Short communication

A model based on clusters of similar color and NIR to estimate oil content of single olives

Claudio Fredes [1,3]*, Constantino Valero [2], Belén Diezma [2], Marco Mora [3,4]*, José Naranjo-Torres [4], Manuel Wilson [3], Gabriel Delgadillo [3]

[1] Departament of Agricultural Science, Universidad Católica del Maule; cfredes@ucm.cl Curicó 3480112, Maule, Chile.

[2] Laboratorio de Propiedades Físicas (LPF_TRAGRALIA), ETSIAAB, Universidad Politécnica de Madrid, constantino.valero@upm.es; belen.diezma@upm.es, Madrid 28040 Spain.

[3] Laboratory of Technological Research in Pattern Recognition, Faculty of Engineering Science, Universidad Católica del Maule, Talca 3480112, Maule, Chile.

[4] Department of Computer Science and Industries, Faculty of Engineering Science, Universidad Católica del Maule, Talca 3480112,Chile.

* cfredes@ucm.cl, mmora@ucm.cl

Abstract: The color and NIR spectrum are key to build an oil estimation model, thus it requires individual olives clustering before the Sohlext oil extraction method can be applied. The objective was to analyze an OC estimation model of individual olives, based on cluster of similar color and NIR spectrum in different combination of the first and/or the second season. This study was performed with Chilean Arbequina olives in 2016 and 2017. The descriptor of the cluster consisted of the 3 color channels of c1, c2, c3 color model plus 11 reflectance points between 1710 and 1735 nm of each olive, normalized with the Z-score index. Clusters of similar color and NIR spectrum were formed with the k-means++ algorithm, leaving a sufficient amount of olives to be able to perform the Sohlext analysis of OC, as reference value. The estimation models were based on the Support Vector Machine. The test was carried out with the Leave One-Out Cross Validation in different training-testing combinations. The best model predicted the OC with 6% and 13%deviation respect to the real value in one season by itself and when one season tested with another season, respectively. The use of clustering in estimation model is discussed.

Keywords: infrared spectroscopy; visible image; support vector machine; olive quality.

1. Introduction

The color of each olive has always been valued for farmers to estimate the harvest because of its close relationship with the quantity and quality of oil. The target olive color classifier is the maturity index (MI) [1] which first considers five sequential colors: green, yellow, red, purple and black, and then takes the advancement of black from skin to drupe as a harvest criterion. The authors [1], developed an automated MI using Computer Vision reaching R2 of 91%. The colors of the olives have been classified using models based on vector supports for the analysis of their image [2]. Currently, color has not stopped being used to build models for estimating oil content (OC), as the first classification criterion although color does not sufficiently explain OC, whereas the NIR spectrum can [3,4].

NIR technology has a high penetration of laboratory use and oil press as a tool for estimation OC in olive paste, but not in fresh olives, because the low sample variability estimates the performance variability of the samples lower the estimation performance [5]. The management of these variables has been approached by applying Partial Least Square regressions (PLS) or principal components regressions, allowing for the control of the variables that affect the OC, such as the number of samples and maturity of the olives [6], variety [7], fresh weight [3], olive size and year [5], among others. Another problem with the OC estimation models in individual olives is that the heteroscedasticity assumption is frequently breached, that is, the variance of errors is not regular in all the observations of individual olives, ultimately threatening the performance of the model. To optimize the management of variability in the OC estimation, a progressive clustering methodology has been proposed by similar color, NIR spectrum and OC subsequently, improving significantly the prediction error of the model [3].

The official measurement of the OC using traditional Sohlext methodology [8] needs more than a single olive to be properly determined, necessarily requiring a batch of them, regardless of the individual characteristics of each olive in the group; before taking the oil measurement, the individuals are ground and mixed until the sample is homogenized [8]. However, when one wants to know the OC of individual olives, the Nuclear Magnetic Resonance (NMR) and/or the micro-Sohlext method give good results [3,5,9]. However, most oil mills do not have this type of microanalysis, let alone NMR, due to its high human and technical resource requirement. And although Sohlext is expensive, labor-intensive, slow, has environmental negative-effects and labor risk compared with non-destructive methods [10], it is still the "ground of truth" that the non-destructive estimation methods should be used to construct estimation models that should be used to construct the non-destructive estimation models.

In this study, non-destructive mathematical clustering by homogeneity will be tested as a way to replace the destructive homogenization of cluster members according to Sohlext [8]. We proposed that OC can be estimated with traditional SOHLEXT from a cluster of olives, previously measured and grouped by color homogeneity and NIR spectrum and it could represent the OC of each olive belonging to the cluster. The objective was to analysis an estimation model of individual olive OC, based on clusters of similar color and NIR spectrum in different combination of the first and/or the second season.

The structure of this paper is as follows. Section 2 presents the materials and methods used in this research. Section 3 shows the experiments performed and presents the results obtained. Section 4 shows the results analysis discussion. Section 5 presents the patent required the procedure proposed.

2. Materials and Methods

As a summary of the method for estimating the oil content of individual olives in this study, Figure 1 shows innovation of the establishment of similar olive clusters to assign oil content of the entire cluster to each olive belonging to it.

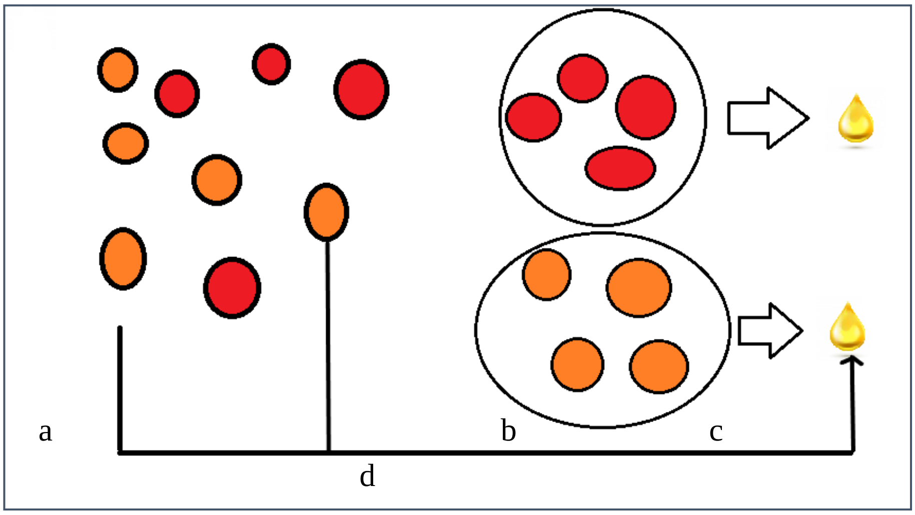

Figure 1. Sequential scheme of the study: a) Fresh olives collected; b) olives grouped by color similarity and NIR using k-means++ clustering; c) Measurement of the real OC of the cluster and its assignment to each olive in the cluster; d) Training and testing of an OC estimation model of two olives, similar in color and NIR, belonging to the same cluster.

2.1 **Obtaining the color and NIR characteristics of the olives**

Between March and June of the 2016 and 2017 seasons, flawless olives of all colors and sizes were randomly collected from all parts of trees from two rows marked from the super intensive olive grove (Olivas Don Rafael", (coordinates -35.1188358, -71.270495,13), Maule Region, Chile. The samples were taken to the Laboratory for Technological Research in Pattern Recognition of the Universidad Católica del Maule, where they were arranged in 24-hole trays to measure their color and NIR characteristics on the same day. The trays containing a wide variety of intact olives were photographed with a Sony Alpha a58 SLR camera placed in controlled environment with diffuse halogen lighting inside. To eliminate errors in the images, such as stakes and imperfections attached to the skin of the fruit, the image was binarized black/white with the OTSU algorithm [11]. Then, using morphological structuring elements with circular closures [12] the images with defects were eliminated (Figure 1). Afterwards, the images were converted to the c1c2c3 color model, which is invariant to lighting, to ensure the color value against possible lighting variations [13]. The mathematical formula of this color model c1c2c3, is expressed by the RGB color model function, which is presented below:

$$\boldsymbol{c1} = \frac{\boldsymbol{R}}{\mathbf{max}(\boldsymbol{G},\boldsymbol{B})};\ \boldsymbol{c2} = \frac{\boldsymbol{G}}{\mathbf{max}(\boldsymbol{R},\boldsymbol{B})};\ \boldsymbol{c3} = \frac{\boldsymbol{B}}{\mathbf{max}(\boldsymbol{R},\boldsymbol{G})} \quad (1)$$

Then, for each identified olive, the NIR spectrum was measured with an Ocean Optics NIR 900-2200 nm spectrophotometer with 512 spectral points with InGaAs array detector in reflectance mode. The wave length between 1710 and 1735 nm was considered, based on the report of [6,14] indicated the spectral absorption of lipids in fresh olives is around 1725 nm. The olives were kept frozen at -20°C waiting for the end of the harvest season and assigning a specific cluster for each olive.

2.2 **Grouping of the olives for their determination of oil**

The clustering of olives were constructed utilizing 14 descriptors corresponding to the c1c2c3 channels of the color model and 11 NIR spectral points every 2.53 nm between 1710 and 1735 nm, normalized according to the z-score indicator (Z), which was calculated as the difference with the mean in respect to the standard deviation of the data sets.

$$Z = \frac{x - \bar{x}}{sD} \quad (2)$$

Clustering was performed by increasing depth levels, respecting the minimum level of 30 olive per group at each level, sufficient amount to perform sample and counter sample of OC by the Sohlext analysis. The clustering was performed calculating a representative point for each group to be obtained (centroid), based on measures of similarity between these points using k-means algorithms that identified k as the number of centroids and allocates every data point to the nearest cluster, while keeping the centroids as small as possible [15]. A report of the olives belonging to each cluster, allowed the cluster to be manually assembled from the olives identified in the trays. The diversity of olives collected allowed for building enough clusters in the 2016 and 2017 seasons.

The OC of each cluster was carried out with a six-unit automatic Sohlext extractor. The analysis consisted of grinding the fresh olives with a hand homogenizer mixer, weighing and drying to obtain between 3 and 5 grams of homogeneous dry mill. The samples were put into paper cartridges and then introduced to the Soxhlet extraction units. After 6 hours of extraction, the samples were cooled in a glass desiccator with a porcelain plate and then gravimetrically weighted with a precision balance to obtain the oil percentage based on dry matter [8]. The Soxhlet oil analysis allowed for the production of six samples per day. This OC was assigned as the reference oil for each olive in the cluster and according to our hypothesis of similarity, of each single olive belonging to it.

2.3 OC estimation model and validation.

The estimation model was based on Support Vector Machines (SVM), which is a powerful methodology for solving problems of non-linear classification. This model is based on supervised training techniques with hyper-parameters for its construction, and that resolves optimization problems associated with the search of vectors [16]. There are two optimization problems of the model; the first is to adjust the method's specific parameters and the second is to find the hyper-parameters of the SVR. In this work, the hyper-parameters were established using the Simulated Annealing heuristic search algorithm (Simulated Annealing) [17]. To search for these parameters, an error and data set were defined, which was the same that generated a smaller error, corresponding to this training set.

To evaluate the model accuracy, the root square mean error of cross-validation (RSMECV) was considered. Calibration models were evaluated using the cross-validation test LOOCV [18]. Samples were taken from each 2016 and 2017 seasons using 70% of the olives for training (E) and 30% for testing (T). Additionally, two ordered seasons were considered in the following set of training (E) and testing (T) were: a) E: 100% Olives 2016 and T: 100% Olives 2017; b) E: 100% Olives 2017 and T: 100% Olives 2016; c) E: 50% Olives 2017 + 50% Olives 2016 and T: 50% Olives 2017 + 50% Olives 2016; d) E: 70% Olives 2017 + 70% Olives 2016 and T: 30% Olives 2017 + 30% Olives 2016; e) E: 80% Olives 2017 + 80% Olives 2016 and T: 20% Olives 2017 + 20% Olives 2016. Sets c, d, and e were trained and tested 10 times selecting random samples. In these sets, the standard deviation and the percentage of deviation in regards to the real value was considered.

3. Results

3.1 Treatment images and spectrum of olives

The removal of distractors, such as peduncles and skin-attached imperfections, was performed without deforming the curvature of the objects [11,12] (Figure 2).

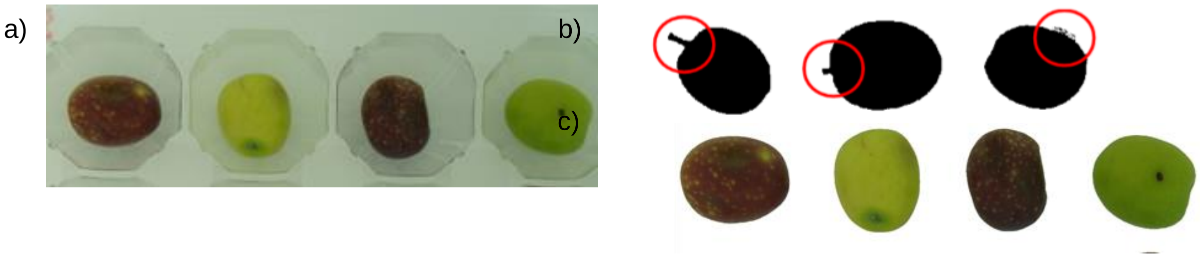

Figure 2. a) Original images of a group of olives inside a tray; b) Black/White image with three defective olives c) Final segmentation of group olives

Segmentation of the images from the bottom was performed with the c1c2c3 model in c3 channel, as proposed [19]. Figure 3 shows a histogram in the c3 channel olive trays, highlighting the classifying power based on the color of the olives [13].

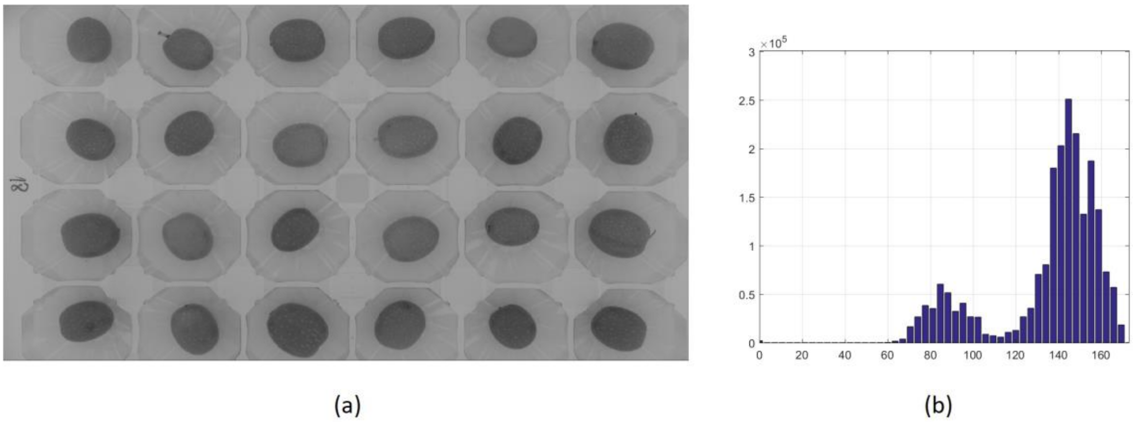

Figure 3. a) Tray of recently collected olives, registered in c3 channel of the c1c2c3 model. b) The c3 Channel Histogram of this tray of olives.

Figure 4 shows the complete NIR spectrum of the olives contained in a tray, the mark (red star) in the graph indicates the range corresponding to the wavelengths between 1710 and 1735 nm, considered for this study

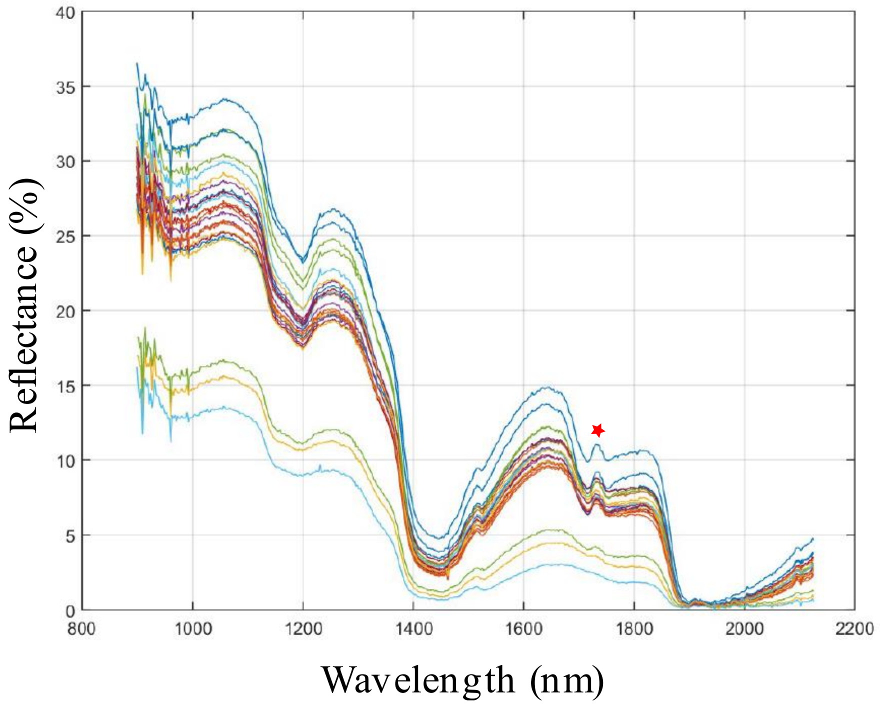

Figure 4. NIR spectrum on a tray of olives, highlighting the spectral range used (red star).

3.3. Cluster of similar olives

The result of the clustering was the formation of 31 and 29 groups of 30 olives with similar NIR/color characteristics in the 2016 and 2017 seasons respectively, which were analyzed for their OC. The clustering algorithm used was the K-means++, which is more robust than the traditional K-means algorithm and is more responsive to the starting position of the centroids, thus allowing for a better choice of the initial values of the K-means clustering, and therefore avoiding the formation of deficient clusters [15]. Figure 5 shows the histograms of the OC of the 2016 and 2017 clusters and their statistical parameters. It can be observed that the variability of the year 2016 was much greater than 2017, although differences between the seasons were expected [5].

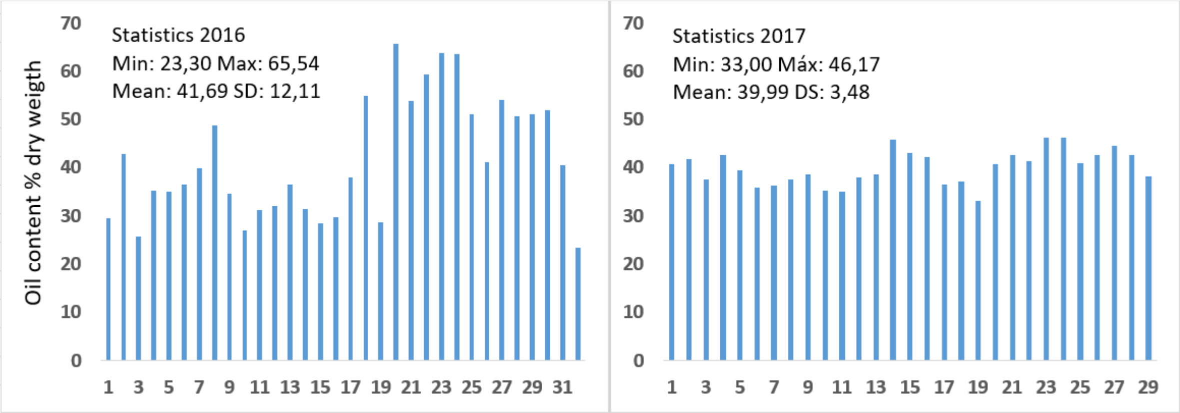

Figure 5. Histograms of OC and statistics of the year 2016 and 2017.

3.4. Estimation model of oil from single olives

The Support Vector Machine (SVR) methodology allowed to build models for estimating the OC of each single olive belonging to a cluster, in the different data set evaluated, generating an error that the SVR minimizes in each set. This error was obtained with LOOCV. Different combinations of training and testing of the model within each set of 2016 and/or 2017 olives were randomly performed 10 times. Table 1 shows the root mean square error of cross-validation (RMSECV) of the estimation model.

Table 1. RMSECV of the OC estimation model based on clustering by similar color and NIR in different training and testing sets with olives from the 2016 and/or 2017 seasons.

Training Sets 2016/17	Testing Sets 2016/17	RMSECV	
100% Olives 2016	100% Olives 2017	7.35	
100% Olives 2017	100% Olives 2016	8.64	
		Mean 10 RMSECV	**SD 10 RMSECV**
70% Olives 2016	30% Olives 2016	3.1	0.4
70% Olives 2017	30% Olives 2017	3.5	0.3
50% Olives 2017 + 50% Olives 2016	50% Olives 2017 + 50% Olives 2016	7.34	0.61
70% Olives 2017 + 70% Olives 2016	30% Olives 2016 + 30% Olives 2017	7.13	0.64
80% Olives 2017 + 80% Olives 2016	20% Olives 2017 + 20% Olives 2016	7.21	0.12

The results of the model in one season, RMSECV were 3.1 and 3.5 as well as average deviation of the real value of 6 and 7%, for 2016 and 2017 respectively. However, in the two seasons, the set with the best performance (80% training 2016 and 2017 and 20% testing 2016 and 2017) obtained a RMSECV 7.21 and average deviation of a real value of 13%.

4. Discussion

The color and NIR spectrum of each olive are characteristics that are highly related to its oil content; however, a single olive is not enough to quantify OC with the official Soxhlet method, thus requiring a group of olives. Previous research [3] pointed out that the current batch-based assessment of the OC (determined by Soxhlet) in mills only reproduces 44% of the underlying heterogeneity, despite being the factory standard, however the incorporation of individual NIR spectra to the model allowed for the increase to 67% explanation of the OC (%) of olives. Clustering by similarity in the OC estimation models of individual olives allowed to control the variability of the sample, resulting in better final performances [6,3]. In this study, olive clustering by similar color and NIR achieved good performances in each season by itself (mean deviation of the real value of 6 and 7% in 2016 and 2017, respectively); this indicates that the hypothesis makes sense for each measurement season. However, the RMSECV practically doubled in two seasons, possibly because the error is multiplied by the combination of two variables. There are not precedents for similar studies to compare these results, but rather a consensus regarding the natural increase in error in more than one season [20,6], which makes it difficult to transfer calibrations year after year in the context of annual changes in the maturation of olives [20,22].

Another angle of this study is related to capturing the NIR and color variables of the model, which in this case were acquired with different equipment, further studies with multi-spectral cameras would allow for simultaneously measuring NIR and color by the same equipment. According to [3], the estimation of the OC from the multi-spectral images would add to the estimation of the weight of the olives, which should provide excellent results in the OC estimation model.

Regarding the criteria to construct the clustering descriptor, in this study comprised of 21% of color characteristics and 79% NIR characteristics around 1,720 nm range. The performance could be improved by including other NIR spectrums related to OC, given the high sensitivity of the model to this variable [9,21]. In [3] it was found that the NIR spectra with more information on OC are between 1153 and 1231 nm, that should be explore in further analysis. Another variation could be to find the weight of each variable descriptor, since color is not as good a descriptor of OC as the NIR spectrum. The maturity entails a color and oil change that depends on the characteristics of the variety, soil and climate of each year [22], thus it is difficult to control multiple variables in one descriptor. Models based on a single variable cluster; the error should be directly proportional to the size of the cluster as well as indirectly proportional to similitude their members. The direct measurement of the OC in each fruit with MNR or micro-Soxhlet is an option, but when by traditional Sohlext is used for measuring, it is necessary to cluster the fruit, in fact many models could not have been made without previously grouping the olives [3,5,7,9,14]. Therefore, future studies should establish mathematical clustering criteria to optimize estimation models.

6. Patents

Estimation method for individual olive oil based on non-destructive technologies (WO2019041055A1). WIPO (PCT) in 2020.

Supplementary Materials:

Table S1: Sohlext oil analysis worksheet by cluster, season 2016.

Table S1: Sohlext oil analysis worksheet by cluster, season 2017.

Author Contributions: C.F. Investigation, Funding acquisition, Writing and original draft; C.V. Funding acquisition, Review & editing; B.D. Funding acquisition, Review & editing; M.M. Funding acquisition, Conceptualization, Methodology, original draft; J.N.-T. Investigation, Software, Resources. All authors have

read and agreed to the published version of the manuscript. M.W. - Investigation, Software, Resources. G.D. Investigation, Software, Resources.

Funding: This research was funded by the foundation for the Promotion of Scientific and Technological Development (Fondo de Fomento al Desarrollo Científico y Tecnológico) from the National Agency of Investigation and Development, Chile, grant number ID15I10142. Partial funding was received from the Universidad Politécnica de Madrid, cooperation grant AL14-PID-11 "Estimación de calidad y contenido en aceite de aceitunas mediante procedimientos ópticos VIS y NIR"

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Guzmán, E.; Baeten, V.; Pierna, J.A.F; García-Mesa J.A. Determination of the olive maturity index of intact fruits using image analysis. J Food Sci Technol. 2015, 52, 1462-1470

2. Avila F.; Mora, M.; Oyarce, M.; Zuñiga, A; Fredes, C. A method to construct fruit maturity color scales based on support machines for regression: Application to olives and grape seeds. J Food Eng. 2015, 162, 9–17.

3. Correa, E.C.; Roger, J.M.; Lleó, L.; Hernández-Sánchez, N.; Barreiro, P.; Diezma, B. Optimal management of oil content variability in olive mill batches by NIR spectroscopy. Sci Rep, 2019, 9, 1-11.

4. Cayuela, J.A; Camino, M.P. Prediction of quality of intact olives by near infrared spectroscopy. Eur. J. Lipid Sci. Technol. 2010, 112, 1209-1217

5. Salguero-Chaparro y Peña-Rodríguez (2014). On-line versus off-line NIRS analysis of intact olives. LWT Food Sc Tech. 2014, 56, 363-369.

6. Fernández-Espinosa, A.J. Combining PLS regression with portable NIR spectroscopy to on-line monitor quality parameters in intact olives for determining optimal harvesting time. Talanta 2015, 148, 216–228.

7. Kavdir, I.; Buyukcan, M.B.; Lu, R.; Kocabiyik, H.; Seker, M.Prediction of olive quality using FT-NIR spectroscopy in reflectance and transmittance modes. Biosyst Eng. 2009, 103, 304-312.

8. AOAC. Official methods and recommended practices of the AOCS of the American Oil Chemists Society, Official methods and recommended practices of the AOCS, 7th ed.; Publisher: The American Oil Chemists' Society, AOAC, USA, 2017; pp. 154–196.

9. Walton, J.H.; Gardner, J.M.; Ferguson, L. Determining the oil and water content of single olives using magnetic resonance imaging (MRI) spectroscopy. In VIII International Olive Symposium. Acta Horticulturae 1199, Split, Croatia, October 2016, Klepo, T; Ferguson, L; Sebastiani, L; Perica, S.; Vuletin, G. (Eds); Publisher: International Society for Horticultural Science. pp. 511-516.

10. Casson, A.; Beghi, R.; Giovenzana, V.; Fiorindo, I.; Tugnolo, A.; Guidetti, R. Environmental advantages of visible and near infrared spectroscopy for the prediction of intact olive ripeness. Biosyst Eng. 2020, 189, 1-10.

11. Otsu, N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979, 9, 62–66.

12. Wong K.; Sahoo, P.K. A gray-level threshold selection method based on maximum entropy principle. IEEE Trans Syst Man Cybern. 1989, 4, 866–871.

13. Ibraheem, N.A.; Hasan, MM.; Khan, R.Z.; Mishra, P.K. Understanding color models: a review. ARPN J Sc Techn. 2012, 2, 265-275.

14. Manley, M. Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. Chem Soc Rev. 2014, 43, 8200-8214

15. Arthur, D.; Vassilvitskii, S. K-means++: the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, Philadelphia, USA, Juanary 2007, Hal Gabow (Ed), Publisher: Society for Industrial and Applied Mathematics, Philadelphia, States United, 2007; pp. 1027–1035.

16. Cortes, C.; Vapnik, V.; Support-vector networks. Mach. Learn. 1995, 20, 273–297.

17. Borgatti SP; Everett MG; Models of core/periphery structures. Soc Net. 1999, 21, 375–95.

18. Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminformatics, 2014, 6, 1-15.

19. Avila, F; Mora, M., Fredes, C. A method to estimate Grape Phenolic Maturity based on seed images. Comp. Electron. Agricult. 2014, 101, 76–83.

20. Stella, E.; Moscetti, R.; Haff, R.P.; Monarca, D.; Cecchini, M.; Contini, M.; Massantini, R. Recent advances in the use of non-destructive near infrared spectroscopy for intact olive fruits. J Near Infrared Spectrosc. 2015, 23, 197-208.

21. Giovenzana, V.; Beghi, R.; Romaniello, R.; Tamborrino, A.; Guidetti, R.; Leone, A. Use of visible and near infrared spectroscopy with a view to on-line evaluation of oil content during olive processing. Biosys Eng. 2018, 172, 102-109.

22. Benito, M.; Lasa, J.M.; Gracia, P.; Oria, R.; Abenoza, M.; Varona, L.; Sánchez-Gimeno, AC. Olive oil quality and ripening in super-high-density Arbequina orchard. J. Sc. Food Agr. 2013, 93(9), 2207-2220.