

Geographical origin identification of teas using UV-VIS spectroscopy

Thi Hue Tran^{1*}, *Quoc Toan Tran*¹, *Thi Thao Ta*² and *Si Hung Le*²

¹Faculty of chemistry, Thai Nguyen University of education, Vietnamese

²Faculty of chemistry, VNU University of science, Vietnamese

Abstract. In this work we proposed a method to verify the differentiating characteristics of simple tea infusions prepared in boiling water alone, which represents the final product as ingested by the consumers. For this purpose, total of 125 tea samples from different geographical provinces of Vietnam have been analyzed in UV-Vis spectroscopy associated with multivariate statistical methods. Principal Component Analysis-Discriminant Analysis (PCA-DA), Partial Least Squares Discriminant Analysis (PLS-DA) and Artificial Neural Network (ANN) were compared to construct the identification model. The experimental results showed that the performance of ANN model was better than PCA-DA and PLS-DA model. The optimal ANN model was achieved when neuron numbers were 200, identification rate being 99% in the training set and 84% prediction set. The proposed methodology provides a simpler, faster and more affordable classification of simple tea infusions, and can be used as an alternative approach to traditional tea quality evaluation.

1 Introduction

The tea (*Camellia sinensis*) has long been a familiar drink among humans and is widely used in many countries around the world. The researches regarding the chemical compositions of tea show that the tea leaves contain approximately 2000 chemical compounds; in which, the polyphenols account for about 25 - 30%, caffeine accounts for about 2 - 5%, the theobromine and theophylline account for about 0.2 - 0.5% of the dry matter content [1]. The special chemical composition, especially the content of methylxanthines, polyphenols determine the quality and economic value of the finished tea [2]. Many studies have shown that besides the seed factors, farming regime, processing techniques, climatic and soil conditions have a great influence on the chemical composition of tea [2]. Therefore, the finished tea products in different geographical regions can have different economic values. Currently, due to the great demand for tea, especially those of high quality, the famous tea products are being counterfeited by the brands with unknown quality and unclear origin. To address the above challenge, the quality control and traceability of the tea products gradually become inevitable needs, bringing practical benefits to the manufacturers and consumers. Vietnam has 35 tea producing provinces with a total area of 125,000 hectares, most of them in the Northern Midlands, North Central and Central Highlands provinces. Vietnam has exported tea to 107 countries, ranking 7th in

* Corresponding author: huettt@tmue.edu.vn

export volume (987.3 thousand tons in 2018), ranking 6th in export value. However, the lacking of quick and reliable analysis methods for determination the origin and quality of tea products is one of bottle-neck to ensure and increase the quality of the products.

The prior studies focused on the differentiation of countries of production or tea varieties according to the mineral composition by using inductively coupled plasma atomic emission spectroscopy or according to the contents of catechins and purine alkaloids by using high performance liquid chromatography (HPLC) with pattern recognition methods [3-5-4]. The main drawbacks of these methods are the use of expensive equipments, usually involving large operational or maintenance cost, and the use of a number of reagents to carry out the extraction of organic compounds, the derivatization of analytes or the mineralization of samples for elemental analysis. Another approach has become more trending recently is the combination of spectral analysis methods. Ultraviolet - Visible spectroscopy (UV-VIS) and Infrared reflectance spectroscopy (IR) combined with multivariate analysis techniques [6-15]. The identification is intended to categorize the food source that is growing vigorously. Compared to other chemical analysis methods, the advantages of the spectral analysis methods are low cost and high-throughput since the tedious sample preparation is not needed. UV-VIS has already been used to differentiate between black, green, and Pu-erh tea varieties [17]. However the method used methanol as the extract solvent, which produced a very broad spectra with highly correlated variables, and required more sophisticated non-linear classifier, Support Vector Machines. Furthermore, the methanol is toxic to humans and the environment [18]. The geographical classification of teas applying IR spectroscopy. Can simultaneously classify both geographic origin and variety for teas has been proposed using digital images [9]. However, these methodologies were carried out directly on the tea as contained in commercialized bags, whereas the infusion represents the final product as ingested by the consumer. In this study, we propose a fast and simultaneous method for identifying the origins of green tea products in several provinces of Vietnam by combining of UV-Vis spectroscopy analysis and Principal Component Analysis-Discriminant Analysis (PCA-DA), Partial Least Squares Discriminant Analysis (PLS-DA), and Artificial Neural Network (ANN).

2. Material and methods

The molecular absorption spectrometer analysis using scientific series 1700 PC - Shimadzu (Japan) was used. Each spectrum consists of 211 values of intensities at 1 nm intervals in the wavelength range 190 to 700 nm with a quartz cuvette with optic path of 10 mm.

A total of 125 green tea samples with known geographical origins was collected from the provinces of Thai Nguyen (35 samples), Lam Dong (25 samples), Phu Tho (31 samples), and other provinces such as Ha Giang, Yen Bai, Tuyen Quang, Hoa Binh (34 samples) was collected. The original information of the samples were recognized by onsite collection (for setting up the model) or based on the package (for comparison of the geographical origin between predicted and training results). Approximately 100 g of air-dried tea-leaves were kept at least 2 days in the dehumidifier at the 30% moisture before analyzing. The dried tea samples were ground in a laboratory grinder to obtain tea powder through to 63 μm mesh BS sieves.

Approximately 1.0 g of tea powders, were accurately weighted, extracted twice with 50.0 mL double-distilled water at temperature of 90 - 95°C for 5 minutes each extraction [10]. The tea solution were filtered with filter paper (0.45 μm), and diluted to 100 mL with double-distilled water. The 1.00 mL of this solution were transferred to a 25mL volumetric flask rising with with distilled water.

Matlab R2016a (The MathWorks, USA) was used for statistical analysis. All 125 samples were randomly divided into two datasets, training and external validating sets with the rate of 80/20. The external validating set consisting of 25 samples was not used for the model construction, hence reflecting the predicting power of the model. A data matrix consisting of columns (absorbance at the selected wavelengths) and rows (the tea samples) was created for the chemometric calculations. PLS-DA, PCA-DA, ANN were applied to obtain the classification models.

3. Results and Discussion

3.1. UV-Vis absorbance spectra

The UV-VIS spectra of the sample is shown in figure.1a. The most informative range of the spectra is found in the UV region (190 – 490 nm), while the information of the region above 490 nm is not important. The absorbance in the range 250 - 350 nm is associated with the $n \rightarrow \pi^*$ transitions of methylxanthines and catechins [16].

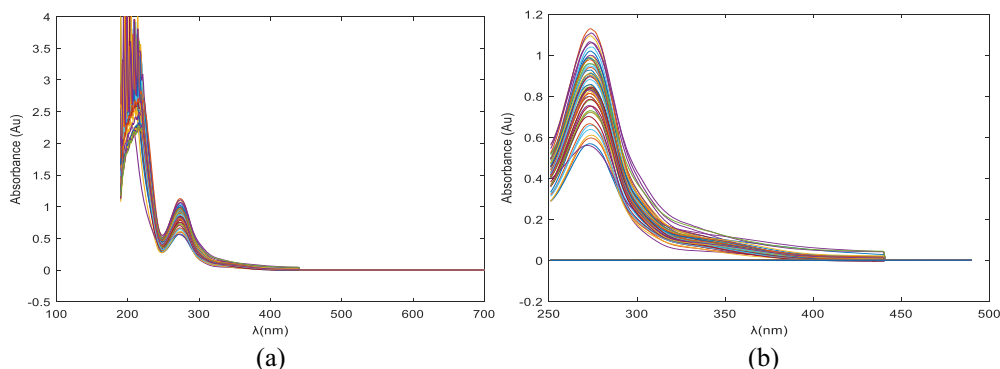


Fig. 1. Raw UV-Vis spectra of all studied tea samples: (a) 190-700nm; (b) 251-490 nm regions

3.2 Identify according by PCA-DA and PLS-DA

3.2.1. Selection of principal components

PCA is a statistical method to transform multiple indicators into several representative aggregative indicators. Redundancy information is reduced from a high-dimensional space to a low dimensional space by using PCA. The vectors obtained from each principal component are orthogonal. As shown in Figure 2, the first principal component (PC1) accounts for 88.76% of the variance. It is explained that the first component represented 88.76% of the information of the green tea samples and only the first PC was used to setup the classification model.

PLS-DA is a modification of Partial Least Squares Algorithm for classification purposes. It is based on the PLS2 algorithm that searches for latent variables with a maximum covariance for the categorical variables (Y) [20]. As shown in Figure 3, PC1 only accounts for 20% of the total variance of the data set. If PC11 is added, the cumulative total variance is over 95%. Therefore, we choose PC number = 11 to start investigating to choose the optimal number of PC.

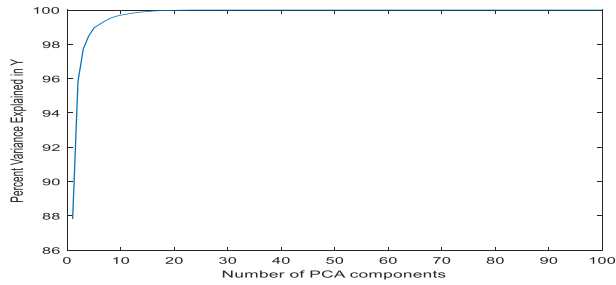


Fig. 2. Accumulated percent variance explained by components for PCA-DA classification model

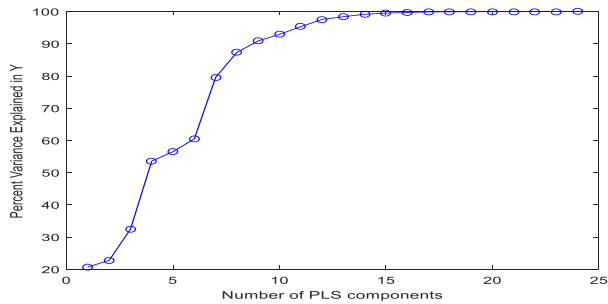


Fig. 3. Accumulated percent variance explained by components for PLS-DA classification mode

3.2.2. PCA-DA and PLS-DA model performance

Two supervised recognition techniques, PCA-DA and PLS-DA, were performed with PC 1~20. Figure 4 and figure 5 represent the recognition results obtained by the PCA-DA and PLS-DA approaches in training and prediction sets. Compared to PCA-DA model, the PLS-DA model was able to deal with imbalance training samples and the prediction set. The ability to determine the origin of a tea sample in the prediction set shows that the PCA-DA model reached the highest classification accuracy of 84% with PC12, while PLS DA performed only 76% with PC18. Figure 6 represents recognition results obtained by the PCA-DA with PC12 model, the geographical origins of the samples TN1, TN4, PT2, and PT6 in the external prediction set were wrongly identified. Thus, the accuracy of the model PCA-DA identifying 25 test samples reaching 84% is not high. Therefore, the study continues to investigate the ability to identify the geographical origin of tea samples based on UV-Vis spectral data using the ANN algorithm.

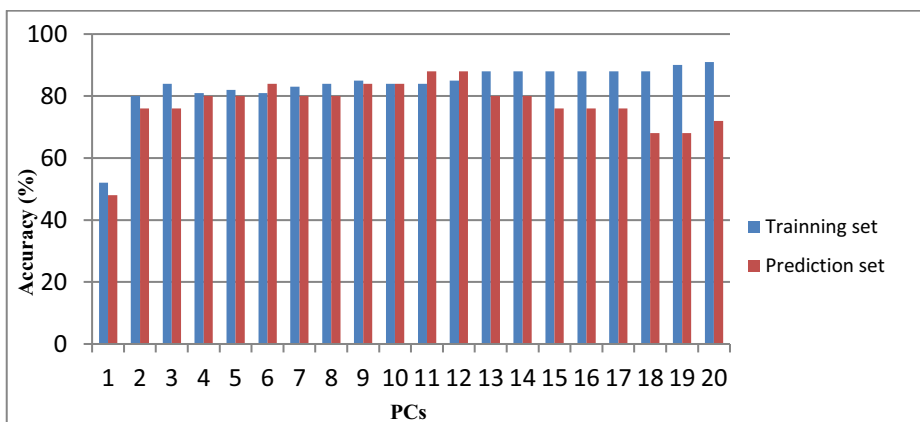


Fig. 4. Accuracy of the PCA-DA model use different PCs

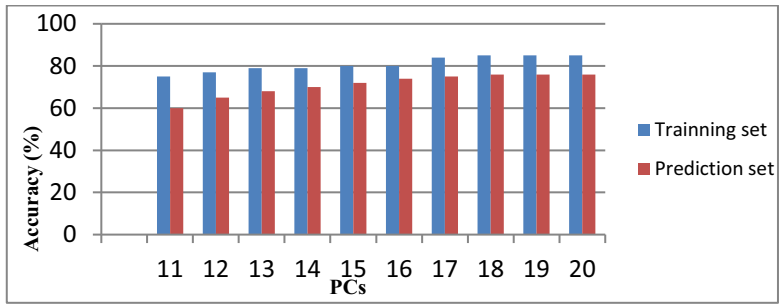


Fig. 5. Accuracy of the PLS-DA model use different PCs

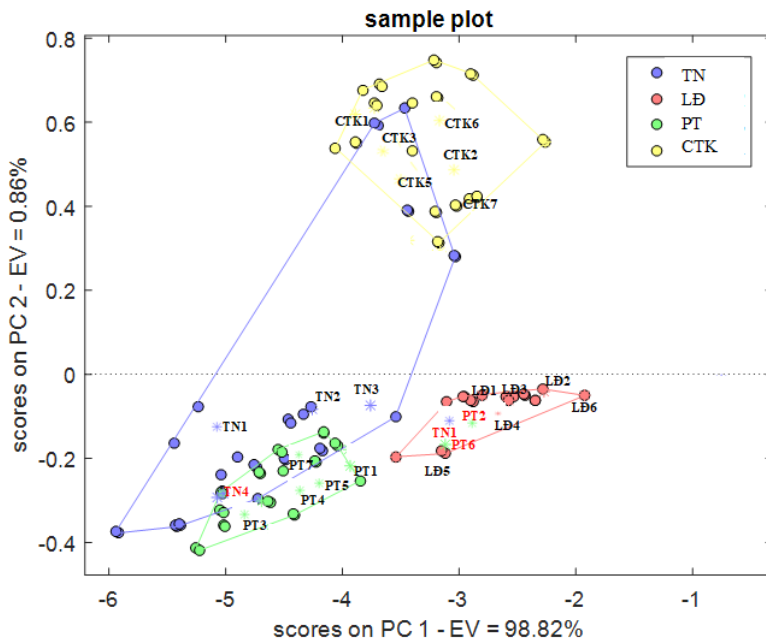


Fig. 6. PCA-DA sample plot for classification of green tea

3.3. Identify according by ANN

In the ANN model, the data of 125 samples is divided into 100 samples belonging to the training dataset and 25 samples belonging to the test set. The training data set is randomly divided into 3 parts at the ratio of 75:15:15, in turn, train, validate, and test. The training dataset is used to calculate gradients and continuously update the weight and error values of the network being trained. Cross evaluation datasets are used to monitor the training process. The results of grouping 100 standard samples in 3 datasets training, validation and testing are shown in the confusion chart in figure 7.

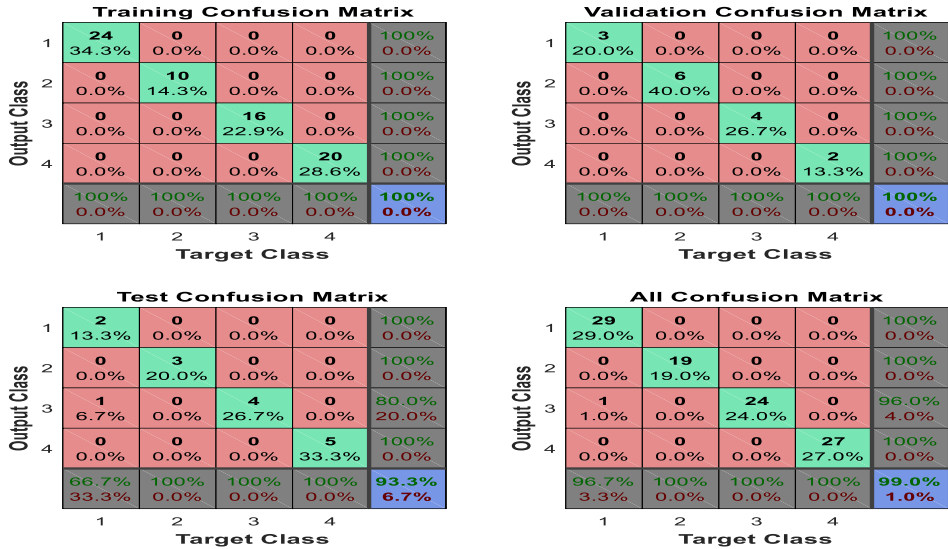


Fig. 7. ANN sample plot for classification of the 100 standard samples

From the results of figure 7, the accuracy in the set of training samples and the set of cross evaluation samples is 100%. After training and cross-evaluation, the process of re-checking the built-in model is done through the test sample set. The accuracy in the test sample is 93.3%. In which, the tea samples in Lam Dong province (group 2), Phu Tho province (group 3) and other provinces (group 4) had 100% accuracy. In 3 samples of Thai Nguyen tea (group 1) belonging to the test set, 2 samples were correctly identified and 1 sample was incorrectly identified in the Phu Tho tea group. Thus, the accuracy of the ANN model based on the database of 100 standard samples reached 99% improved compared with PLS-DA and PCA-DA methods.

After building an identification model, test with 25 samples. The results of grouping 25 test samples are shown in the Confusion chart in Figure 8.

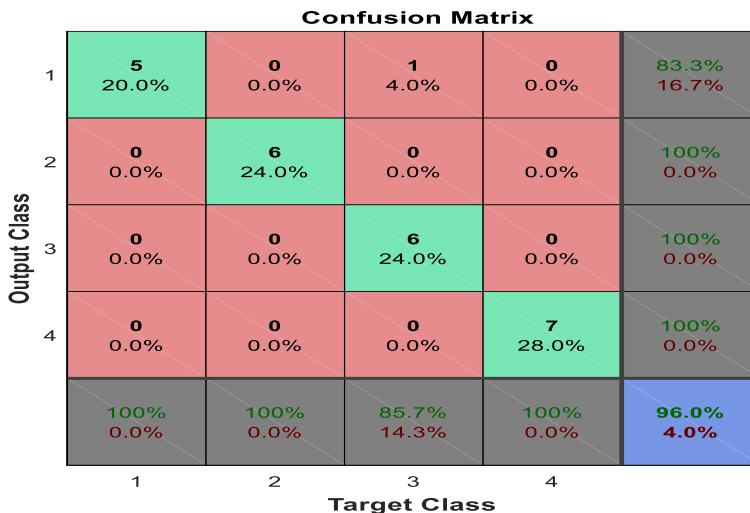


Fig. 8. ANN sample plot for classification of the 25 test samples

The results of figure 8 is found that the results of classifying the test samples from 4 different geographical regions have a high accuracy of 96%. Out of 25 samples, the ANN model correctly identified 24 samples, and 1 sample in Phu Tho province incorrectly identified Thai Nguyen sample. Therefore, ANN is the suitable method to determine the origin of a tea sample.

4. Conclusions

This study used ANN as a pattern recognition tool to develop an identification model. We show that UV-Vis spectra coupled with ANN pattern recognition can identify the country of origin of green tea samples. The ANN algorithm outperforms the PCA-DA and PLS-DA approaches in identifying the geographical origin of the tea samples. Therefore, UV-Vis spectra analysis with ANN pattern recognition could be used to identify the origin of other agricultural products. The proposed classification method is an alternative to those using expensive equipments or involving tedious sample preparation. UV-Vis spectra are easily obtained with operational simplicity and low cost and the applied pattern recognition calculations are well known.

This study is a part of project DH2018-TN04-03 funded by Thai Nguyen of Education.

References

1. A. Y. Yashin, B. V. Nemzer, I. Y. Yakov, *J. of Food Research*, **4(3)**, 56-88 (2015)
2. L. F. Pedro, J. M. Mari´a, F. Pablos, G. A Gustavo, *J. Agric. Food Chem.*, **50**, 1833-1839 (2002)
3. L. F. Pedro, J. M. Mari´a, F. Pablos, G. A. Gustavo, *J. Agric. Food Chem.*, **49**, 4775-4779 (2001).
4. A. M. Pineiro, A. S. Fisher, S. J. Hill, *J. of Food Composition and Analysis*, **16(2)**, 195-211 (2003).
5. M. A. Szymczycha, M. Welna, P. Pohl, *Microchemical Journal*, **121**, 122-129 (2015)
6. Q.S. Chen, Z.M. Guo, J.W Zhao, *J. Pharmaceut. Biomed.*, **48**, 1321-1325 (2008)
7. Q. S. Chen, J. W. Zhao, H. Lin, *Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy*, **72(4)**, 845-850 (2009)
8. A. Dankowska, W. Kowalewski, *Spectrochimica Acta A: Molecular and Biomolecular Spectroscopy* **211**, 195-202 (2019)
9. P. H. G. D. Diniz, A. A. Gomes, M. F. Pistonesi, B. S. F. Band, *Food Analytical Methods*, **7(8)**, 1712-1718 (2014)
10. P.H.G. Diniz, M.F. Barbosa, K.D.T. Melo Milanez, M.F. Pistonesi, *Food Chemistry*, **192**, 374-379 (2016)
11. P. Lin, Y. Chen, Y. He. *Food and Bioprocess Technology* **5(1)**, 235-242, (2009)
12. J. M. Ning, J. Sun, X. Zhu, Z. Zhang. *J. Food Sc. and Techn.*, **11(4)**, 337-343 (2016)
13. J. M. Ning, J. J. Sun, S. H. Li, M. Sheng, Z. Zhang, *Int. J. of Food Properties*, **20(2)**, 1515-1522 (2017)
14. A. Moreda-Piñeiro, A. Fisher, S.J. Hill, *J. Food Compos. Anal.*, **16**, 195-211 (2003)
15. S.H. Yan, *J. Near Infrared Spec.*, **6**, 313-325 (2005)
16. X. Wang, J. Huang, W. Fana, H. Lu, *Anal. Methods*, **201**, 787-792 (2015)

17. A. Pallacios-Morillo, A. Alcázar, F. de Pablos, J. M. Jurado, *Spectrochimica Acta A*, **103**, 79–83 (2013)
18. J. J. Clary, *The Toxicology of Methanol* (1st ed., Hoboken: Wiley, 2013)