Discrimination of Brazilian propolis according to the seasoning using chemometrics and machine learning based on UV-Vis scanning data

Maíra M. Tomazzoli[1,*], Remi D. Pai Neto[1], Rodolfo Moresco 1, Larissa Westphal[1], Amélia R. S. Zeggio[1], Leandro Specht[2], Christopher Costa[3], Miguel Rocha[3] and Marcelo Maraschin[1]

[1]Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil

[2]Environmental Military Police, Florianopolis, Brazil

[3]Centre Biological Engineering, School of Engineering, University of Minho, Braga, Portugal

Summary

Propolis is a chemically complex biomass produced by honeybees (*Apis mellifera*) from plant resins added of salivary enzymes, beeswax, and pollen. The biological activities described for propolis were also identified for donor plant's resin, but a big challenge for the standardization of the chemical composition and biological effects of propolis remains on a better understanding of the influence of seasonality on the chemical constituents of that raw material. Since propolis quality depends, among other variables, on the local flora which is strongly influenced by (a)biotic factors over the seasons, to unravel the harvest season effect on the propolis' chemical profile is an issue of recognized importance. For that, fast, cheap, and robust analytical techniques seem to be the best choice for large scale quality control processes in the most demanding markets, e.g., human health applications. For that, UV-Visible (UV-Vis) scanning spectrophotometry of hydroalcoholic extracts (HE) of seventy-three propolis samples, collected over the seasons in 2014 (summer, spring, autumn, and winter) and 2015 (summer and autumn) in Southern Brazil was adopted. Further machine learning and chemometrics techniques were applied to the UV-Vis dataset aiming to gain insights as to the seasonality effect on the claimed chemical heterogeneity of propolis samples determined by changes in the flora of the geographic region under study. Descriptive and classification models were built following a chemometric approach, i.e. principal component analysis (PCA) and hierarchical clustering analysis (HCA) supported by scripts written in the R language. The UV-Vis profiles associated with chemometric analysis allowed identifying a typical pattern in propolis samples collected in the summer. Importantly, the discrimination based on PCA could be improved by using the dataset of the fingerprint region of phenolic compounds (λ = 280-400ηm), suggesting that besides the biological activities of those secondary metabolites, they also play a relevant role for the discrimination and classification of that complex matrix through bioinformatics tools. Finally, a series of machine learning approaches, e.g., partial least square-discriminant analysis (PLS-DA), *k*-Nearest Neighbors (kNN), and Decision Trees showed to be complementary to PCA and HCA, allowing to obtain relevant information as to the sample discrimination.

*Email: mairatomazzoli@gmail.com

Keywords: Propolis, UV-Vis scanning spectrophotometry, chemometrics, metabolic profile, botanical source, seasonality.

1 Introduction

Propolis is a resinous substance collected by honeybees *Apis mellifera* from various plant sources and added to salivary enzymes, beeswax, and pollen. Bees use propolis to seal openings in their honeycombs and to protect them against microorganisms and insects. Many studies have reported a broad spectrum of propolis' biological activities, such as cytotoxic, antiherpes, free radical scavenging, antimicrobial, and anti-HIV activities [1, 2]. More recently, a classification system has been proposed where Brazilian propolis samples fit into 12 groups based on their physiochemical traits and botanical origins [3]. The botanical origin of propolis is extremely important to guarantee raw materials of superior quality to supply demanding markets as cosmetics and pharmaceutical drugs. The biological activities described for propolis were also identified for resin donor plant, however a common challenge for the standardization of propolis samples is to understand the influence of seasonality on its chemical composition, which, in its turn, can modify its biological actions [4]. Previous studies of our research group (unpublished data) have identified a series of compounds in propolis produced in highland areas (i.e., São Joaquim county - altitude 1,360m) in Southern Brazil, which could be hypothetically associated to the native flora [5], giving rise to a typical propolis chemotype.

In this study, a bioinformatics approach was used, applying multivariate statistical techniques (principal component analysis - PCA and hierarchical clustering analysis - HCA) and machine learning to a UV-Visible scanning dataset (n = 73 samples, λ = 280-800 ηm) of propolis hydroalcoholic extracts (HE) samples. The analytical strategy herein adopted aimed to gain insights as to the claimed chemical heterogeneity of propolis samples collected over the seasons, in connection with the changes in the flora of the geographic region under study. Currently, the development of descriptive and classification models based on fast, cheap, and robust analytical techniques such as UV-Vis spectrophotometry is of interest to the pharmaceutical industry, for instance, since more detailed techniques (liquid or gas chromatography, coupled or not to mass spectrometry detectors) present important constraints for the routine analysis and quality control of complex matrices like propolis. On the other hand, the large amount of data afforded by UV-Vis scanning spectrophotometry and the eventual similarity of the spectral profiles of the samples turns the adoption of bioinformatics tools compulsory to obtain relevant and additional information.

2 Materials and Methods

2.1 Propolis samples and selection

Propolis samples from *A. mellifera* (n = 73) were collected in São Joaquim county (28° 17' 38" S, 49° 55' 54" W, Santa Catarina state, Southern Brazil) during 2014 in the summer, spring,

autumn, and winter and during 2015 in two seasons, i.e., summer and autumn. The samples were classified by visual analysis according to their colors as red, green, brown, and light brown propolis, taking into account that the resins collected by bees present a color peculiar to the plant donor.

2.2 Propolis extraction and UV-visible scanning spectrophotometry

The preparation of HE was performed as described by Popova et al (2004), with modifications [6]. Propolis samples (500 mg) were added of 25mL ethanol 70% (v/v) and incubated (24h, darkness). The extracts were filtered on cellulose support under vacuum, completing the final volume to 25 mL with EtOH 70% (v/v). The UV-visible spectra of propolis HE were performed by adding a 50µL aliquot of the extract (EtOH 70%) in 3mL of EtOH 70%. Absorbance values were recorded on a UV-visible spectrophotometer (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) using a spectral window of 280 to 800 ηm (2 ηm resolution/-data point).

2.3 Chemometric analysis and machine learning

The UV-Vis data set of the propolis HE was processed considering the definition of the spectral window of interest (280-800ηm), baseline correction, normalization, and optimization of the signal/noise ratio (smoothing). Further, the data matrix was exported to Excel®datasheet as a *.csv* format file and subjected to multivariate statistical analysis, using PCA and HCA. For that, scripts were written in R language (v. 3.1.1) using tools developed by our research group and some functions from the packages Chemospec [7] and HyperSpec [8]. PCA and HCA can help one to extract relevant features from a given dataset, minimizing the redundant information and characterizing the relationship between the variables studied.

For machine learning analysis, classification models were built to try to discriminate the propolis samples by their harvest season. Three models were chosen, e.g., partial least square-discriminant analysis (PLS-DA), *k*-Nearest Neighbors (kNN), and Decision Trees (as implemented in the rpart package from R), using a repeated cross-validation with 10 folds, with 10 repetitions. The models' parameters were optimized considering 10 different values, using error estimation procedures implemented in the *caret* R package. The scripts, raw data and chemometric analysis are available in supplementary material in the site: `http://darwin.di.uminho.pt/metabolomicspackage`. The report of analysis generated from the scripts provided by the R Markdown are available in `http://darwin.di.uminho.pt/metabolomics/dataset/Maira_PropolisUV`.

3 Results and Discussion

3.1 UV-Vis scanning spectrophotometry and chemometric analysis

Propolis is not used as a raw material directly in industry; rather, it is preprocessed by removing inert material, wax, dirt, and insoluble material, followed by the extraction of its bioactive compounds with suitable solvents. This process must preserve bioactive compounds, particularly phenolic ones. The UV absorption at 290-400 ηm is typical of phenolic compounds such as flavonoids [9] and all the spectral profiles (280-800 ηm) of the studied samples showed absorbance signals in that spectral window (**Fig.** 1), indicating that the extraction solution (EtOH: water, 70: 30, v/v) was able to recover the phenolic compounds from propolis. Besides, the spectral profiles showed to be somewhat similar, suggesting a homogeneous chemical composition among the samples, despite their collecting season. Thus, the UV-Vis spectral dataset was used for calculation of the principal components and for hierarchical clustering analysis, in order to tentatively classify the propolis samples into homogeneous groups according to the harvest season.

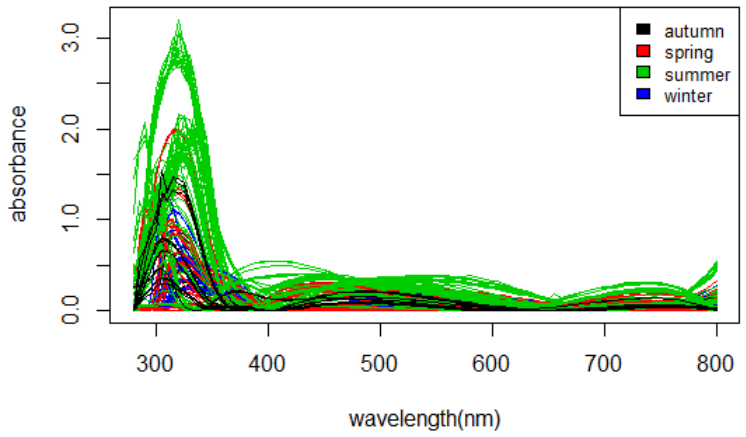

Figure 1: UV-Vis spectral profile (λ = 280-800 ηm) of seventy-three hydroalcoholic extracts (70% EtOH, v/v) of propolis samples collected in São Joaquim county during 2014, in all seasons (summer, spring, autumn, and winter) and during 2015 in two seasons (summer and autumn).

Hierarchical clustering analysis (HCA) was applied to the UV-Vis dataset (λ = 280-800 ηm). In this analysis, the objects in each cluster tend to be similar, but different from objects in other clusters, with no initial information on group composition [10]. The Euclidean distance between two samples was used as the similarity metric, while the method unweighted arithmetic average (UPGMA) was used for the hierarchical clustering process. In the method UPGMA, the highest similarity identifies the next cluster to be formed, estimating the arithmetic average of the similarities or distances between a candidate object and each of the cluster members. In the case of a previously formed cluster, the calculation is between all members of the two clusters. All objects receive equal weights in the computation [11]. The resulting tree revealed samples discriminated into two main groups, the first one having samples collected in the four seasons, but with few samples collected in the summer (**Fig.** 2). The second group, however, contain almost exclusively propolis samples produced in the summer, revealing an interesting separation.

Further, by applying HCA to the UV-Vis dataset (λ = 280-800 ηm) added of the 2015 samples (summer and autumn) the summer propolis samples did not group so well as previously observed (**Fig.** 3). The differences between the two hierarchical analyses may be linked to the extensive chemical variability that plants might present, as result of secondary metabolites biosynthesis pathways sensitive to the regulatory effects of many environmental factors as, for example, climate and interactions with insects and pathogens [12]. The chemical composition of propolis is directly related to the donor plant resin. In its turn, the chemical profile of the donor plant may suffer directly from changes in the climate as observed over the seasons in Southern Brazil, mostly in highlands areas such as the site of the present study. Indeed, in São Joaquim county, the greatest production of propolis occurs in the summer, coinciding with the budding of new plant species potentially donors of resin. In the other seasons, the propolis production drops because of the low temperatures and of the caducifolious habit of several plant species [13].

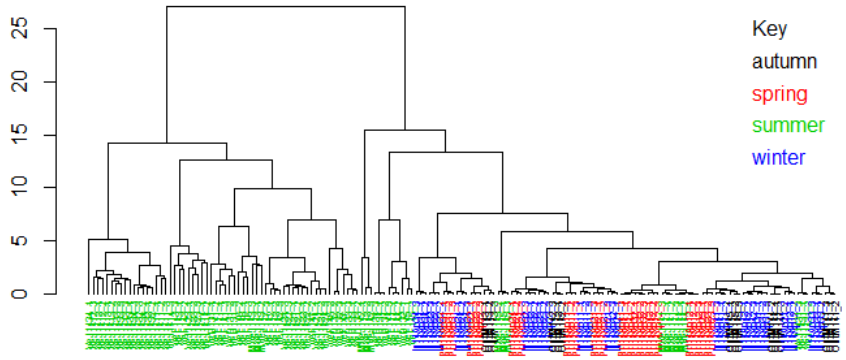

Figure 2: Hierarchical clustering dendrogram (UPGMA method) of fifty-five samples from Southern Brazil, collected during 2014 in all seasons: summer, spring, autumn, and winter.

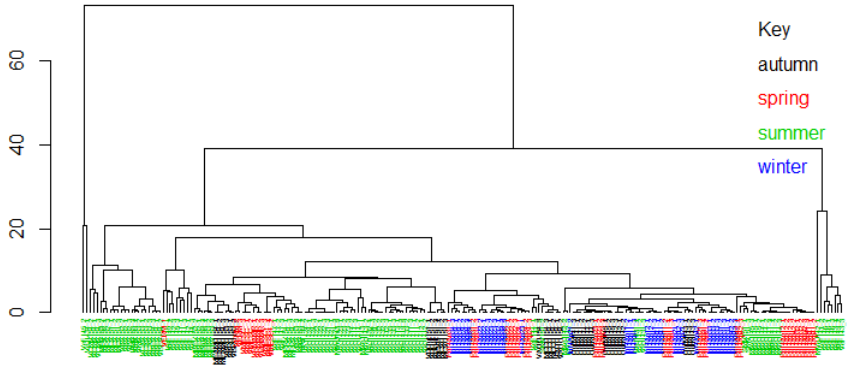

Figure 3: Hierarchical clustering dendrogram (UPGMA method) of seventy-three samples collected in São Joaquim county during 2014, in all seasons (summer, spring, autumn, and winter) and during 2015 (summer and autumn).

In order to get a better understanding of the harvest season effect indicated by HCA, the UV-Vis dataset was used for the calculation of the principal components (PCA). The main objective of the PCA is to reduce the size of the data without loss of information. PCA turn variables with high correlation in latent variables uncorrelated, allowing the separation and extraction of relevant information [10]. In general, the results of PCA and HCA are complementary and

when employed in tandem constitute an interesting tool to construct reliable models [14]. The first two components PC1 (68.0%) and PC2 (13.9%) explained 81.9% of the total variance of the dataset (**Fig.** 4) for samples collected in 2014 seasons. By expanding the model and including the contribution of the PC3 (12.2%), it was possible to cover 94.1% of dataset's variability. The PCA results have confirmed the sample discrimination by seasons into two groups, as observed in the HCA. The summer samples dispersed in the two components, while the remaining ones overlaid and centered on the graphic. Similarly to HCA, after adding the 2015 samples (summer and autumn) to the PCA model (**Fig.** 5), it was not possible to classify so well the summer samples in a distinct group, the most samples are overlaid and centered on the graphic. For this analysis the first two components PC1 (64.56%) and PC2 (17.6%) explained 82.16% of the total variance of the dataset. Regarding the color variable of the samples, both HCA and PCA did not allow discriminating the samples (data not shown).

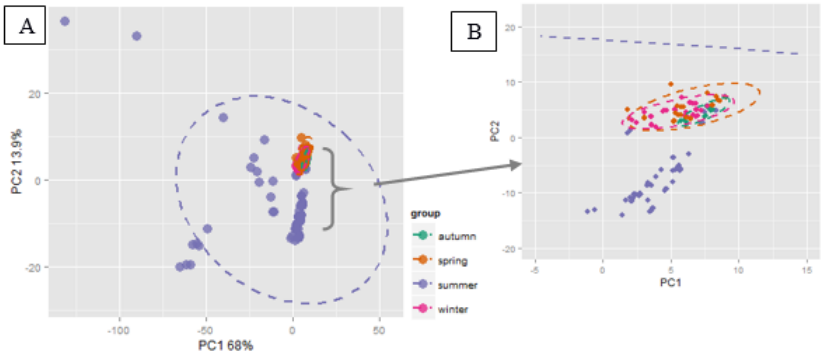

Figure 4: (A) Principal components analysis (PCA) scores scatter plot of the UV-Vis spectral profile (λ = 280-800 ηm) of propolis samples collected in 2014 (summer, spring, autumn, and winter) in São Joaquim county, Southern Brazil. (B) Amplification of the overlapping samples in PCA.

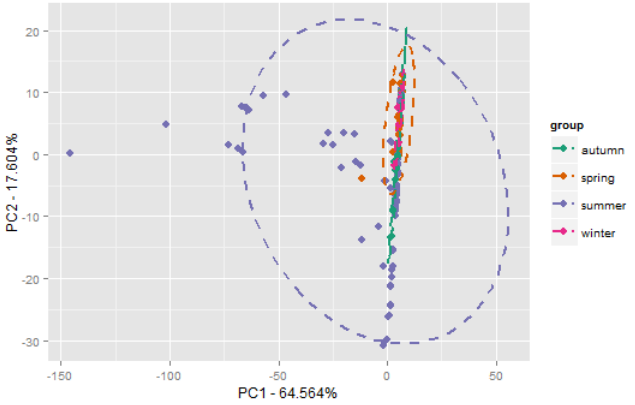

Figure 5: Principal components analysis (PCA) scores scatter plot of the UV-Vis spectral profile (λ = 280-800 ηm) of samples collected during 2014, in all seasons and 2015 (summer and autumn).

Since phenolic compounds have been claimed as the most important bioactive metabolites in propolis, in a second approach we investigated the harvest season effect on the phenolic composition of that biomass. Thus, the UV-Vis dataset in the region of absorption of those secondary metabolites, i.e., 280-400 ηm, was used for further HCA and PCA. Again, two groups were detected by HCA (**Fig.** 6) and PCA (**Fig.** 7) and both methods discriminated the summer propolis

samples as a result of their phenolic composition. In the PCA model, the first two components comprised for 92.8% of the total variance of the data set, suggesting that phenolic compounds seem to be an interesting class of metabolites for discrimination of propolis. Indeed, taking into account the improved discrimination shown in the PCA results using the UV-Vis fingerprint region of phenolic compounds, one could speculate that by targeting those compounds in propolis extracts better classification models would come about.

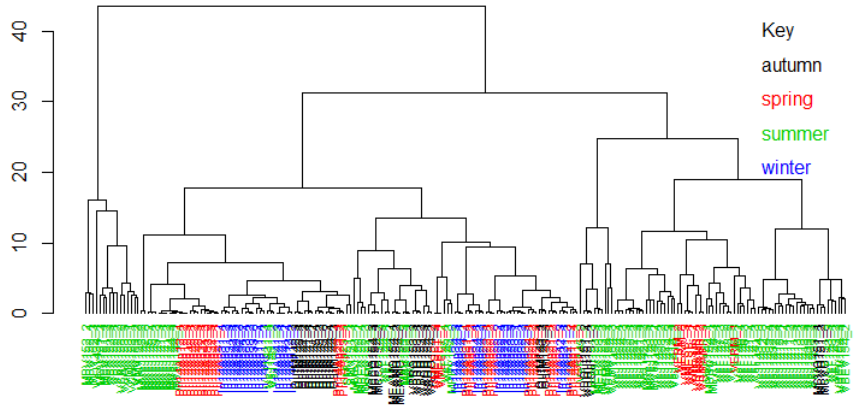

Figure 6: Hierarchical clustering dendrogram (UPGMA method) of the fingerprint region of absorbances of phenolic compounds (UV-Vis, λ = 280-400 ηm) of samples collected over all seasons in 2014 and 2015 (summer and autumn) in São Joaquim county.

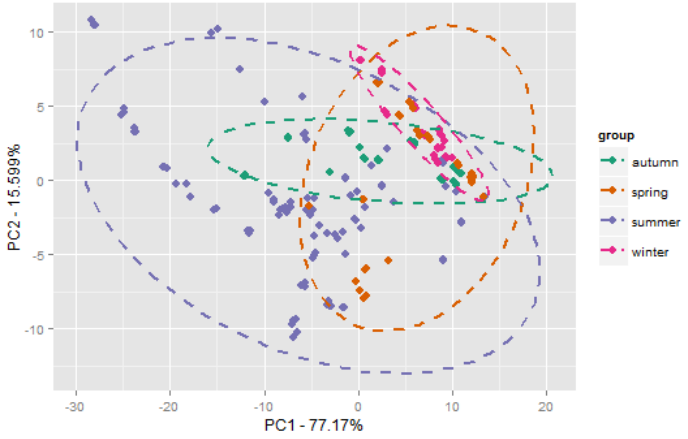

Figure 7: Principal components analysis scores scatter plot of the fingerprint region of absorbances of phenolic compounds (UV-Vis, λ = 280-400 ηm) of samples collected over all seasons in 2014 and 2015 (summer and autumn) in São Joaquim county.

These findings are of interest for the purpose of quality control processes of propolis extracts in industry, based on the fact that most of their well-known pharmacological activities rely on those secondary metabolites. In general, the majority of phytochemicals belong to the groups of phenolic compounds, alkaloids, and terpenes [15]. Nonetheless, flavonoids, phenolic acids and their ester derivatives are the major metabolites found in propolis [16]. For instance, the European propolis is characterized by their prominent amounts of flavonoids, which are not often found in tropical samples [17]. In the later, prenylated phenylpropanoids are often present, the best known is the (3,5-diprenyl-4-hydroxycinnamic acid) [18], a high valuable compound

(£ 315/10mg) also known as Artepillin C®, who has been patented for the treatment of tumors [19, 20]. *Baccharis dracunculifolia* is a native plant to Brazil commonly found in Minas Gerais state (Southeastern Brazil) and source of a green resin, the main source of Artepillin C® [21, 22]. Considering the interaction between *B. dracunculifolia* and *A. mellifera*, the best period to produce propolis rich in Artepillin C® is from December to April, i.e., summer time in south hemisphere [23]. In this context, one can note the importance of identifying the seasonality effect on the propolis chemical profile and its resulting quality as source of important secondary metabolites. Finally, despite the fact that UV-Vis scanning spectrophotometry is a fast, cheap, and reliable analytical technique, the amount of data afforded makes unfeasible the selection of propolis samples according to their spectral profile by visual inspection, turning the bioinformatics tools mandatory for the recovery of important features for the classification of heterogeneous samples into similar groups.

3.2 Machine learning

Machine learning techniques have become popular in recent years for decision support, predicting events, and data analysis. In this context, supervised models using classification algorithms inductors such as partial least square-discriminant analysis (PLS-DA), *k*-Nearest Neighbors (kNN) and decision trees were built to discriminate São Joaquim propolis according to the seasonality. PLS-DA is a classic PLS method where the variable y is categorical and represents the class samples. Using the class information, PLS-DA tends to improve the separation between two groups of samples. It is commonly used to classification and selection of biomarkers [24]. For the kNN algorithm (k-Nearest Neighbors) when a new object is presented to the classification, a set of similar examples is retrieved from the training set, being used to classify the new object. These similar examples, have the shortest distance in the dimensional space and therefore the algorithm is known as the "nearest neighbor" [25]. The rpart algorithm is associated with the inductors algorithms of classification trees, that implements the CART methodology (Classification and Regression Trees). Classification trees are binary and its growth is limited to 31 (thirty one) depth levels, the algorithm also implements pruning process to minimize the error estimate [26]. Machine learning was applied to the UV-Vis dataset (λ = 280-800 ηm) and the PLS-DA, kNN, and decision tree models showed an accuracy of 67.5%, 77.5% and 81.4%, respectively, for predictive analytics of seasonality (**Tab.** 1). Repeated cross-validation with 10 folds and 10 repetitions was used as parameters for estimating the models' performance. The Kappa statistic, a measure of how closely the instances classified by the classifier matched the actual data label and the accuracy found are shown in **Tab.** 1.

Table 1: Accuracy and kappa indexes using PLS-DA algorithms, *k*-Nearest Neighbor, and rpart as classification inductors for propolis' UV-Vis data set (λ= 280-800 ηm). The last two columns show the standard deviation of the accuracy and Kappa statistics.

	Accuracy	*Kappa*	Accuracy SD	*Kappa* SD
pls	0.67497	0.50574	0.09191	0.13401
knn	0.77486	0.66985	0.07874	0.11435
rpart	0.81435	0.72467	0.08599	0.12595

Further, the variable importance analysis revealed that mostly the absorbances in the phenolic spectral window seemed to be more useful for prediction (e.g., λ= 284-288 ηm; 342-343 ηm; and 364-366 ηm). This findings prompted us to repeat the machine learning analysis but just considering the UV-Vis data set of the phenolic region in the spectrum (λ = 280-400 ηm). The performance of the classification models (**Tab.** 2) was slightly inferior to the previous analysis (**Tab.** 1), but it is indicative that eventually only the absorbance signals of the phenolic region seems to be sufficient for propolis classification.

Table 2: Accuracy and kappa indexes using PLS-DA algorithms, *k*-Nearest Neighbor, and rpart as classification inductors for UV-Vis data set (λ= 280-400 ηm). The last two columns show the standard deviation of the accuracy and Kappa statistics.

	Accuracy	*Kappa*	Accuracy SD	*Kappa* SD
pls	0.66691	0.47750	0.06198	0.09703
knn	0.74258	0.61540	0.08085	0.11903
rpart	0.78989	0.68993	0.08383	0.12182

In this context, over the past years, UV-Vis spectrophotometry has been an analytical technique to guarantee quality control of chemically complex matrices as propolis and plants extracts for metabolomic studies. Characteristics of the absorption spectra indicative of the chemical composition of the sample may be used as the basis for the construction of descriptive and predictive models, including machine learning. Thus, the selection of specific sample characteristics can be used to improve the accuracy of the classification model or by establishing a subset of classes discriminating characteristics [27]. The propolis profile is well known for its high chemical heterogeneity considering the huge biodiversity of plant species found in some producer regions [28] in Brazil, e.g., Atlantic Rainforest in Santa Catarina state, Southern Brazil. Because of this, the effect of flora composition on the propolis' chemical profile has great influence and we could expect a high chemical heterogeneity among samples from distinct geographic regions where propolis has been collected.

Studies using UV-Vis scanning spectrophotometry to obtain the chemical profiles of propolis HE found 100% of accuracy using the PLS-DA algorithm model for predictive analysis of regions (South and Southeast Brazil [9]). Similarly, nuclear magnetic resonance (NMR) of propolis samples has showed to be effective to classify that biomass in relation to its geographical region. Through the use of large NMR data set and data mining techniques the construction of descriptive (PCA) and predictive (PLS-DA) models achieved a good performance, i.e., accuracy $\sim$ 83% [27].

4 Conclusions

The UV-Vis spectrophotometric profile approach associated with chemometric analysis (PCA and HCA) allowed identifying a different grouping pattern in samples of propolis produced during the summer season over the other seasons, inferring the importance of the seasonality effect on the propolis chemical profile and its resulting quality as source of important secondary metabolites. The classification model based on chemometrics herein described could even be

improved by using the dataset of the fingerprint region of phenolic compounds, suggesting that besides their biological activities they are also compounds relevant for the discrimination and classification of that complex matrix through bioinformatics tools. The use of machine learning tools showed to be complementary to the descriptive PCA and HCA models, allowing to obtain a better classification of the studied samples.

Acknowledgements

Financial support and the research fellowship to the later author from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) are acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq. The authors also thank the FCT Strategic Project of UID/BIO/04469/2013 unit.

References

[1] M. Marcucci. Propolis: chemical composition, biological properties and therapeutic activity. *Apidologie*, 26(2):83–99, 1995. URL `http://www.apidologie.org/index.php?option=com_article&access=standard&Itemid=129&url=/articles/apido/pdf/1995/02/Apidologie_0044-8435_1995_26_2_ART0002.pdf`.

[2] A. H. Banskota, Y. Tezuka and S. Kadota. Recent progress in pharmacological research of propolis. *Phytotherapy Research*, 15(7):561–571, 2001.

[3] Y. K. Park, S. M. Alencar and C. L. Aguiar. Botanical origin and chemical composition of Brazilian propolis. *Journal of Agricultural and Food Chemistry*, 50(9):2502–2506, 2002.

[4] L. Gobbo-Neto and N. P. Lopes. Plantas medicinais: Fatores de influência no conteúdo de metabólitos secundários. *Quimica Nova*, 30(2):374–381, 2007.

[5] A. R. S. Zeggio. *Estudo da própolis Catarinense, com ênfase na sua composição fenólica: uma estratégia à tipificação regional*. Ph.D. thesis, Universidade Federal de Santa Catarina, 2014.

[6] M. Popova, V. Bankova, D. Butovska, V. Petrov, B. N. Damyanova, a. G. Sabatini, G. L. Marcazzan and S. Bogdanov. Validated methods for the quantification of biologically active constituents of poplar-type propolis. *Phytochemical Analysis*, 5(4):235–240, 2004.

[7] B. A. Hanson. ChemoSpec : An R Package for Chemometric Analysis of Spectroscopic Data and Chromatograms (Package Version 2 . 0-2). pages 1–41, 2014.

[8] C. Beleites. Import and Export of Spectra Files. *Vignette for the R package hyperSpec*, pages 1–20, 2011.

[9] R. S. N. Paganotti, J. C. Rezende and P. J. D. Barbeira. Discrimination between producing regions of Brazilian propolis by UV-VIis spectroscopy and partial least squares discriminant analysis. *Current Analytical Chemistry*, 10(4):537–544, 2014.

[10] K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton: CRC Press, 2009.

[11] P. Legendre and L. Legendre. *Numerical Ecology*. Amsterdam: Elsevier, 1998.

[12] R. Bauer and G. Tittel. Quality assessment of herbal preparations as a precondition of pharmacological and clinical studies. *Phytomedicine*, 2(3):193–198, 1996. URL `http://dx.doi.org/10.1016/S0944-7113(96)80041-9`.

[13] E. Bastos, R. Santana, A. Calaºa-Costa and P. Thiago. Interaction between Apis mellifera L. and Baccharis dracunculifolia DC, that favours green propolis production in Minas Gerais. *Brazilian Journal of Biology*, 71(3):727–734, 2011.

[14] R. R. Meglen. Chemometrics: Its role in chemistry and measurement sciences. *Chemometrics and Intelligent Laboratory Systems*, 3(1-2):17–29, 1988.

[15] D. Kennedy and E. Wightman. Herbal extracts and phytochemicals: plant secondary metabolites and the enhancement of human brain function. *Advances in Nutrition*, 2:32–50, 2011. URL `http://advances.nutrition.org/content/2/1/32.full.pdf`.

[16] A. H. Banskota, Y. Tezuka, J. K. Prasain, K. Matsushige, I. Saiki and S. Kadota. Chemical constituents of Brazilian propolis and their cytotoxic activities. *Journal of Natural Products*, 61(7):896–900, 1998.

[17] V. Bankova. Chemical diversity of propolis and the problem of standardization. *Journal of Ethnopharmacology*, 100(1-2):114–117, 2005.

[18] V. S. Bankova, S. L. de Castro and M. C. Marcucci. Propolis: recent advances in chemistry and plant origin. *Apidologie*, 31(1):3–15, 2000.

[19] Y. K. Park, J. F. Paredes-Guzman, C. L. Aguiar, S. M. Alencar and F. Y. Fujiwara. Chemical constituents in Baccharis dracunculifolia as the main botanical origin of Southeastern Brazilian propolis. *Journal of Agricultural and Food Chemistry*, 52(5):1100–1103, 2004.

[20] S. Kumazawa, M. Yoneda, I. Shibata, J. Kanaeda, T. Hamasaka and T. Nakayama. Direct evidence for the plant origin of Brazilian propolis by the observation of honeybee behavior and phytochemical analysis. *Chemical & Pharmaceutical Bulletin*, 51(6):740–742, 2003.

[21] T. Hata, S. Tazawa, S. Ohta, M. R. Rhyu, T. Misaka and K. Ichihara. Artepillin C, a major ingredient of Brazilian propolis, induces a pungent taste by activating TRPA1 channels. *PLoS ONE*, 7(11):1–9, 2012.

[22] S. M. D. Alencar, C. L. D. Aguiar, J. Paredes-Guzmán and Y. K. Park. Composição química de Baccharis dracunculifolia, fonte botânica das própolis dos estados de São Paulo e Minas Gerais. *Ciência Rural*, 35(4):909–915, 2005.

[23] J. M. Sforcin, J. P. B. Sousa, A. A. Filho, J. K. Bastos, M. C. Búfalo and L. R. S. Tonuci. *Baccharis dracunculifolia: uma das principais fontes vegetais da própolis brasileira*. São Paulo: Fundação Editora da UNESP (FEU), 2012.

[24] E. Szymańska, E. Saccenti, A. K. Smilde and J. a. Westerhuis. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8:3–16, 2012.

[25] S. D. C. D. O. Galvão and E. R. H. Júnior. Ponderação do knn utilizando-se de métodos tradicionais de selecção de atributos. *da Vinci, Curitiba*, 1(1):115–125, 2004.

[26] C. Frizzarini and M. S. Lauretto. Proposta de um algoritmo para indução de árvores de classificação para dados desbalanceados. pages 722–733, 2013.

[27] M. Maraschin, A. Somensi-Zeggio, S. K. Oliveira, S. Kuhnen, M. M. Tomazzoli, A. C. M. Zeri, R. Carreira and M. Rocha. A machine learning and chemometrics assisted interpretation of spectroscopic data - A NMR-based metabolomics platform for the assessment of Brazilian propolis. *7th IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB 2012*, LNBI 7632:129–140, 2012.

[28] D. G. Watson, E. Peyfoon, L. Zheng, D. Lu, V. Seidel, B. Jonhston, J. a. Parkinson and J. Fearnley. Application of principal components analysis to 1H-NMR data obtained from propolis samples of different geographical origin. *Phytochemical Analysis*, 17(5):323–331, 2006.