COFFEE SCIENCE
ISSN: 1984-3909

eISSN: 1984-3909
Coffee Science, e161710, 2021

Doi: https://doi.org/10.25186/.v16i.1710

Classification of the maturity stage of coffee cherries using comparative feature and machine learning

Sebastián Velásquez[1*], Arlet Patricia Franco[2], Néstor Peña[3], Juan Carlos Bohórquez[3], Nelson Gutiérrez[4]

[1]Research & Development department, Industria Colombiana de Café, Medellin, Antioquia, Colombia
[2]Universidad Pontificia Bolivariana, Desarrollo y Aplicación de Nuevos Materiales/DANM, School of Engineering and Architecture, Montería, Colombia
[3]Universidad de los Andes, Department of Electrical and Electronics Engineering, Bogotá, Colombia
[4]Universidad Surcolombiana, Department of Agriculture Engineering, Huila, Colombia
Contact autors: svelasquez@colcafe.com.co, arletfranco@gmail.com, jubohorq@uniandes.edu.co, npena@uniandes.edu.co, ngutierrezg@usco.edu.co
Received in January 21, 2020 and approved in January 7, 2021

ABSTRACT

This work presents the use of multiple techniques (i.e., physicochemical and spectral) applied to harvested coffee cherries for the postharvest classification of the maturity stage. The moisture content (MC), total soluble solids (TSS), bulk density, fruits' hardness, CIEL*a*b parameters and the dielectric spectroscopy methods were applied on coffee cherries at seven maturity stages. These maturity stages were assessed according to the days after flowering (DAF) and the physical appearance as traditionally performed by growers. An increase of the green-to-red ratio (i.e., a*) parameter was perceived, accompanied by a monotonic response for the hardness, TSS and bulk density with a maximum moisture content at stage 5. In the case of the dielectric spectroscopy technique, the loss parameter presented higher losses for unripe stages at the ionic conduction region. To compare the individual performance of each of the techniques, three machine learning methods were used: random forest (RF), support vector machine (SVM) and k-nearest neighbours (k-NN). The meta-parameters for these techniques were optimized for each case to achieve the best performance possible. Furthermore, as the dielectric response is of spectral nature, recursive feature selection was applied and the 500 MHz to 1.3 GHz frequency range selected for the task. The highest performance was obtained for the colorimetric (75.1%) and hardness (72.5%) responses, while the lowest was obtained for the moisture content (45.5%). The dielectric spectroscopy response presented a promising response (56.8%), that achieved a clear separation of unripe from ripe stages, except for stage 5 in which some of the samples were classified as stage 2. Most techniques studied are compatible with field conditions, and the dielectric technique shows potential to be transferred based on available software-radio defined platforms.

Key words: Dielectric spectroscopy; Coffee maturity; Postharvest classification; Physicochemical analysis.

1 INTRODUCTION

The processing of coffee represents an important challenge to the industry as it considerably affects the quality and represents the highest costs to the producers (Arcila et al., 2007; Farah, 2012; Sunarharum; Williams; Smyth, 2014). Colombia has traditionally produced coffee by the wet method. In this method, once the cherries are hand-picked, they are pulped, fermented, and dried in the coffee farm or processing units. The dry parchment coffee beans are then hulled and roasted or exported. Most of these processing practices are performed by traditional artisan techniques.

One of the most essential stages is the hand-picking and sorting of the ripe coffee cherries. For instance, the Colombian coffee industry performs a selective and manual harvesting of ripe cherries to avoid affecting the sensory attributes of the beverage (Marín et al., 2004; Puerta-Quintero, 2000). As the maturity of the cherries if not achieved uniformly, several harvest sessions are required (Craig et al., 2015). The selection process is performed according to the appearance of the cherry and completely depends on skilful human labour. The coffee cherry harvesting represents around 40% of the production costs for Colombian coffee grower (Federación Nacional de Cafeteros - FNC, 2010).

The use of field-compatible approaches as the total soluble solids measured by the refractive index which correlates to the total soluble solids, quantitative measurement of colour, titratable acidity, and equatorial firmness are appealing approaches that might be closer to the growers' requirements (Arcila et al., 2007; Silva et al., 2014). These are variables that can increase or reduce with the cherry development; for example, the sugar content of the pulp increases with maturity and the pericarp softens (De Castro; Marraccini, 2006). One of the drawbacks of these techniques is that some of them are not batch compatible.

The dielectric spectroscopy technique at the radio-frequency and microwave range has been of interest for quality assessment applications of multiple crops (Sosa-Morales et al., 2010), it can process batches and simple samples, and it can be transferred to field compatible equipment. The complex permittivity of the samples is indicative of the molecular composition, and is represented by a complex number (Equation 1) that depends on the frequency and temperature (Franco et al., 2015).

$$\hat{\varepsilon} = \varepsilon' - j\varepsilon'' \quad (1)$$

The real part of $\hat{\varepsilon}$ is the relative electrical permittivity or dielectric constant (ε'), and the imaginary part of is the dielectric loss factor (ε'') with $j = \sqrt{-1}$. The electrical permittivity represents the ability of the material to polarize and to store electric energy in response to an applied electric field, and the loss factor is associated with the dissipation of energy as heat (Sosa-Morales et al., 2017).

As the maturity of many fruits and vegetables consists on the formation of sugars or the increase or loss of bounded and unbounded water, the dielectric spectroscopy approach has been used to assess the maturity of different fruits (Castro-Giraldez et al., 2010; Guo et al., 2015). This method has been also considered to study the water features of green and roasted coffee (Berbert et al., 2008; Iaccheri et al., 2015).

Finally, the development of analytical based models has employed machine learning techniques for this purpose. As a considerable wavelength and frequency information is generated, the selection for critical features and qualitatively describe the reasons for this selection should be considered during the construction (Li et al., 2018). To allow efficient technological transfer, the meta-parameters considered for different machine learning methods have to be optimized and selected to produce the best model possible (Fashi; Naderloo; Javadikia, 2019). Furthermore, it is ideal not to rely on a single scheme but to explore as many techniques as possible as some techniques have better behavior than others for specific tasks (Yang et al., 2019). Finally, all these elements merge together in the construction of a pipelined structure where feature selection, dimension reduction, and meta-parameter optimization are combined to produce simple and accurate models (Piedad et al., 2018). This strategy can be combined to consider multiple factors or in this case processing stages that support growers' in the complicated task of providing quality consistence (Rungpichayapichet et al., 2016).

Successfully results in maturity prediction of fruits using machine learning have been previously reported. For papaya (Santos Pereira et al., 2018), combined digital image features and random forest to develop a model to the ripening prediction. Combination of hand-crafted image features with machine learning techniques allowed to reach higher accuracy, improving the classification models for papaya. On the other hand, the work by (Behera; Rath; Sethy, 2020) employed local binary pattern (LBP), histogram of oriented gradients (HOG), Gray Level Co-occurrence Matrix (GLCM) and k-nearest neighbor (KNN), and support vector machine (SVM) among others, and it was possible the development of a classification model for maturity prediction with 100% of accuracy and 0.0995 s training time. Classification of cherries regular and irregular shaped was developed (Momeny et al., 2020) employing HOG and LBP to extract the features of the images of cherries and KNN, ANN, Fuzzy and EDT algorithms to classify them. Results were compared with the Convolutional Neural Network (CNN) method, which was able to classify cherries with accuracy around 99% in all image sizes.

The aim of this study was to evaluate the physical features of coffee cherries that could operate as global maturity classifiers, regardless of the cultivar. Different physicochemical approaches compatible with the coffee cherry and the dielectric spectroscopy technique were used for this purpose. Machine learning was employed to the feature's comparison and classification according to defined maturity stages: days after flowering and physical appearance of the cherries.

2 MATERIAL AND METHODS

2.1 Coffee samples

Four coffee cultivars were used for this study: two from northern Huila (NH1 / NH2), one from southern Huila (SH1), and one from Caldas (CA), all coffee producing regions in Colombia. All coffee cherries were collected during the 2018 harvest, and their features are presented in Table 1.

Around 10 kg of fresh cherries collected at the plantations were stored at 8 °C and transported to the South Colombian Coffee Research Centre – CESURCAFE pilot plant during a period not superior to 6 – 8 h. Consequently, the sample set was constituted by 84 coffee cherry samples: 7 maturity stages, 4 cultivars and 3 biological replicates. Stage 1 represents green unripe cherries, which appear 196 d after flowering (Marín et al., 2004). Stage 2 corresponds to green-yellow cherries (i.e., 203 d), Stages 3 and 4 to almost ripe or "pintón" cherries (i.e., 208-215 d. Stages 5 and 6 represent ripe cherries (i.e., 217-224 d) while Stage 7 (i.e., after 224 d) represented overripe fruits. The difference between stages 3 and 4 were assessed according to physical appearance, as traditionally performed by growers. The same strategy was considered for stages 5 and 6. All cultivars considered were red varietals.

Table 1: Coffee cultivars information: Northern Huila (NH1 / NH2), Southern Huila (SH1) and Caldas (CA).

Label	Variety	Municipality	City	Department	Altitude [amsl]	Average temperature
NH1	Colombia Rojo	Las Juntas	Santa María	Huila	1750	21.6 °C
NH2	Colombia Rojo	Los Pinos	Santa María	Huila	1630	22.5 °C
SH	Castillo	El Piñal	Gigante	Huila	1450	22.0 °C
CA	Castillo Naranjal	Canaan	Viterbo	Caldas	1300	29.0 °C

2.2 Colorimetric and physical analysis

The colorimetric analysis was performed with a Konica Minolta CR-410 chroma-meter (New Jersey, USA) evaluating parameters as *L**(lightness), *a**(redness-greenness) and *b**(yellowness-blueness). Analysis were performed in triplicate form, recording 9 measurements for each sample. Colour difference was estimated using Equation 2.

$$\Delta E = \sqrt{\left(\Delta L^*\right)^2 + \left(\Delta a^*\right)^2 + \left(\Delta b^*\right)^2} \quad (2)$$

All physical analyses were performed in triplicate. Total soluble solids as percentage were determined by an Atago PR 201 portable digital refractometer (Tokyo, Japan) and are presented as °Brix. Total soluble solids for Stages 1 and 2 could not be obtained as mucilage could not be extracted. The moisture content for the samples was calculated according to the wet basis gravimetrical method based on the ISO 6673 standard (Adnan et al., 2017). Empty aluminium tins to contain the samples were weighed with a 4-digit precision scale (i.e., WT). Then, a 5 g coffee cherry sample was deposited into the tin to obtain the wet weight (i.e., WW+WT). These full tins were heated inside a Memmert 55 oven (Schwabach, Germany) set at 105 °C, and the samples were dried for 24 h. After removing the samples from the oven, the tins were weighed again to obtain the dry weight (i.e., DW+WT). The moisture content of each sample was computed according to Equation 3 (Adnan et al., 2017).

$$MC(\%) = \frac{(WW + WT) - (DW + WT)}{(DW + WT) - WT} \cdot 100 \quad (3)$$

The traditional weight over volume method was used to calculate the bulk density. The samples were deposited in a beaker. The volume that the samples occupied was measured, and the weight of the samples was recorded. The bulk density was calculated as the division of the weight of the coffee cherries by their volume.

The hardness measurements were completed with a Brookfield CT3 texture analyser (Pennsylvania, USA) equipped with a 50 kg load cell and a TA25/1000 cylindrical plane plunger. The plunger was set to advance at a velocity of 0.8 mm•s^{-1} until the pulp failed.

2.3 Dielectric spectroscopy characterization

The dielectric spectroscopy characterization was carried out using a Keysight 85070E open circuit coaxial probe (Santa Rosa, USA). Before performing the measurements, the vector network analyser was warmed up for at least 90 min and subsequently calibrated using the standard given by the equipment: air – open circuit, short block and load - distilled water. Three to four measurements were performed before the tests to verify the calibration and stability of the equipment.

Furthermore, after calibration, the probe and cable were not manipulated to avoid interferences. The 85070 v.E06.01.36 software (Agilent Technologies, Malaysia) controlled the network analyser to perform the frequency sweep and measure the complex reflection coefficient of the sample around the probe tip, which was then converted to the complex permittivity. The measurements were performed from 0.5 to 20 GHz. For each measurement, 1001 points were acquired with an IF-bandwidth of 10 kHz.

A mass of 40 g of coffee cherries stored at 4 °C for no longer than 1-2 weeks was mashed with a ceramic mortar until a homogenous paste was obtained. The coffee cherries paste was poured into a beaker and carefully placed on and around the probe tip, to avoid air bubbles around the sensing zone. All measurements were performed at 20 ºC and a relative humidity of 65%.

After each measurement, the probe was cleaned with distilled water and dried with soft paper. All the dielectric properties determinations were carried out in independent triplicates. Average values and standard deviations were calculated.

2.4 Data analysis

The first part of the study consisted in the evaluation of the seven maturity stages. Accordingly, a random effect multilevel model was considered. Maturity was considered as the fixed effect, and the variety was considered as the random effect (random intercept and slope). Then, Tukey pairwise comparisons were performed to identify the groups that differentiated the stages. For the dielectric spectroscopy response, the analysis was split in two parts. The first portion consisted in the identification of the most critical features. For this purpose, the technique of extremely randomized trees was used, in tandem with recursive variable elimination. In the first run, five-fold cross validation was performed to determine the number of trees that resulted in the highest maturity stage classification accuracy. Then, thirty random seeds were generated, and the process of recursive feature elimination performed for each seed. Interactions across some of the frequencies were also considered.

The remaining features were scaled and analysed with principal component analysis (PCA), because of the high collinearity across frequencies, and the first 50 components were obtained. The traditional SVD decomposition was used for this purpose, and the biplot of the scores and loadings presented.

The third stage consisted in the development of the classification models. In the case of the physicochemical attributes, each technique was individually considered.

Three types of classification techniques were considered: random forests, support vector machine (SVM) and k nearest neighbours (k-NN). The data was partitioned into two sets: 56 samples for training and 28 for validation, stratifying the observations by the interaction between maturity and variety. This partition was randomly performed 30 times, using the seeds previously computed, to obtain the statistics of the classification accuracy and confusion matrix results. This recalls the bootstrapping technique.

The hyper-parameters for each of the techniques were obtained by five-fold cross validation. The parameters to be calculated for each method were: number of trees for random forests, the sparsity regularization parameter and kernel function for SVM, and the number of neighbours used for k-NN. Then, each resulting model was evaluated with the validation set, and the classification accuracy and confusion matrix recorded. The latter were then averaged, and these correspond to the reported values.

In the case of the dielectric spectroscopy technique, an additional layer was considered in the pipelined classifier: the PCA components as exogenous variables. Cross validation was also considered to select the optimal number of components for each technique. Once selected, the process is identical to the approach used for the physicochemical features. As the dielectric spectroscopy response generates three different features (i.e., dielectric permittivity, loss factor and loss tangent), the feature with the best accuracy was selected. Multi-level random effects models were calculated with the *lm4* and *lsmeans* R packages version 3.3.6. The machine learning models were implemented with the *scikit-learn* package in Anaconda with Python 3.7.

3 RESULTS AND DISCUSSION

3.1 Colorimetric response

As the results for L* (*lightness*) and *b**(yellowness-blueness) are colinear, the plot between *a** and *b** are presented in Figure 1.

This plot presents an adequate distribution for each maturity stage as all varieties were red, and the only sample set that presented undesirable behaviour is the NH2- Stage 2 set, which recalls the NH2-Stage 3 set. Yet, having four varieties accounts for possible variations across samples and the biological samples present the consistency required. The dominating parameter for this experiment was *a**. Although the presence of yellow during stages 2 to 4 dominated the variation on *b** (yellowness-blueness), which slightly changed with the proportion of yellow in the cherries. The parameter *a**(redness-greenness) changed from green to red with the progress of maturity. The lightness *L** depicted a monotonically response, decreasing with the ripening process (Velásquez et al., 2019).

3.2 Physical analyses: the effects of maturity

The results for the features considered are presented in Table 2. As can be seen, all features were distributed between three to five groups. The moisture content and bulk density values agree with those reported in the literature (Aristizábal-Torres et al., 2012; Farah, 2012).

The hardness and TSS values variations are related to the natural ripening process of fruits, which resulted in fruit softening and increase of sugar content other compounds water-soluble (Bashir; Abu-Goukh, 2003). Identical behaviour for hardness and TSS was previously reported as well (Marín-López et al., 2003). Due to the spread response of the original

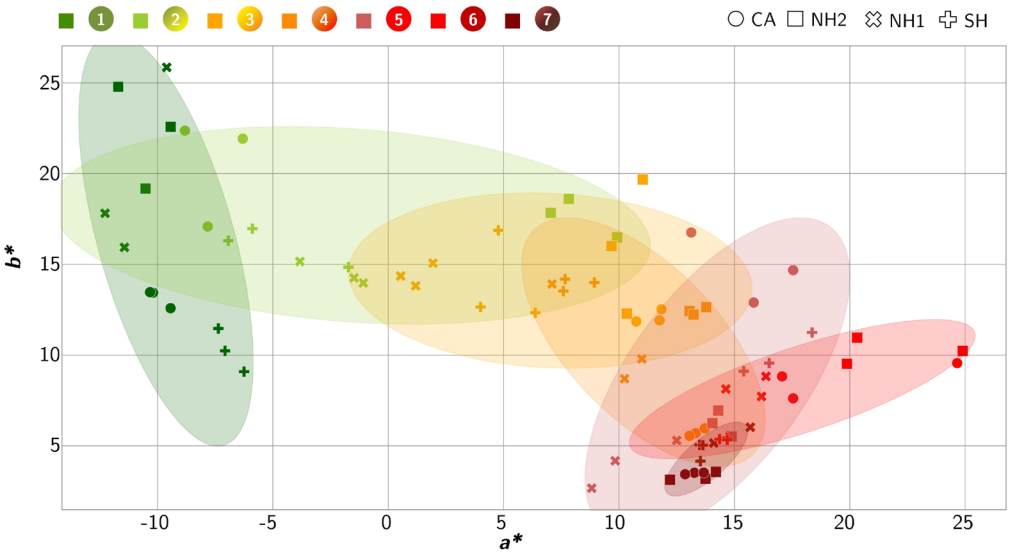

Figure 1: Colorimetric study: b* vs. a*, grouping of maturity stages for Northern Huila (NH1 / NH2), Southern Huila (SH1) and Caldas (CA) samples.

values, a natural logarithm transformation was performed on the data for modelling. This feature and the total soluble solids (TSS) were distributed in five groups.

In the case of the moisture content, stage 1 presented a value close to 70.5% and decreased in stages 2 and 3. Then, the moisture reached a maximum at stage 5 which corresponds with stage 6 to the ripe stages. Finally, it decreased at stage 7 due to pericarp absorption. Similar results were previously reported (Marín-López et al., 2003) for coffee fruits var. Colombia red, presenting the highest moisture content at stage 5 and then diminishing until stage 7.

In the case of bulk density, the higher bulk density values were seen for fully and green-yellow immature stages, and the lowest for mature stages. As can be seen in Table 2, this parameter increased with moisture content. This behaviour was previously reported (Chandrasekar; Viswanathan, 1999) coffee beans arabica and robusta varieties.

3.2.1 Permittivity and maturity

Figure 2 presents the dielectric permittivity response for all maturity stages and the frequency range between 0.5 GHz and 20 GHz. The x-axis, which represents the frequency, is in logarithmic scale.

The response varies from 60 down to 20 for the complete frequency range. The relative dielectric permittivity presented no statistical difference across the maturity stages. The lowest dielectric permittivity was seen for stage 7. Figure 3 presents the loss factor response.

Table 2: Mean values and standard errors for each considered technique. Superscripts represent categories across the calculated values and P-val represents the p-value of the analysis of variance for each model.

Stage	Moisture content (%)	Bulk Density [kg•m^{-3}]	L*	a*	b*	ΔE	TSS (°Brix)	Hardness [N]
Min	62.5	0.454	17.1	-12.3	2.7	19.4	4.9	3.1
1	70.4 ± 0.4^{ab}	0.527 ± 0.006^{b}	34.8 ± 4.2^{cd}	-9.6 ± 1.0^{a}	16.4 ± 2.8^{cd}	39.7 ± 5.1^{cd}	-	86.3 ± 6.3^{f}
2	68.1 ± 1.1^{a}	0.522 ± 0.010^{ab}	36.9 ± 6.0^{d}	-1.6 ± 3.9^{ab}	17.1 ± 3.4^{d}	41.2 ± 6.9^{d}	-	46.8 ± 8.9^{ef}
3	69.9 ± 1.1^{b}	0.501 ± 0.012^{ab}	33.0 ± 3.5^{cd}	7.0 ± 2.7^{bc}	14.1 ± 2.4^{cd}	36.8 ± 4.2^{bd}	8.0 ± 0.9^{a}	32.6 ± 9.2^{cdf}
4	71.8 ± 0.8^{b}	0.496 ± 0.008^{ab}	27.8 ± 4.3^{bd}	11.1 ± 2.1^{cd}	10.7 ± 3.2^{bcd}	32.0 ± 5.0^{cd}	10.4 ± 0.4^{b}	20.3 ± 7.3^{bc}
5	74.0 ± 0.6^{c}	0.494 ± 0.010^{ab}	25.5 ± 7.0^{abc}	14.3 ± 1.0^{de}	8.8 ± 4.8^{ac}	30.7 ± 8.4^{abc}	12.1 ± 0.4^{c}	14.6 ± 7.1^{abc}
6	70.7 ± 0.7^{ab}	0.493 ± 0.010^{ab}	24.2 ± 3.5^{b}	17.9 ± 2.4^{e}	8.1 ± 2.2^{ab}	31.3 ± 4.1^{ac}	14.4 ± 0.7^{d}	10.0 ± 6.5^{ab}
7	67.3 ± 1.6^{ab}	0.491 ± 0.010^{a}	18.6 ± 4.2^{a}	13.8 ± 1.3^{ce}	4.3 ± 3.0^{a}	23.3 ± 5.0^{a}	15.4 ± 0.9^{d}	5.9 ± 6.4^{a}
Max	76.1	0.544	49.5	24.9	25.9	56.6	18.0	114.6
P-val	0.002569	0.01727	0.0007161	1.985×10^{-5}	0.001134	0.002484	4.127×10^{-3}	5.57×10^{-6}

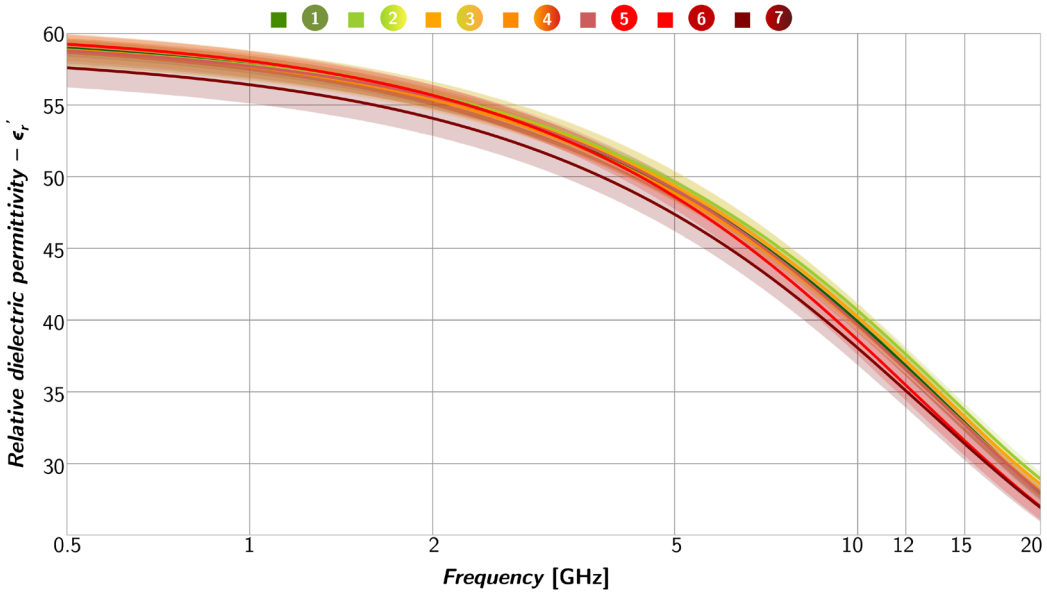

Figure 2: Dielectric permittivity response between 0.5 GHz and 20 GHz for coffee samples at each maturity stage. Solid lines represent mean values at each stage, and shadowed regions represent the standard errors.

Opposed to the dielectric permittivity, the loss factor presented a difference across maturity stages, especially for stage 1 (green unripe). A second group was seen for stages 2 to 4, and a third for stages 5 to 7. The frequency regions where this difference was more evident was between 0.5 MHz and 2 GHz, that corresponds to the ionic conduction region (Keysight Technologies, 2015). The relaxation frequency, which was close to 15 GHz for all stages, presented no difference across stages. The dipolar rotation region, which is mainly dominated by the presence of water, is not the predominating phenomena in the discrimination of the maturity stage. The response is typical of a salt-mineral solution (Franco et al., 2015), which might be indicative that the maturity in coffee responds to the mineral content of the pulp rather than to its water content (Farah, 2012).

Finally, the loss tangent is presented in Figure 4. As with the loss factor, the main separation across maturity stages was perceived between 0.5 and 2 GHz.

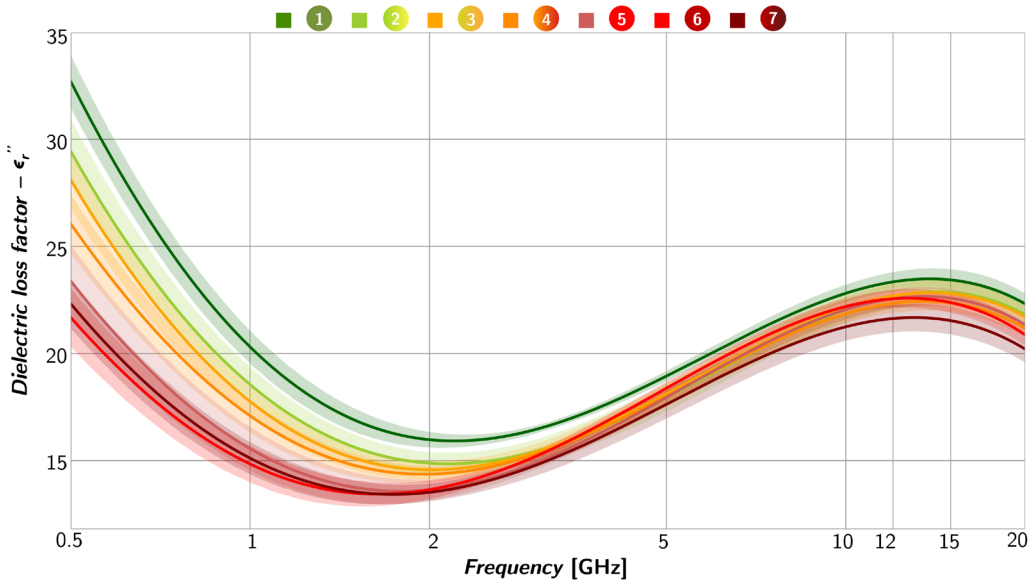

Figure 3: Loss factor response between 0.5 GHz and 20 GHz for coffee samples at each maturity stage. Solid lines represent mean values at each stage, and shadowed regions represent the standard errors.

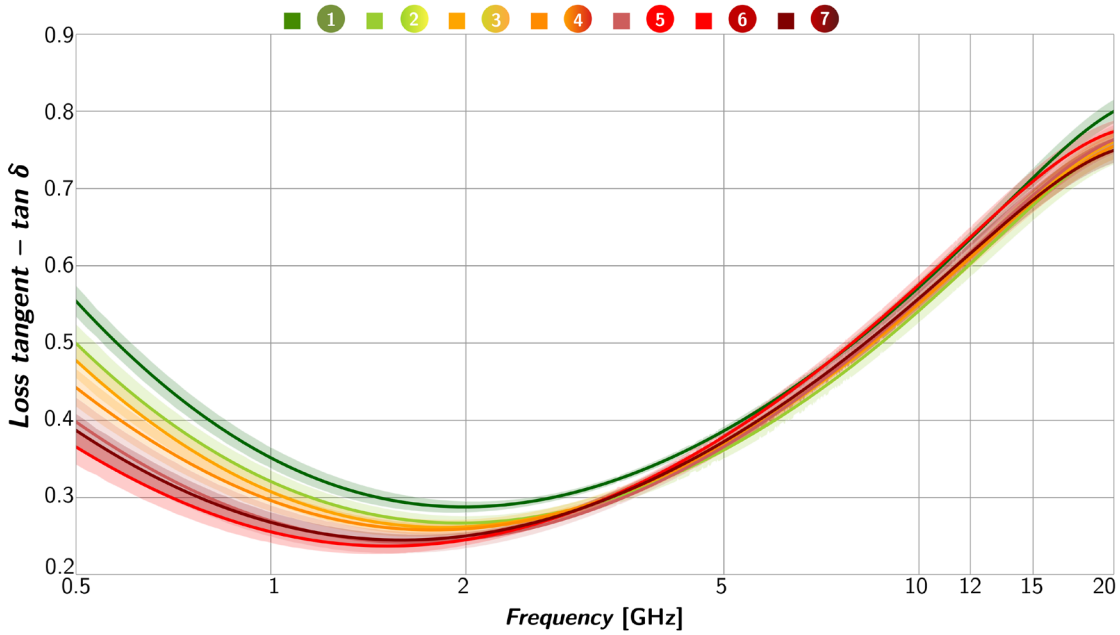

Figure 4: Loss tangent response between 0.5 GHz and 20 GHz for coffee samples at each maturity stage. Solid lines represent mean values at each stage, and shadowed regions represent the standard errors.

3.3 Feature selection and principal component analysis (PCA)

The first cross validation showed that the optimal number of design trees was 300. After randomly running the recursive elimination scheme, 42 out of the 1001 frequencies were included in all models for maturity classification: frequencies between 0.5 GHz to 1.3 GHz. When running the PCA algorithm with these 42 frequencies, 948.5 MHz was the highest PC1 loading frequency. This first component represented almost 98% of the variability of the data. After analysing the remaining frequency ranges: 2 – 10 GHz and 10 GHz to 20 GHz, the highest ranked in these ranges were frequencies between 2 to 3 GHz and those superior to 19 GHz. The interactions between these ranges and 948.5 MHz were included in the final model. A total of 62 features were used for the final principal component analysis. The bi-plot, which presents both the scores and loadings plot is presented in Figure 5.

The first component represents 97% of the variance while the second represents 1.6%. The high variance represented in component 1 shows that the high collinearity between frequencies. Yet, the discrimination capacity of the technique to classify the maturity stages is good although only 6% of the features are retained. The interactions mostly contributed in the direction of component 2. The first component was related to the maturity stage, while the second component presented a slight discrimination across varietals.

The highest dispersion was perceived for stage 1 and stage 2, as with the colorimetric analysis. Positive values of the first component are indicative of green underripe to almost ripe cherries, while negative values indicate ripe to overripe cherries. Samples below 0 for this component are related to good quality attributes (Puerta-Quintero, 2000).

Figure 6 presents the loss tangent values at 948.5 MHz and the groups according to the random effects model. The maturity stages were separated in five different groups at this frequency, with the lowest values for stages 5 to 7 and the higher values for stages 1 to 4, being stage 1 the stage with the highest value. Consequently, Figures 3 to 6 depict the potential that the dielectric spectroscopy technique has for the classification of the maturity stage of coffee cherries.

3.4 Classification models

According to the pipeline structure presented in section 2.6, Table 3 presents the general results for the techniques used and all the dielectric properties.

The best classification results for the dielectric spectroscopy technique were obtained for the SVM (support vector machine) model, that used the first five components of the PCA analysis and used the radial basis kernel with regularization parameter equivalent to 10. Random Forest models present overfitting of the validation sets, and k-NN presented intermediate performance. Consequently, the SVM models were considered and the dielectric spectroscopy response was represented by the loss tangent. Table 4 presents the accuracy obtained for each of the measurements.

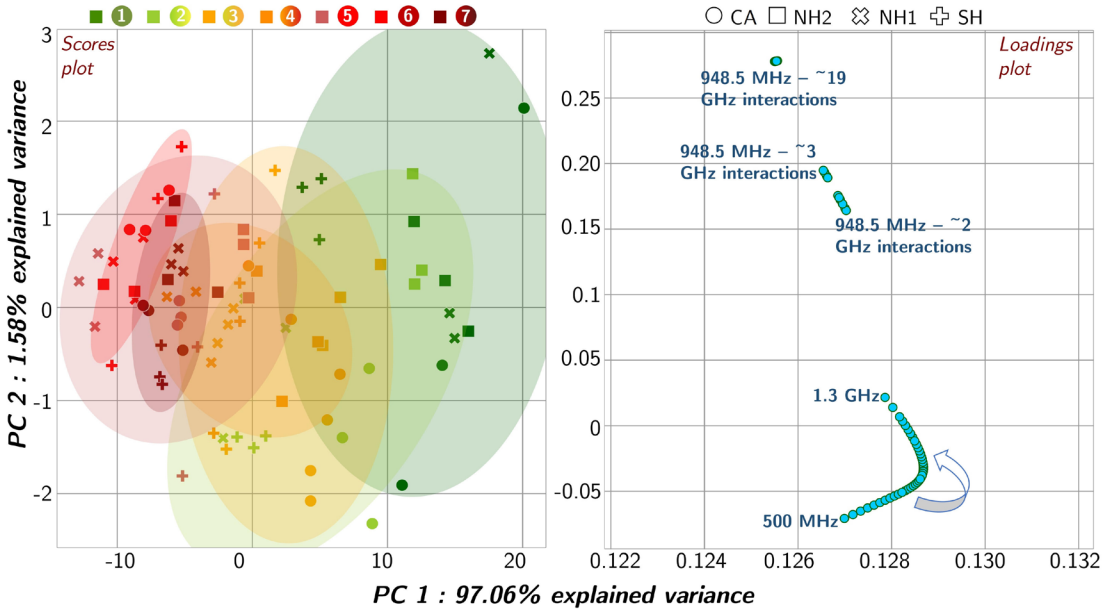

Figure 5: PCA bi-plot for the loss tangent response and the recursive selection features: scores and loadings plot for Northern Huila (NH1 / NH2), Southern Huila (SH1) and Caldas (CA) samples.

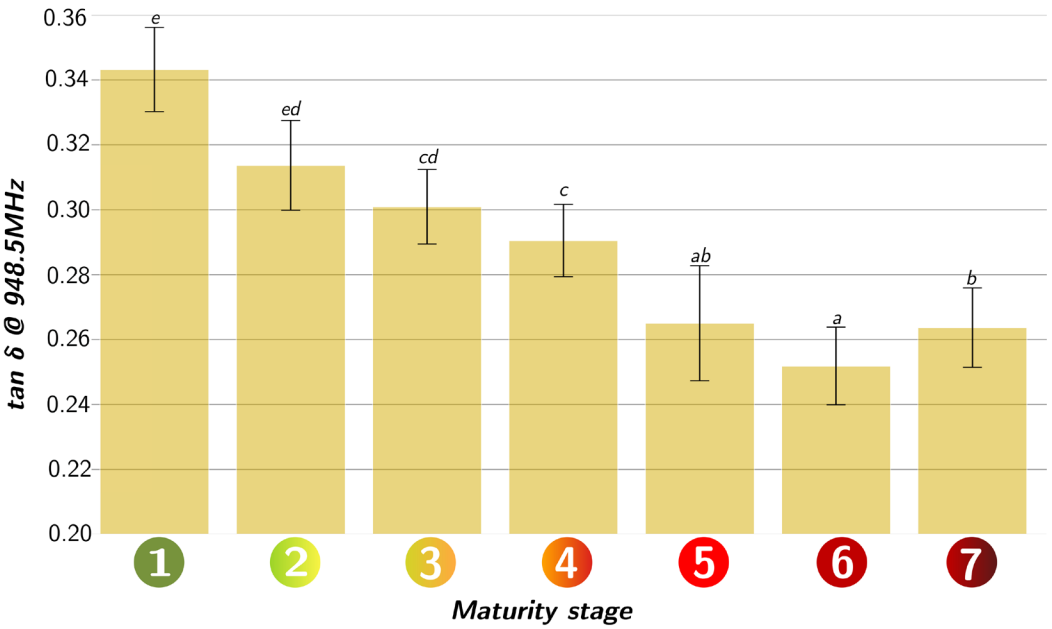

Figure 6: Loss tangent at 948.5 MHz for each maturity stage. Bars represents mean value and error bars represent standard error (p-value: 0.0144.

Table 3: Method and dielectric feature comparison: tree number for Random Forest, regularization parameter (Reg. P) for support vector machine (SVM) and neighbor number for k nearest neighbours (k-NN).

		Random Forest		SVM		k-NN	
		PCA	Tree N.	PCA.	Reg. P	PCA.	Neigh.
		3	175	*5*	*10*	*1*	*4*
tan δ	**Train**	100%		90.7%		95.0%	
	Test	50.7%		56.8%		54.7%	
		PCA	Tree N.	PCA.	Reg. P	PCA.	Neigh.
		2	*200*	*4*	*1000*	*1*	*4*
ε'	**Train**	100%		89.8%		87.0%	
	Test	41.9%		45.1%		41.0%	
		PCA	Tree N.	PCA.	Reg. P	PCA.	Neigh.
		150	*5*	*2*	*1000*	*4*	*2*
ε''	**Train**	100%		48.8%		54.0%	
	Test	21.8%		20.0%		17.3%	

Table 4: Accuracy for the classification of the seven maturity stages: dielectric spectroscopy (tan δ), colorimetric analysis (CIEL*a*b*), moisture content (MC), total soluble solids (TSS) and hardness

Technique	tan d	CIEL*a*b*	MC	TSS	Hardness
Accuracy [%]	56.8	75.1	45.5	48.9	72.5

The highest ranked techniques were the hardness and colorimetric response. For the latter, the result is valid while analysed cultivars are red at ripeness. Yellow, orange or pink varieties might result in a lower global classification accuracy. These two techniques are followed by the loss tangent, which presents an accuracy close to 60 %: The lowest ranked techniques are the total soluble content and the moisture content. However, the normalized confusion matrix is more indicative of the real accuracy at each stage. Figure 7 presents these results, were the columns represent the actual stage, and the rows represent the stage predicted by the model.

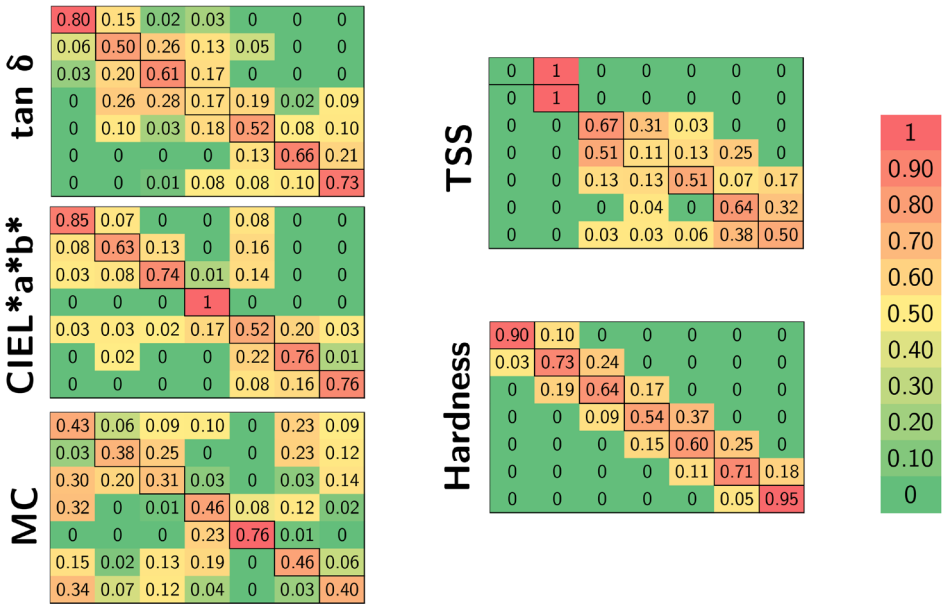

Figure 7: Normalized confusion matrices for the dielectric spectroscopy (tan δ), colorimetric analysis, moisture content (MC), total soluble solids (TSS) and hardness classification models.

The desired response for an ideal model is an identity matrix. False positives are represented in values outside the diagonal of the rows. In the case of the dielectric spectroscopy, the highest true positive values were seen for stages 1, 5 and 6. In particular, the model for stage 1 did not present false positives from stages 5, 6, and 7. Hence, no underripe cherries will be considered as ripe. Stages 2 and 3 present a performance decrease compared to stage 1. The lowest performance samples corresponded to stages 4 and 5. In general, stage 4 is confused with stages 2,3 and 5. In the case of stage 5, samples are mostly confused with stages 4, 5 and 2. It is important to recall that this is independent of the cultivar.

In the case of colour, some of the stage 5 samples were classified as stage 1. The best accuracy was achieved for stage 4. Stage 5 was misclassified at every stage, mostly for stages 4 and 5. Colour is definitely an important variable for maturity assessment, but is important to consider other alternatives (Arcila et al., 2007).

The moisture content is the technique that presented the weakest performance. Yet, the stage that had the best performance was stage 5. Hence, the separation of stage 5 could be achieved by threshold verification. Nevertheless, performing moisture content measurements in the field is not an easy task.

The lack of measurements in stages 1 and 2 might affect the accuracy of the total soluble solid model. The remaining stages present adequate classification values for stages 3, 5 and 6. Stages 6 and 7 present similar performances and are eventually confused. Finally, the hardness presents the best performance. All stages are misclassified with at most 2 of their neighbour stages. Stages 1 and 7 presented the best performance.

4 CONCLUSIONS

Global maturity classifiers for coffee cherries were proposed. The validity of the sample collection was verified with the colorimetric analysis, which presented the highest variance for stages 1 and 2 and a reduction of the variance as the maturity advances. Cultivars with other colours should be considered to evaluate the consistency of the technique.

The physicochemical features associated to the maturity stages agree with previously reported results. Three groups were obtained for all considered technologies due to the most attributes presented a monotonic response. Even though the moisture content presented a maximum at stage 5, threshold strategies could be verified for this approach.

The dielectric spectroscopy technique presents an important potential for this application. Although the dielectric permittivity did not present a difference across the maturity stages, the loss factor and loss tangent did. The difference was mainly evidenced at the ionic conduction region. The loss tangent should be employed as dielectric spectroscopy marker at frequencies from 300 MHz to 1.3 GHz. The higher frequency interactions contributed to the separation across the second PCA component, which was mostly related to the cultivar, than to the first that responded to the maturity stage. This technique has potential for performing singular and mass (i.e., group of cherries) measurements. Furthermore, this could be transferred to simpler devices that can be carried to the field for cherry sorting considering that planar technologies (i.e., antennas and filters) and current software defined platforms operate well at the selected frequencies. This could be further extended to implement RF imaging equipment for online automatic sorting.

From the three machine learning techniques, the best results were obtained for SVM. The training and test accuracy sets agreed well for this technique. As seen in the confusion matrices, the dielectric spectroscopy approach performed well for stages 1, 6 and 7, and reduced performance for stages 3 and 5. According to this, the technique could be used to predict the moisture content.

Although techniques such as the colorimetric analysis and hardness presented reliable performance, the analysis must be performed on a unique cherry basis or require high accuracy that cannot be easily transferred to portable field compatible technology.

5 ACKNOWLEDGMENTS

Physicochemical analyses of coffee and sample collection were performed by members of the Cesurcafé research centre at Universidad Surcolombiana. Dielectric properties measurements were performed by William Romero at Universidad de los Andes.

6 FUNDING

This project was supported by the Colombian Science Council - Colciencias (The science, technology, and development management department) ["Proyecto CAFES", grant number 120471551893].

7 REFERENCES

ADNAN, A. et al. Rapid prediction of moisture content in intact green coffee beans using near infrared spectroscopy. **Foods**, 6(5):38, 2017.

ARCILA, J. et al. **Sistemas de producción de café en Colombia**. Chinchiná, Cenicafé. 2007. 309p.

ARISTIZÁBAL-TORRES, I. D. et al. Physical and mechanical properties correlation of coffee fruit (Coffea Arabica) during its ripening. **Dyna**, 72(172):148-155, 2012.

BASHIR, H. A.; ABU-GOUKH, A. B. A. Compositional changes during guava fruit ripening. **Food Chemistry**, 80(4):557-563, 2003.

BEHERA, S. K.; RATH, A. K.; SETHY, P. K. Maturity status classification of papaya fruits based on machine learning and transfer learning approach. **Information Processing in Agriculture**, 2214-3173, 2020.

BERBERT, P. A. et al. Use of a dielectric function for determination of coffee seeds moisture content. **Bragantia**, 67(2):541-548, 2008.

CASTRO-GIRALDEZ, M. et al. Development of a dielectric spectroscopy technique for the determination of apple (Granny Smith) maturity. **Innovative Food Science & Emerging Technologies**, 11(4):749-754, 2010.

CHANDRASEKAR, V.; VISWANATHAN, R. Physical and thermal properties of coffee. **Journal of Agricultural Engineering Research**, 73(3):227-234, 1999.

CRAIG, A. P. et al. Fourier transform infrared spectroscopy and near infrared spectroscopy for the quantification of defects in roasted coffees. **Talanta**, 134:379-386, 2015.

DE CASTRO, R. D.; MARRACCINI, P. Cytology, biochemistry and molecular changes during coffee fruit development. **Brazilian Journal of Plant Physiology**, 18(1):175-199, 2006.

FARAH, A. Coffee constituents. In: CHU, Y. F. **Coffee:** Emerging health effects and disease prevention. wiley-blackwell, new york, p. 21-58, 2012.

FASHI, M.; NADERLOO, L.; JAVADIKIA, H. The relationship between the appearance of pomegranate fruit and color and size of arils based on image processing. **Postharvest Biology and Technology**, 154:52-57, 2019.

FEDERACIÓN NACIONAL DE CAFETEROS - FNC. **Sustainability that matters 1927-2010, FNC - Management reports**. 2010. Available in: https://www.federaciondecafeteros.org/static/files/informe_sostenibilidad_eng.pdf, Access in: March, 03, 2017.

FRANCO, A. P. et al. Dielectric properties of green coconut water relevant to microwave processing: Effect of temperature and field frequency. **Journal of Food Engineering**, 155:69-78, 2015.

GUO, W. et al. Determination of soluble solids content and firmness of pears during ripening by using dielectric spectroscopy. **Computers and Electronics in Agriculture**, 117:226-233, 2015.

IACCHERI, E. et al. Different analytical approaches for the study of water features in green and roasted coffee beans. **Journal of Food Engineering**, 146:28-35, 2015.

KEYSIGHT TECHNOLOGIES, **Basics of measuring the dielectric properties of materials.** Application note. 2015. Available in: https://www.keysight.com/us/en/assets/7018-01284/application-notes/5989-2589.pdf, Access in: February, 20, 2019.

LI, X. et al. SSC and pH for sweet assessment and maturity classification of harvested cherry fruit based on NIR hyperspectral imaging technology. **Postharvest Biology and Technology**, 143:112-118, 2018.

MARÍN-LÓPEZ, S. M. et al. Cambios físicos y químicos durante la maduración del fruto de Café (Coffea arabica L. var. Colombia). **Cenicafé**, 54(3):208-225, 2003.

MARÍN-LÓPEZ, S.M. et al. Relación entre el estado de madurez del fruto del café y las características de beneficio rendimiento y calidad de la bebida. **Cenicafé**, 54(4):297-315, 2004.

MOMENY, M. et al. Accurate classification of cherry fruit using deep CNN based on hybrid pooling approach. **Postharvest Biology and Technology**, 166:e111204, 2020.

PIEDAD, E. et al. Postharvest classification of banana (Musa acuminata) using tier-based machine learning. **Postharvest Biology and Technology**, 145:93-100, 2018.

PUERTA-QUINTERO, G. I. Influencia de los granos de café cosechados verdes en la calidad física y organoléptica de la bebida. **Cenicafé**, 51(2):136-150, 2000.

RUNGPICHAYAPICHET, P. et al. Robust NIRS models for non-destructive prediction of postharvest fruit ripeness and quality in mango. **Postharvest Biology and Technology**, 111:31-40, 2016.

SANTOS PEREIRA, L. F. et al. Predicting the ripening of papaya fruit with digital imaging and random forests. **Computers and Electronics in Agriculture**, 145:76-82, 2018.

SILVA, S. D. et al. Coffee quality and its relationship with brix degree and colorimetric information of coffee cherries. **Precision Agriculture**, 15:543-554, 2014.

SOSA-MORALES, M. E. et al. Dielectric properties of berries in the microwave range at variable temperature. **Journal of Berry Research**, 7(4):239-247, 2017.

SOSA-MORALES, M. E. et al. Dielectric properties of foods: Reported data in the 21st Century and their potential applications. **LWT - Food Science and Technology**, 43(8):1169-1179, 2010.

SUNARHARUM, W. B.; WILLIAMS, D. J.; SMYTH, H. E. Complexity of coffee flavor: A compositional and sensory perspective. **Food Research International**, 62:315-325, 2014

VELÁSQUEZ, S. et al. Volatile and sensory characterization of roast coffees - Effects of cherry maturity. **Food Chemistry**, 274:137-145, 2019.

YANG, X. et al. Machine learning for cultivar classification of apricots (Prunus armeniaca L.) based on shape features. **Scientia Horticulturae**, 256:e108524, 2019.