA Labortechnik, Staufen, Germany) and then sieved using 30-mesh sieves. Analyses were carried out in triplicate for free acidity, and in duplicate for acidity, ash, lipid, protein, glucose and fructose (Supporting Table 1).

#### 2.1.1. Free acidity

Free acidity was determined by potentiometric titration of the pollen samples with a 0.05 mol $L^{-1}$NaOH standard alkali solution, employing an automatic potentiometric titrator (785 DPM Titrino, Metrohm AG, Herisau, Switzerland), titration endpoint at pH 8.5 (Horwitz, 2006).

#### 2.1.2. Ash

A portion of 2 g of the samples were weighed in a porcelain crucible; then the samples were incinerated on a hotplate and in a muffle furnace (550 °C) for 6 h (Zenebon & Pascuet, 2005, p. 1018).

#### 2.1.3. Lipid

Total lipid was determined after acid (HCl) thermal hydrolysis of the 5 g of the samples. Petroleum ether solvent in a butt extractor

**Table 1**
Mean values of constituent's concentration in bee pollen samples by State and the experimental limits of detection for all constituents. Adapted from Martins et al. (2011).

| State[a] | N[b] | Statistic | Ash | Lipid | Protein | Glucose | Fructose | Free acidity |
|---|---|---|---|---|---|---|---|---|
| | | | (g/100 g) | | | | | (meq/kg) |
| BA | 37 | Mean ± SD | 3.01 ± 0.40 | 7.08 ± 0.75 | 19.03 ± 2.13 | 14.16 ± 3.04 | 19.41 ± 2.07 | 411.0 ± 110.7 |
| SC | 30 | Mean ± SD | 2.29 ± 0.42 | 8.01 ± 1.56 | 19.39 ± 3.23 | 16.06 ± 2.93 | 19.37 ± 1.76 | 264.7 ± 80.5 |
| SP | 23 | Mean ± SD | 2.36 ± 0.22 | 6.14 ± 1.52 | 20.03 ± 3.05 | 15.20 ± 2.32 | 18.81 ± 2.27 | 304.6 ± 71.7b |
| SE | 18 | Mean ± SD | 3.61 ± 0.32 | 6.89 ± 0.57 | 21.91 ± 2.52 | 12.02 ± 2.55 | 17.81 ± 1.26 | 429.0 ± 72.8 |
| PR | 10 | Mean ± SD | 2.55 ± 0.58 | 9.23 ± 1.67 | 18.68 ± 2.91 | 13.95 ± 3.36 | 17.37 ± 3.26 | 292.0 ± 97.8 |
| MG | 10 | Mean ± SD | 1.89 ± 0.29 | 7.09 ± 1.85 | 17.58 ± 3.53 | 17.24 ± 3.42 | 20.23 ± 3.13 | 218.2 ± 79.6 |
| ES | 10 | Mean ± SD | 1.93 ± 0.43 | 7.55 ± 0.76 | 19.90 ± 2.03 | 16.68 ± 3.25 | 20.05 ± 2.44 | 306.9 ± 119.6 |
| RS | 9 | Mean ± SD | 2.70 ± 0.24 | 8.30 ± 2.39 | 21.71 ± 3.20 | 15.43 ± 2.06 | 18.38 ± 1.57 | 263.2 ± 44.0 |
| DF | 3 | Mean ± SD | 2.77 ± 0.16 | 6.94 ± 0.27 | 21.27 ± 1.34 | 19.12 ± 0.35 | 20.87 ± 0.23 | 142.5 ± 8.6 |
| MT | 2 | Mean ± SD | 2.91 ± 0.06 | 7.13 ± 0.25 | 20.09 ± 0.48 | 8.77 ± 1.59 | 14.45 ± 2.63 | 444.8 ± 16.5 |
| PI | 1 | Mean ± SD | 3.37 ± 0.01 | 7.78 ± 0.04 | 20.61 ± 0.11 | 13.94 ± 0.09 | 18.57 ± 0.21 | 326.7 ± 18.3 |
| CE | 1 | Mean ± SD | 3.39 ± 0.04 | 6.44 ± 0.05b | 21.59 ± 0.11 | 14.84 ± 0.35 | 20.12 ± 0.16 | 382.3 ± 1.1 |
| Total | 154 | Overall mean | 2.65 ± 0.62 | 7.34 ± 1.52 | 19.57 ± 3.00 | 14.89 ± 3.22 | 18.99 ± 2.28 | 327.6 ± 114.4 |
| ***LOD*** | | | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.3 |

*LOD*: experimental limit of detection.
[a] BA (Bahia); SC (Santa Catarina); SE (Sergipe); MG (Minas Gerais); ES (Espírito Santo); RS (Rio Grande do Sul); DF (Distrito Federal); MT (Mato Grosso); PI (Piauí); CE (Ceará).
[b] Number of samples per State.

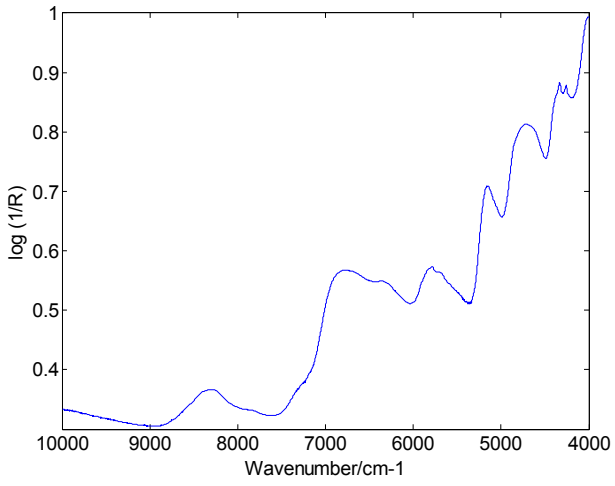


**Fig. 1.** Generic NIR spectrum of pollen samples.

was used to extracted lipids. After removal of the solvent (ether) the lipid content was determined by weight difference (Zenebon & Pascuet, 2005, p. 1018).

### 2.1.4. Protein

The protein content was determined by Kjeldahl method. Nitrogen content was determined through acid hydrolysis and distillation of 0.5 g of sample. Ammonia was distilled and collected in boric acid solution, and later titrated with a standard solution of hydrochloric acid (Zenebon & Pascuet, 2005, p. 1018). According Rabie, Wells, and Dent (1983) the factor 5.60 should be used for the conversion of nitrogen level to protein.

### 2.1.5. Glucose and frutose

After extraction of sugars, glucose and fructose were quantified by liquid chromatography (HPLC) according Burgner and Feinberg (1992). Portions of 2.5 g of samples were mixed with 25 mL of deionized water and stirred for 2 h. Solutions of zinc acetate and potassium ferricyanide were used to clarifying. The sugar solution was then filtered in a 0.45 μm diameter membrane and injected into a HPLC (model Pro Star, Varian, Mulgrave, Australia), equipped with a refractive index detector, model 350 RI, and a column, Luna $NH_2$, 250 × 4.6 mm, 5 μm (Phenomenex Inc., Torrance, USA). The analyses were carried out at 40 °C. The mobile phase was acetonitrile:water (85:15 v/v), the flow rate was 1 mL/min, and the injection volume was 20 μL.

## 2.2. Near infrared spectroscopy

The diffuse reflectance spectra in the near infrared region were acquired by using an Antaris II FT-NIR spectrometer (Thermo Fisher Scientific, Verona, USA) in the 4500–10,000 $cm^{-1}$ range. The spectra were generated by averaging sixteen successive scans with 4 $cm^{-1}$ nominal resolution (yielding 3112 wavenumbers). The signals were obtained in reflectance mode (% R) and transformed into absorbance by using a log transformation. Fig. 1 shows a generic spectrum. Three spectra were recorded for each sample, and the average spectrum was used for data analysis.

### 2.2.1. NIR spectroscopy data analysis

The mean spectra of the 154 samples were organized in a matrix format **X** (154, 3112) where each row corresponds to a sample and the columns correspond to the absorbance (log 1/R) values. Different mathematical transformations such as first and second derivatives, baseline correction, smoothing, standard normal variate (SNV) and multiplicative signal correction (MSC) were tested. The vector **y** of concentrations was correlated with spectral information through the partial least squares (PLS) regression method on the mean centered data using the Pirouette®4.5 software. Leave-one-out cross-validation was used to determine the number of factors (latent variables) in the calibration model. The presence of outliers was investigated analyzing the plot of leverage *vs.* Studentized residuals (Fig. S1–S3) and by the score plots using spectral data. The confidence ellipses in PCA analysis were constructed by the Mahalanobis distance (generalized statistical distance), $d_{AB}^{M} = [(\mathbf{X}_A - \mathbf{X}_B)^{T}\mathbf{Var}^{-1}(\mathbf{X}_A - \mathbf{X}_B)]^{1/2}$, where $\mathbf{Var}^{-1}$ is the inverse of variance-covariance matrix (Ferreira, 2015). After removing outliers, the data sets were randomly split into two subsets: training (calibration) and test sets. The final regression model was evaluated by analyzing the values of coefficient of determination for the training set ($R^2$ Cal), standard error of calibration (*SEC*), standard error of cross validation (*SECV*) and the relative error (*RE%*). External validation set was used to evaluate the prediction ability of the model through the standard error of prediction (*SEP*), the relative standard deviation (*RSD%* = *SEP* × 100/mean, where the mean is taken from reference values in the validation set), the range error ratio, *RER* (AACC, 1999), where *RER* = $\text{Range}_y$ of reference data/SEP and the coefficient of

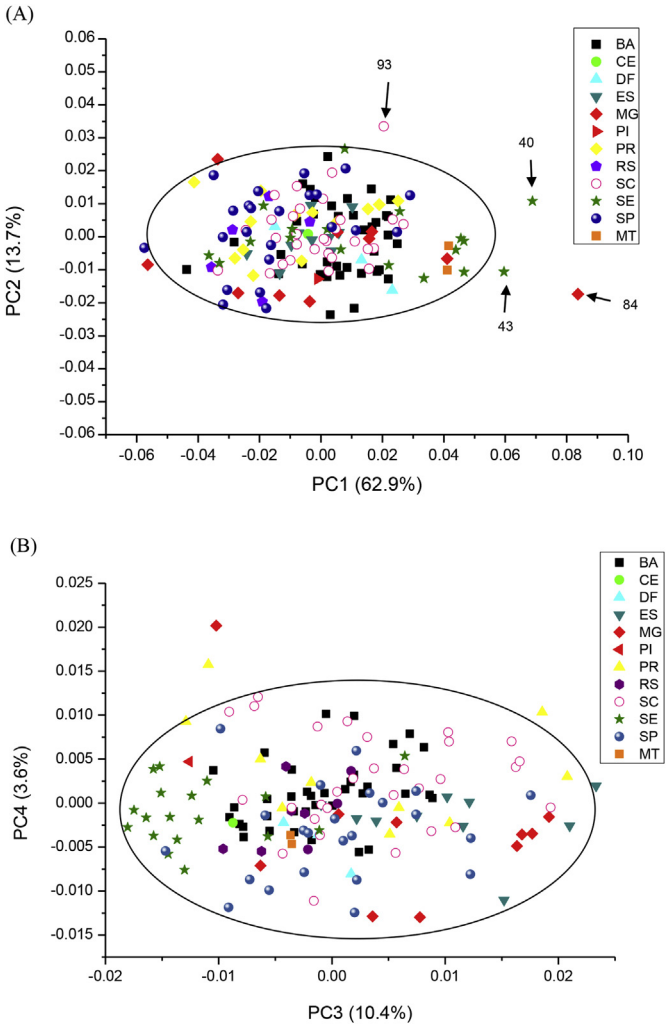


**Fig. 2.** (A) PC1 *vs.* PC2 and (B) PC3 vs. PC4 scores plot of meancentered NIR spectra of bee pollen samples collected at twelve different Brazilian States.

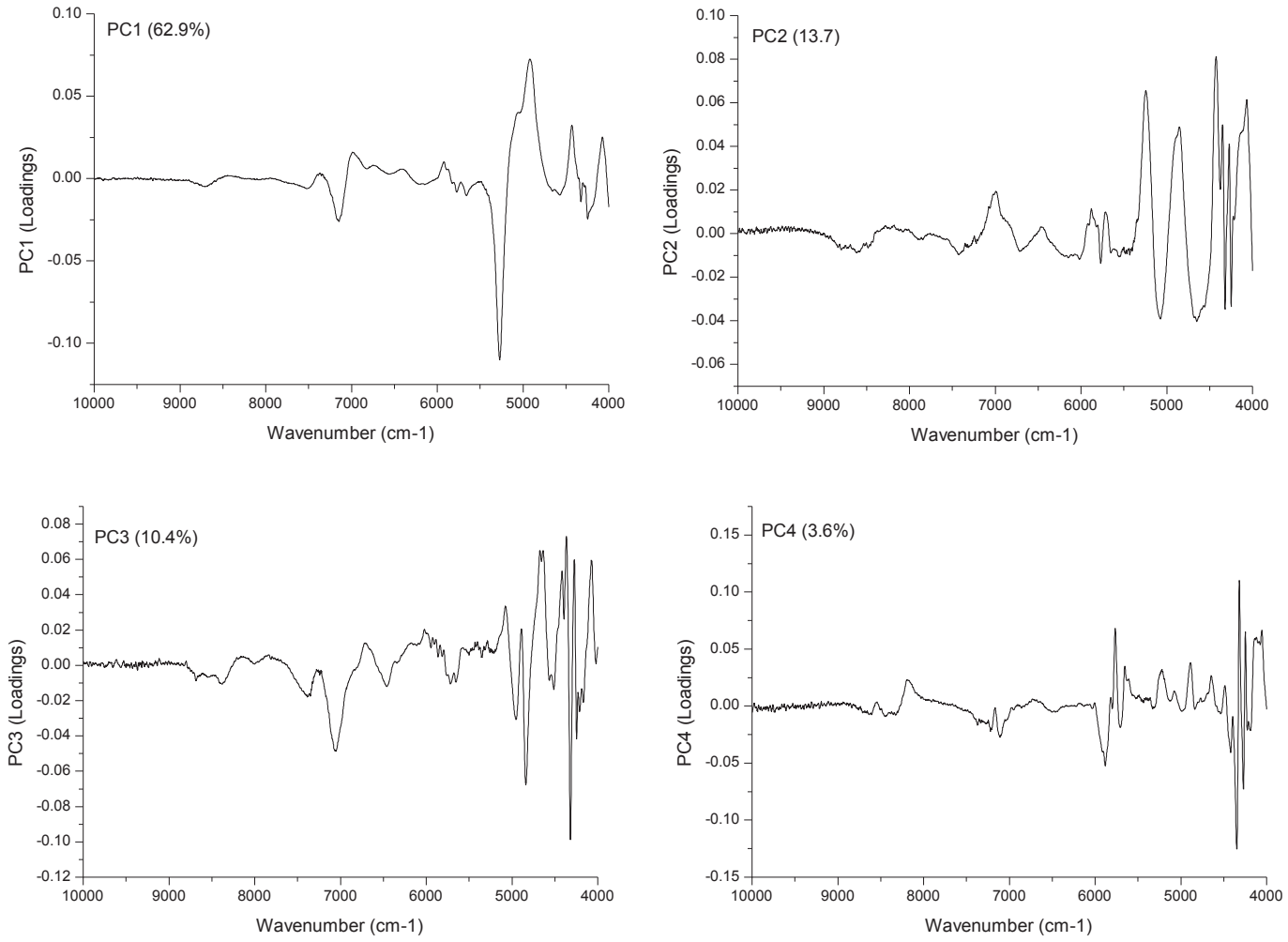


**Fig. 3.** Loadings plots from PCA analysis.

**Table 2**
Parameters for models evaluation and validation[a] of the best PLS models obtained.

| y | Training matrix size | Pre-treatment[b] | F | Out | $R^2$ | | | *SEC* | *SECV* | *SEP* | *RSD%* | *RER* | Mean *RE%* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Cal | Val | Pred | | | | | | Training | Test |
| ash | 108 × 3112 | S + 1D + MSC | 8 | 6 | 0.9635 | 0.9160 | 0.9088 | 0.12 | 0.18 | 0.18 | 6.81 | 12.74 | 3.76 | 4.50 |
| lipid | 109 × 3112 | 1D | 7 | 6 | 0.9708 | 0.9382 | 0.9534 | 0.29 | 0.40 | 0.31 | 4.41 | 9.56 | 3.08 | 4.40 |
| protein | 113 × 3112 | 1D + SNV | 5 | 3 | 0.9853 | 0.9602 | 0.9770 | 0.39 | 0.63 | 0.46 | 2.51 | 28.97 | 1.49 | 2.63 |
| glucose | 111 × 3112 | S + 1D | 9 | 12 | 0.9575 | 0.8855 | 0.9487 | 0.66 | 1.04 | 0.70 | 4.39 | 15.76 | 3.39 | 4.35 |
| fructose | 103 × 3112 | 1D | 8 | 13 | 0.9417 | 0.8696 | 0.8901 | 0.57 | 0.81 | 0.73 | 3.75 | 11.88 | 2.38 | 4.69 |
| Free acidity[a] | 114 × 3112 | 1D | 9 | 11 | 0.9646 | 0.9212 | 0.9621 | 25.63 | 24.48 | 27.85 | 9.23 | 12.43 | 5.58 | 5.79 |

Original matrix size 154 × 3112. y, dependent variable: analyzed bee polen constituents.
[a] For free acidity, dependent variable was transformed to build the model. The results are shown in original units.
[b] S: smoothing; 1D: first derivative; MSC: multiplicative signal correction; SNV: standard normal variate. F: number of factors in the model; Out: outliers.

determination for the test set ($R^2$ Pred).

The analytical quality of the models was evaluated by the determination of multivariate figures of merit, through the net analyte signal (NAS) (Bro & Andersen, 2003). The sensitivity (*SEN*), the inverse of analytical sensitivity ($\gamma^{-1}$), the selectivity (*SEL*) and the limits of detection (*LOD*) and quantification (*LOQ*) were estimated (Ferreira, 2015). The equations for the calculation of validation parameters are shown in Supporting Table 2.

## 3. Results and discussion

### 3.1. Exploratory data analysis

Prior to regression analysis, an exploratory analysis using the spectral data was performed in order to investigate the presence of outliers and any trend of discrimination among bee pollen samples from different Brazilian regions. The original matrix **X** data were transformed by SNV and first derivative. The scores plot of the two first principal components (PC1 vs PC2, in Fig. 2A) shows that some

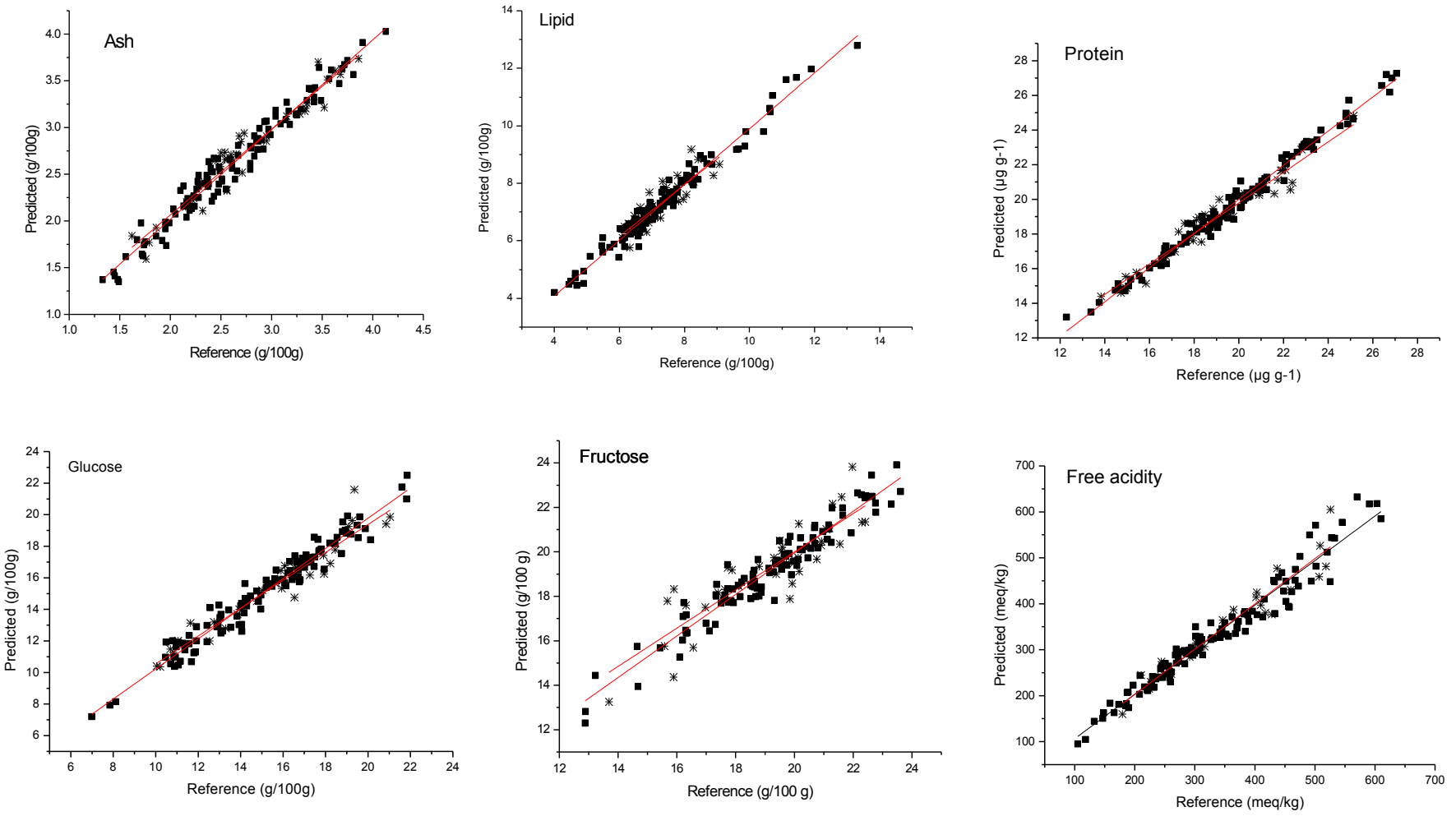


**Fig. 4.** Plots of reference *vs.* predicted values of calibration (■) and validation (*) sets for all PLS models built.

**Table 3**
Figures of merit for the PLS models built for each analyzed pollen constituent.

| | Free acidity | Ash | Fructose | Glucose | Lipid | Protein |
|---|---|---|---|---|---|---|
| *SEN* | 0.0031 | 9.0815e-04 | 1.9684e-04 | 1.3230e-04 | 4.7623e-04 | 0.6027 |
| $SEN_{NAS}$ | 0.0031 | 9.0796e-04 | 1.9682e-04 | 1.3135e-04 | 4.7635e-04 | 0.6027 |
| $\gamma^{-1}$ | 0.0068 | 0.0202 | 0.1262 | 0.1695 | 0.0595 | 0.8662 |
| $\gamma^{-1}{}_{NAS}$ | 0.0068 | 0.0202 | 0.1262 | 0.1708 | 0.0595 | 0.8662 |
| *SEL* | 0.0663 | 0.1139 | 0.0576 | 0.0579 | 0.0930 | 0.2085 |
| *LOD* | 0.0225 | 0.0664 | 0.4152 | 0.5577 | 0.1958 | 2.8499 |
| $LOD_{NAS}$ | 0.0225 | 0.0664 | 0.4152 | 0.5618 | 0.1958 | 2.8499 |
| *LOQ* | 6.7300e-07 | 1.6635e-07 | 4.8897e-08 | 2.9672e-08 | 1.3501e-07 | 3.1469 |
| $LOQ_{NAS}$ | 6.7307e-07 | 1.6632e-07 | 4.8892e-08 | 2.9458e-08 | 1.3504e-07 | 3.1469 |

*SEN*: sensitivity; NAS: net analyte signal; $\gamma^{-1}$: inverse of the analytical sensitivity; *SEL*: selectivity; *LOD*: limit of detection; *LOQ*: limit of quantification.

samples are outside the 95% confidence ellipse. According to Table 1, among the total of 154 bee pollen samples, 147 are distributed by eight Brazilian States; the seven remaining samples are from Distrito Federal (DF, 3 samples), Mato Grosso (MT, 2 samples), Piauí (PI, 1) and Ceará (CE, 1). Although PC3 and PC4 accumulate small percentage of the variance (10.4 and 3.6) and the majority of the samples are overlapped, this graphic shows a tendency to discriminate them by the States. To improve visualization, the sample number 138, which is outside the confidence ellipse, was removed (Fig. 2B) although it is not an outlier. Samples from Sergipe (SE) are grouped on the left of PC3 *vs* PC4 scores plot, being clearly discriminated from the others by their negative scores on PC3, as the unique sample from PI. Samples from RS, CE, MT and DF also present negative scores in PC3, but with lower values, while samples from Minas Gerais (MG) and Espírito Santo (ES) are displayed on the right side of the graphic (positive scores). PC3 loadings (Fig. 3) show high negative and positive values at approximately 4850 and 4650 $cm^{-1}$, due to combination bands of N-H from amides, present in protein, and combination bands of N-H (amines) + O-H where OH is from R-OH that can be attributed to glucose and fructose. A peak of minor intensity is present at approximately 7000 $cm^{-1}$ due to the first overtone of N-H and O-H from amides/amines and R-OH, respectively. These observations indicate that the samples are discriminated by the amount of proteins, glucose and fructose. In fact, bee pollen samples collected in the locations that appear to the left of the graphic have high protein content (Table 1), while Minas Gerais (MG) and Espírito Santo (ES) contain samples with lower values of protein. The samples from MG and ES are also among those with higher glucose and fructose contents (Table 1). The mentioned peaks related to amines, amides and the R-OH signals from glucose and fructose are not present in PC4 loadings. On the other hand, PC4 loadings graphic shows two peaks at 5760 and 5880 $cm^{-1}$ due to C-H first overtone, which are not present in PC3. The highest values of loadings in PC4 and PC3 are in the range from about 4070 to 4400 $cm^{-1}$, typical vibrational bands of combination from C-H and C-C present in the structure of the constituents of pollen, and in fibers.

The majority of Santa Catarina (SC) and Bahia (BA) samples present positive scores in PC4, while for São Paulo (SP) and Minas Gerais (MG) the PC4 scores tend to be negative. Martins et al. (2011) found that samples from Bahia and Santa Catarina have the highest levels of free acidity and those from Minas Gerais had the lowest levels. The loadings of PC4 shows two small peaks at approximately 5100 and 5200 $cm^{-1}$ where the signals of C=O stretch and O-H combinations from carboxylic acids are present (Shenk, Workman, & Westerhaus, 2008; Workman, 1996). These peaks are present in the graphics of the first four principal components (Fig. 3), but it is more strongly pronounced in the PC1 loadings. Although the discrimination of the samples is less noticeable on PC1 vs. PC2 graph, BA samples, which have the highest free acidity levels, have positive scores in the PC1, while most of the samples from SP and RS that exhibit lower levels of free acidity (Table 1) have negative scores on PC1. Based on PCA analysis one can have some idea about the region where the bee pollen has been collected. In order to confirm these observations, another PCA analysis was performed with the autoscaled concentrations of the constituents, instead of the spectral data (Fig. S4). The discrimination by the regions where the bee pollen samples were collected followed the same tendency observed previously by NIR spectroscopy data, with BA samples discriminated from MG and SP on the fourth principal component. The SE, MT, CE and PI samples were discriminated from ES and MG samples mainly in PC1. The loadings plots show that ash, protein and acidity have high positive values in PC1, while glucose and fructose have high negative loadings. In PC4, ash and protein are the main constituents responsible for the discrimination of the samples, where ash has the highest positive values, and protein presents negative loadings (Fig. S5).

The hierarchical clustering analysis (Fig. S6) shows five groups at approximately 0.7 similarity values. The first group, G1, is a large group composed mainly by samples from Bahia (BA) and Santa Catarina (SC) that have the highest levels of free acidity (Table 1). The samples from the other Brazilian States are present in small number. G1 is grouping the majority of BA samples (23 in a total of 37), one third of SC (11 in 30), the only one PI sample and almost all DF samples (2 in 3). G5, the largest of the groups, contains the majority of the samples from SE, RS and SP and half of the PR samples. The only sample from Ceará (CE) is also present in G5. The groups G2, G3 and G4 are smaller and are composed by a mixture of a few samples of the various states.

The results of hierarchical analysis considering the largest groups, G1 and G5, are similar to those obtained by discriminations of PC4 and PC1. Samples from BA and SC that contain the highest levels of acidity are grouped in G1 and tend to be discriminated from the samples of SP (G5), with lower levels of acidity, by PC4. The discrimination of these samples is also observed in the PC1 vs PC2 graph, where most of BA samples present positive scores in PC1 (on the right of the graph) in opposite to those from SP and RS (with lower levels of free acidity) with negative scores.

### 3.2. Partial least square regression modeling

PLS models were built for each one of the analyzed constituents of the bee pollen samples, which were ash, lipid, glucose, fructose, protein and free acidity. The fraction of outliers removed from the models varied from 2 to 8%. After removal of outliers, the models were built with the training sets that are the original data set from which some samples were excluded to comprise the test sets for external validation. The training matrix sizes for each model are shown in Table 2.

Among all pretreatments applied to the mean spectra, the smoothing, first derivative, multiplicative signal correction and

standard normal variate in different combinations for each model, produced the best results. The optimum number of factors for the models was defined by applying leave-one-out cross-validation during PLS modeling of the training sets. This number is indicated by the minimum in the plot of SECV *versus* number of factors. Table 2 shows the coefficients of determination for calibration and validation ($R^2$ *Cal* and $R^2$ *Val*) and the standard error of cross validation and calibration (*SECV* and *SEC*, respectively) for the final models.

The regression vectors for the models (Fig. S7) show that the concentration of the constituents increases with the increase of the negative coefficients whereas the data were transformed by the first derivative. All of the models present high negative coefficients between 4100 and 4500 $cm^{-1}$, which are correlated to the concentration of the studied components of bee pollen. The signals observed in this region of the spectra are due to combination bands of C-H and C-C stretching from CH, $CH_2$ and $CH_3$ (Shenk et al., 2008; Workman, 1996). For ash, besides the mentioned signals, a peak close to 6000 $cm^{-1}$ presents high contribution to the model and it is the first overtone region due to C-H stretching from $CH_3$ and ArCH. The regression vector for lipids also shows a negative band in this region (from 5890 to 5950 $cm^{-1}$), in addition to a peak of low intensity around 5200 $cm^{-1}$ due to O-H combination bands from esters.

For protein, the regression vector shows two high negative peaks, one between 4600 and 4700 $cm^{-1}$, and the other at approximately 4900 $cm^{-1}$. The first one corresponds to the region of combination bands from N-H and O-H stretching of amides and amines and the other is due to N-H stretching from amides. Both peaks are also negative in the correlation spectrum what is in agreement with the use of first derivative pre-processing data. The correlation spectrum presents another high negative peak between 5900 and 6000 $cm^{-1}$, the first overtone region of the C-H stretching vibrations from ArCH, found in aminoacids. This peak is also present in the regression vector, but with low intensity.

The main peaks related to fructose and glucose are the negatives around 4300, 4400 and 4800 $cm^{-1}$, where the last is due to the combination bands of O-H stretching in ROH bond of sugars.

In the free acidity model, a peak at approximately 5300 $cm^{-1}$ shows the contribution of C═O stretch of acids to build this model. The other two indicated peaks at approximately 5600 and 5800 $cm^{-1}$are in the region of the first overtone of C-H stretching from CH and $CH_2$.

Fig. 4 shows the reference vs. predicted values of calibration and validation sets for all PLS models built according to the conditions of Table 2. Linearity is observed in all plots, but the best fitted models are those for protein, lipid, ash and glucose. For fructose and free acidity it is observed some spreading of the data, but the reasonable agreement between the plots of reference vs. predicted values for calibration and external validation indicate that the final models can be used for an approximate prediction of new samples even for fructose and free acidity, what is confirmed by the relative standard deviation (*RSD*%) and RER values for these properties (Table 2).

The best model was obtained for protein, for which only three outliers (2%) were found and the validation parameters were the best, with the lowest *RSD*% (relative standard deviation) and highest *RER* (range error ratio). Free acidity was the most difficult property to model, probably due to the wide range, and a previous log transformation of the concentrations was necessary; eleven outliers were excluded from original matrix (7%) and nine factors were used in the model. This property presented the highest range of variation (105.3–609.9). After building the model for free acidity training set, the results were back transformed to original units. The *RSD*% for free acidity was the highest obtained (9.23%), but the *RER* > 10.0 indicated that the model is acceptable for quality control. For other models, *RSD* were lower than 5% and approximately 7% for ash. The parameter *RER* was slightly below 10 only for lipid. According to American Association of Cereal Chemists (AACC) – Method 39-00 (AACC, 1999) the model is acceptable for quality control when $RER \geq 10$, and if $RER \geq 15$ the model is good for research quantification. Among the six models built, two presented this parameter above 15 and the others equal or higher than 10, indicating the good quality of the models. The protein model presented the lowest mean relative errors (RE%) for both, training and test sets, while the highest were found for free acidity, as expected (Table 2). However, all of the models presented low percent relative errors ranging from 0.11 to 16.93 for acidity and from 0.01 to 7.87 for protein, as shown in the Supporting Tables 3–8.

### 3.3. Figures of merit

The reliability of the results can also be evaluated through the figures of merit (Table S2), such as sensitivity, selectivity, limit of detection and limit of quantification, which can be estimated by the net analyte signal (NAS) (Bro & Andersen, 2003). With the exception of the selectivity, all others can be calculated by the conventional method (Ferreira, 2015), and therefore, the results of both methods were included in Table 3. With respect to the values of the parameters in Table 3, it can be seen that they are greater for protein than for the others, especially for sensitivity (*SEN*). These figures of merit are influenced by the spectral pretreatment.

The inverse of analytical sensitivity ($\gamma^{-1}$) has a direct relationship with the concentration and indicates the smallest difference in concentration that can be distinguished by the method, being a slightly high in the case of protein. However, the minor concentration found for protein was 13.38 g/100 g, which is much higher than the estimated limits of detection and quantification. *LOD* and *LOQ* are much below of the minimal values for all analyzed constituents, indicating that NIR-PLS method is useful to detect and quantify ash, free acidity, protein, lipid, fructose and glucose in concentrations above these minima.

## 4. Conclusions

The results obtained for calibration and prediction parameters indicated that the models are validated and can be used for quantification of protein and glucose of new bee pollen samples. The models for ash, lipid, fructose and free acidity are indicated for quality control. The difficulty to model the free acidity was circumvented by using a log transformation of the reference data, which do not affected the analysis result since the predicted values were accounted with the untransformed data.

The exploratory analysis of the data indicated that it is possible to have some idea about the area of bee pollen collect, using only the spectral data of the samples.

Finally, this work showed that NIR spectroscopy associated to PLS regression can be successfully used for determination of the major components in bee pollen.

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgments

MAM acknowledges the Brazilian governmental agencies, CNPq (proc. n. 473108/2006-2) and FAPESP (proc. n. 59551-2/2006), for financial support.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.lwt.2017.02.003.

## References


AACC. (1999). *Near-infrared methods - guidelines for model development and maintenance* (St. Paul, MN).

Bagchi, T. B., Sharma, S., & Chattopadhyay, K. (2016). Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. *Food Chemistry, 191*, 21–27.

Baldi, C., Grasso, D., Pereira, S. C., & Fernández, G. (2004). Caracterización bromatológica del pólen apícola argentino. *Ciencia, Docencia y Tecnología, 29*, 145–181.

Bro, R., & Andersen, C. M. (2003). Theory of net analyte signal vectors in inverse regression. *Journal of Chemometrics, 17*, 646–652.

Burgner, E., & Feinberg, M. (1992). Determination of mono and disaccharides in foods by interlaboratory study: Quantitation of bias components for liquid chromatography. *Journal AOAC International, 75*, 443–464.

Calderon, F. J., Reeves, J. B., Foster, J. G., Clapham, W. M., Fedders, J. M., Vigil, M. F., et al. (2007). Comparison of diffuse reflectance fourier transform mid-infrared and near-infrared spectroscopy with grating-based near-infrared for the determination of fatty acids in forages. *Journal of Agricultural and Food Chemistry, 55*, 8302–8309.

Campos, M. G. R., Bogdanov, S., Almeida-Muradian, L. B., Szczęsna, T., Mancebo, Y., Christian, F., et al. (2008). Pollen composition and standardisation of analytical methods. *Journal of Apicultural Research, 47*, 154–161.

Codigo Alimentario Argentino. (1998). *Buenos aires.* Argentina: La Canal y Asociados.

Costa, M. C. A., Matallo, M. B., Ferreira, M. M. C., Queiroz, S. C. N., Almeida, S. D. B., & Franco, D. A. S. (2016). *Brachiaria plantaginea as a potential (new) source of shikimic acid*. Quantification by NIR and PLS regression. *Planta Medica Letters* 01: e1 https://www.thieme-connect.de/DOI/DOI?10.1055/s-0042-102202 (Last Accessed 16 April 2016).

Ferreira, M. M. C. (2015). *Quimiometria: Conceitos, Métodos e Aplicações* (Vol. 159, pp. 380–388). Campinas, Brazil: Editora da Unicamp.

Hooton, D. E. (1978). The versatility of near-infrared reflectance devices. *Cereal Food World, 23*, 176–179.

Horwitz, W. (2006). *Official methods of analysis of the association of official analytical Chemists* (18th ed., Vol. 44, p. 37). Gaithersburg, Maryland: AOAC. met. 962.19.

Komosinska-Vassev, K., Olczyk, P., Kafmierczak, J., Mencner, L., & Olczyk, K. (2015). Bee pollen: Chemical composition and therapeutic application. *Evidence-Based Complementary and Alternative Medicine*, 6. ID 297425.

Kostic, A. Z., Pesic, M. B., Mosic, M. D., Dojcinovic, B. P., Natic, M. M., & Trifkovic, J. D. (2015). Mineral content of bee pollen from Serbia. *Archives of Industrial Hygiene and Toxicology, 66*, 251–258.

Luo, W., Wu, J., Wang, X. K., Lin, X., & Li, H. (2013). Near infrared spectroscopy combination with PLS to monitor the parameters of naproxen tablet preparation process. *Analytical Methods, 5*, 1337–1345.

MAPA - Brazilian Ministry of Agriculture. (23/01/2001). *Livestock and Supplies – instrução Normativa n. 3, de 19 de janeiro de 2001, Seção Vol. 1* pp. 18–23). Published in: Diário Oficial da União.

Martins, M. C. T., Morgano, M. A., Vicente, E., Baggio, S. R., & Rodriguez-Amaya, D. B. (2011). Physicochemical composition of bee pollen from eleven brazilian states. *Journal of Apicultural Research, 55*, 107–115.

Morgano, M. A., Martins, M. C. T., Rabonato, L. C., Milani, R. F., Yotsuyanagi, K., & Rodriguez-Amaya, D. B. (2010). Inorganic contaminants in bee pollen from southeastern Brazil. *Journal of Agricultural and Food Chemistry, 58*, 6876–6883.

Pedro, A. M. K., & Ferreira, M. M. C. (2007). Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy. *Analytica Chimica Acta, 595*, 221–227.

Rabie, A. L., Wells, J. D., & Dent, L. K. (1983). The nitrogen content of pollen protein. *Journal of Apicultural Research, 22*, 119–123.

Rambo, M. K. D., Amorim, E. P., & Ferreira, M. M. C. (2013). Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. *Analytica Chimica Acta, 775*, 41–49.

Ribeiro, J. G., & Silva, R. A. (2007). *Estudo comparativo da qualidade de pólen apícola fresco, recém processado, não processado e armazenado em freezer e pólen de marca comercial através de análises fisico-químicas*. Tecnologia & Desenvolvimento Sustentável, Ano 1. https://www.academia.edu/5809746/estudo_comparativo_da_qualidade_de_p%c3%93len_ap%c3%8dcola_fresco_rec%c3%89m_processado_n%c3%83o_processado_e_armazenado_em_freezer_e_p%c3%93len_de_marca_comercial_atrav%c3%89s_de_an%c3%81lises_fisico-qu%c3%8dmicas (Accessed 18 March 2016).

Serra Bonvehí, J., & Escolà Jordà, R. (1997). Nutrient composition and microbiological quality of honey-bee-collected pollen in Spain. *Journal of Agricultural and Food Chemistry, 45*, 725–732.

Serra Bonvehi, J., Gonell Galindo, J., & Gomez Pajuelo, A. (1986). Estúdio de la composición y características físico-quimicas del pólen de abejas. *Alimentaria, 176*, 63–67.

Shenk, J. S., Workman, J. J., & Westerhaus, M. O. (2008). In D. A. Burns, & E. W. Ciurczak (Eds.), *Handbook of near- infrared analysis* (pp. 347–383). Florida: CRC Press, Inc.

Szczesna, T. (2007). Study on the sugar composition of honeybee-collected pollen. *Journal of Apicultural Science, 51*, 5–13.

Teye, E., Huang, X., Sam-Amoah, L. K., Takrama, J., Boison, D., Botchway, F., et al. (2015). Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis. *Food Chemistry, 176*, 403–410.

Viegas, T. R., Mata, A. L., Duarte, M. M. L., Kássio, M. G., & Lima, K. M. G. (2016). Determination of quality attributes in wax jambu fruit using NIRS and PLS. *Food Chemistry, 190*, 1–4.

Workman, J. J., Jr. (1996). *Applied spectroscopy reviews* (Vol. 31, pp. 251–320). Norwalk, CT: The Perkin Elmer Corporation.

Yang, K., Wu, D., Ye, X., Liu, D., Chen, J., & Sun, P. (2013). Characterization of chemical composition of bee pollen in China. *Journal of Agricultural and Food Chemistry, 61*, 708–718.

Zenebon, O., & Pascuet, N. S. (2005). *Métodos físico-químicos para análise de alimentos* (4 ed.). Brasília: Ministério da Saúde/ANVISA. São Paulo: Instituto Adolfo Lutz.