Biomatemática ***28*** *(2018), 49–64* ***ISSN 1679-365X***
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Multivariate analysis of trunk texture patterns for supporting tree species identification using computational intelligence

Adriano Bressane[1],
Faculdade de Engenharia de Sorocaba; 18.087-125, Sorocaba/SP.

Felipe H. Fengler[2], Sandra R. M. M. Roveda[3], Jose A. F. Roveda[4],
Antonio C. G. Martins[5],
UNESP–Sorocaba; 18.087-180, Sorocaba/SP.

Abstract. The texture patterns recognition in the tree trunk has been evaluated as an alternative to support species identification. However, the growing demand for extracting more patterns requires an approach able to treat redundant information. The present study aims at evaluating the use of multivariate analysis for improving the performance of trunk texture patterns as tree species indicators. For the experimental procedures, 1188 samples were obtained from 11 arboreal species. By processing on grayscale images, texture patterns were extracted based on first and second order statistics. Then, synthetic variables were obtained by transformations using multivariate analyzes, and used as input in a predictive modeling process. As a result, the multivariate analysis provided an expressive dimensionality reduction, decreasing the number of predictor variables in 85.7%. By optimizing the computational effort, the fall in the error rate achieved 71.4% during the machine learning. Furthermore, a significant increase in the generalization capability was observed during the validation test, achieving 98.6% accuracy. In conclusion, multivariate analysis can be considered a promising approach, but in future studies the use of soft class labels could also be evaluated, to further improving the arboreal identification using computational intelligence.

[1]adriano.bressane@facens.br
[2]felipe.fengler@unesp.br
[3]sandra.regina@unesp.br
[4]jose.roveda@unesp.br
[5]antonio.martins@unesp.br

***Keyword**: biometrics; predictive modeling; computational ecology.*

1. Introduction

The arboreal identification can be difficult and even unfeasible in certain conditions, fostering the development of methods based on computational intelligence, but there are still issues to overcome (Bressane et al., 2015; Yanikoglu et al., 2014; Machado et al., 2013). The current computer-based techniques have focused on leaves features, leading to limitations in cases that those structures are not available. In these cases, the pattern recognition of tree trunk texture could be an alternative, but it is still an ongoing research issue.

The tree trunk features are relatively uniform by species, so that can be useful for a broad identification (Wojtech and Wessels, 2011; Vaucher, 2010). Roughness, thickness, presence of lenticels, aculeus, and stretch marks, among other morphological features, in different directions and denseness, create trunk textures characteristics of each tree species. Nevertheless, taking into account that the trunk texture is a biological feature, its natural variability requires the continuous evaluation of new patterns to overcome the dissimilarity within species, even as the similarity between some of them.

On the other hand, the extraction and inclusion of more patterns also requires an approach able to treat redundant information, owing to the possibility of these new patterns are correlated. Thus, the use multivariate analysis techniques, as the Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), and Exploratory Factor Analysis (EFA), could be experienced.

An important difference among such techniques is that the PCA operates without foreknowledge on class labels (unsupervised). In turn, FDA is a supervised technique in which the class information is considered. Similarly, the EFA also considers the data structure. In spite of this, the performance afforded by a given technique is not necessarily superior to another, being recommended a comparative assessment case-by-case (Martinez and Kak, 2001). In common, such techniques find a coordinate system that maximizes the variance explained in the data, producing synthetic variables by linear combinations of original measured variables. Thus, synthetic variables produced by such techniques could avoid the use of predictors with little explanatory power, allowing the compress information and dimensionality reduction, even as optimizing the computational effort during machine learning (Bro and k. Smilde, 2014; Abdi

and Williams, 2010; Jolliffe, 2002). Hence, the use of the synthetic variables as indicators may provide better results than the original variables.

The present study aims to evaluate the use of multivariate analysis for improving the performance of trunk texture patterns as tree species indicators features, in order to support its identification using computational intelligence.

2. Methods

2.1 Data collection for the experimental analysis

The experimental analysis were performed using outer bark images of 11 deciduous tree species, native from the Brazilian flora. These images were taken at different heights of the trunk, all around the trees, with 50 mm of distance from the digital camera to the target. Then, a central area was cut from each image and, using a moving mask with 512 x 512 pixels, 108 samples per species were thus obtained (see Figure 1).

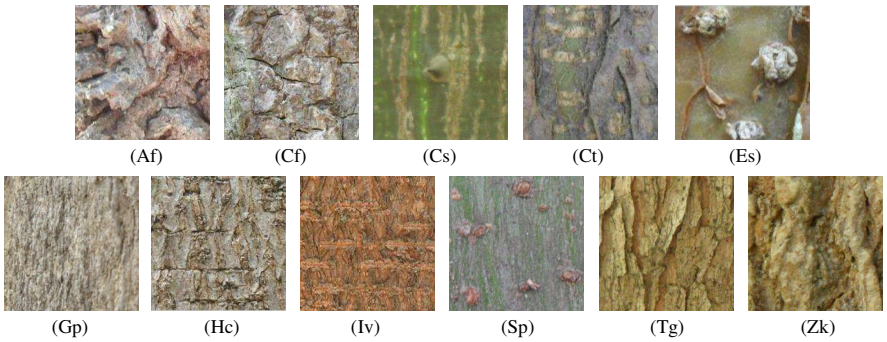

Figure 1: Outer bark images (512 x 512 pixels) of the tree trunk from: *Anadenanthera falcata* (Af), *Cedrela fissilis* (Cf), *Ceiba speciosa* (Cs), *Centrolobium tomentosum* (Ct), *Erythrina speciosa* (Es), *Gochnatia polymorpha* (Gp), *Hymenaea courbaril* (Hc), *Inga vera* (Iv), *Schizolobiun parahyba* (Sp), *Tibouchina granulosa* (Tg), and *Zanthoxylum kleinii* (Zk).

In doing that, 1188 samples were obtained for the experimental analysis, so that 70% were used for the machine learning (training dataset), and 30% during the performance assessment (testing dataset, randomly selected).

2.2 Original variables extraction based on trunk texture

Although some studies have obtained better results in the pattern recognition using color information, it can be more susceptible to variations due to environmental conditions and image acquisition settings. Moreover, from a biological point of view, it's still important to consider that the color features of the same tree may vary depending on the season. Therefore, in order to obtain results for supporting the species identification, the images were transformed from RGB system to HSV space.

Then, using values in the V channel from gray-level images, original variables (z_i) based on first and second order statistics were extracted. The first-order statistics included the uniformity, entropy, skewness, smoothness, intensity, and standard deviation. In turn, the second ones were the contrast, correlation, energy, and homogeneity (Table 1). In the second-order statistics extraction, the values of each of the four parameters were measured at 16 relative positions (θ), equivalents to distance between pixels equal to 1, 3, 5 and 7, in the directions 0, 45, 90 and 135 degrees, so that were generated 64 co-occurrence descriptors. Thus, taking in to account the 6 first-order statistics, the total number of original variables was 70 texture patterns.

2.3 Synthetic variables from multivariate analysis

From the PCA, the synthetic variables ($z_i^{'}$) called principal components (PC) were obtained by uncorrelated linear combinations of the original variables (z_i), i.e, of the texture patterns, and generated in decreasing order of variance, by solving the characteristic equation of the correlation matrix (R), given by Bro and k. Smilde (2014):

$$\det\left(R - \lambda I\right) = 0 \tag{2.1}$$

where λ_i are the eigenvalues, for each of which there is an eigenvector w_i, such that the synthetic variables $z_i^{'}$ are determined as:

$$z_i^{'} = w_{i1}z_1 + w_{i2}z_2 + ... + w_{ip}z_{p'}\,,\ (i = 1, ..., p^{'}) \tag{2.2}$$

where p is the number of original variables.

Table 1: Original variables based on first and second order statistics, considering: grey levels number (L), pixel intensity (ϕ_i), image histogram ($p(\phi_i)$), matrix dimension (δ), relative position (θ), probability of satisfying θ (p_{ij}), mean of rows (m_r) and columns (m_c).

Pattern	Feature	Function
Uniformity	gray levels proximity	$u = \sum_{i=0}^{L-1} p^2(z_i)$
Entropy	image randomness	$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$
Smoothness	gray shades transition	$s = 1 - \frac{1}{1+\mu_2^2}$
Intensity	average gray level	$\mu_1 = \sum_{i=0}^{L-1} z_i p(z_i)$
St. deviation	gray levels dispersion	$\mu_2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \mu_1)^2$
Skewness	asymmetry measure	$\mu_3 = \sum_{i=0}^{L-1} (z_i - \mu_1)^2 p(z_i)$
Contrast	pixels comparison	$c_\phi = \sum_{i=1}^{k} \sum_{j=1}^{k} (i-j)^2 p_{ij}$
Correlation	pixel joint occurrence	$r_\phi = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{(i-m_r)(j-m_c)}{\sigma_r - \sigma_c} p_{ij}$
Energy	local intensity variation	$\varepsilon_\phi = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}^2$
Homogeneity	gray levels closeness	$h_\phi = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{p_{ij}}{1+\lvert i-j \rvert}$

In addition, synthetic variables based on oblique components (OC) was also extracted by means of rotation after PCA, using oblimin method (τ equal to 0) by allowing orthogonal dimensions (when existing) and at the same time does not require independent dimensions. In the FDA the wi is also known as weight vector (or weighting coefficients) of the discriminant functions (DF), similarly considered as synthetic variables ($z_i^{'}$), but with p limited a condition ($p^{'}$) as in (Russell et al., 2000):

$$p^{'} = min(g-1,\ p), \tag{2.3}$$

where g is the number of classes.

As an unsupervised technique, the PCA finds the largest total scatter (S_T) in the data. In turn, the FDA takes into account the data structure, focusing on maximizing between-classes-scatter (S_B), while at the same time the within-classes-scatter (S_W) is minimized (see Figure 2), finding the eigenvector (w) associated with the largest eigenvalue (λ) that maximizes the Fischer's objective function (F), given by:

$$F(w) = w^T S_B w \ (w^T S_w w)^{-1}, \ S_T = S_B + S_w. \tag{2.4}$$

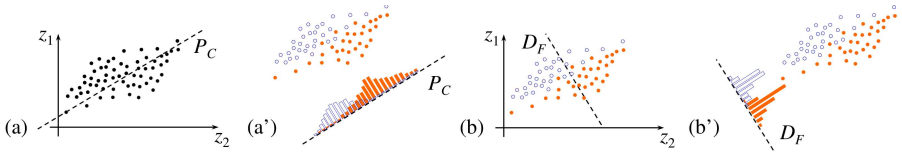

Figure 2: Synthetic variables from linear combination of the original variables (z_1 and z_2), correspondent to principal components - P_C (a) and discriminant functions - D_F (b), even as their directions with the largest total scatter (S_T) projected by PCA (a'), and maximum $F(w)$ given by FDA (b').

Similarly the FDA, the EFA aims to provide a causal modeling considering the data structure. By contrast, whereas the synthetic variables ($z_i^{'}$) produced by PCA and FDA can be considered a composite of the original variables (z_i), in the EFA the opposite occurs (Beavers et al., 2013), as can be seen in Figure 3.

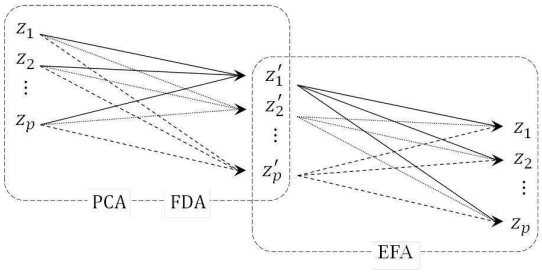

Figure 3: Causal relationships between synthetic variables ($z_i^{'}$) and original ones (z_i) in Principal Component Analysis (PCA), Fischer Discriminant Analysis (FDA), and Exploratory Factor Analysis (EFA).

Statistically, the main difference is that the EFA criterion is based on the communality or common-scatter (S_C), i.e, the variance shared among variables. In the present study, the extraction method was based on principal factors (PF), and the communalities (h_i) were measured by squared multiple correlations, in order to find eigenvector (w) associated with the largest eigenvalue (λ) that maximizes the total communality, given by:

$$\sum_{i=1}^{p} h_i = \sum_{i=1}^{m} \lambda_i, \tag{2.5}$$

where p is the number of original variables, m is the number of synthetic variables, and

$$h_i = \sum_{i=1}^{m} l_{ij}^2, \tag{2.6}$$

where l_{ij} is correlation between the i^{th} principal factor with j^{th} original variable.

Taking into account that such techniques are sensitive to the relative scaling of the original variables, before starting multivariate analysis the data set (x) was standardized, converting all texture patterns to a common scale with an average ($\overline{x}$) of zero and standard deviation (σ) of one, given as in:

$$z = (x - \overline{x})\sigma^{-1}. \tag{2.7}$$

Furthermore, as these multivariate analysis operate over the relationship measures between variables, for non-normal data the synthetic variables are not necessarily statistically independent, i.e., the mutual information is minimized, but some redundancy may remain. Therefore, in the present analysis was used the Spearman's coefficient, a non-parametric surrogate of the Pearson's one, regarded robust for general distributions (distribution-free), and less sensitive to outliers due to inherent variability of the phenomenon. Thus, the multivariate analysis allowed extracting the most important information from the texture patterns (original variables), in order to represent it as synthetic variables ($z_i^{'}$), correspondent to the P_C, D_F and P_F, used as indicators of the tree species in the predictive modeling.

2.4 Predictive modeling and performance assessment

Providing a suitable basis to compare the predictive performance of the features based on original and synthetic variables was mandatory for assessing the prospective improvement by using multivariate analysis. Therefore, the predictive modeling procedure was based on a k-Nearest Neighbor (k-NN) classifier, once it is quite sensitive to features relevance (Lovrek et al., 2008; Ramirez and Puiggros, 2007; Bao et al., 2002). The k-NN is a non-linear and non-parametric supervised machine learning method, requiring for the training process a learning data set (L) composed by pre-classified samples (l_i) in their respective arboreal species (A):

$$L = (l_1, sp(l_1)), ..., (l_N, sp(l_N)), \tag{2.8}$$

where $f(l_i)$ denotes the class (or arboreal species) of the learning sample l_i, so that the $f \in A = (\alpha_1, ..., \alpha_{n_{sp}})$, and n_sp is the total number of tree species.

To determine the tree species of the testing sample in the query point (t_q), the similarity was evaluated considering the k closest points, and the inverse squared distance as weighting factor, so that the nearer neighbors were more influential than the more distant ones. Thereby, t_q correspond to majority class given by:

$$f(t_q) = argmax_{\alpha \in A}(\sum_{i=1}^{k} \delta(\alpha, f(l_i)), \tag{2.9}$$

where $\delta(\alpha, f(l_i))$ is equal to 1 if α correspond to $f(l_i)$, or is equal to 0, otherwise.

As similarity measure between two instances x_i and x_j in the n-dimensional features space (f) was used the Euclidean distance function (d_E), given by:

$$D_E(x_i, x_j) = \sqrt{\sum_{f=1}^{n} (x_i - x_j)^2}. \tag{2.10}$$

A smaller k nearest points may provide a less stable classifier, but a larger k tends to be less precise. Therefore, an error rate (E_{rate}) estimated through v-fold cross-validation (20 folds) was carried out over training data set, in order to identify the best k neighbors and predictor variables amount, even as the number of factors to retain. Then, a hold-out validation using the testing data set also was performed for assessing the generalization ability of synthetic variables as indicators of tree species, even as the prospective improvement in comparison with the use of original variables, according to the metrics of

overall accuracy, precision, sensitivity, specificity, and area under the Receiver Operating Characteristic (ROC) curve.

Considering all species, the overall accuracy (θ) measures the ratio of samples correctly classified by the total number of samples (n_T), given by:

$$\theta = n_T^{-1} \sum_{i=1}^{n_{sp}} TP_{spi}, \tag{2.11}$$

where TP_{spi} is the total number of true positive samples, and n_{sp} is the total number of tree species.

Precision (P) measures the hit rate for each species (spi), take into account the total number of samples identified as belonging to spi (I_{spi}), as in:

$$P(spi) = TP_{spi}.I_{spi}^{-1} = TP_{spi}(TP_{spi} + FP_{spi})^{-1}, \tag{2.12}$$

where FP_{spi} is the total number of false positive samples.

Sensitivity, or true positive rate (tp_{rate}), measures the proportion of positives samples correctly identified as such, taking into account the total number of samples actually belonging to spi (V_{spi}), as in:

$$tp_{rate}(spi) = TP_{spi}.V_{spi}^{-1} = TP_{spi}(TP_{spi} + FN_{spi})^{-1}, \tag{2.13}$$

where FN_{spi} is the total number of false positive samples.

Specificity, or true negative rate (tn_{rate}), measures the proportion of negatives samples correctly identified as such, taking into account the total number of samples actually belonging to others species, as in:

$$tn_{rate}(spi) = 1 - fp_{rate} = TN_{spi}(TN_{spi} + FP_{spi})^{-1}, \tag{2.14}$$

where TN_{spi} is the total number of true positive samples, and fp_{rate} is the false positive rate.

From these metrics, the area under the curve (AUC) based on ROC method (Fawcett, 2005; Landgrebe and Duin, 2007), which provides an integrated measure of true and false positive rates (sensitivity, 1-specificity), was used to further comparative evaluation among predictor variables with the best overall accuracy.

3. Results and discussion

By analyzing the Kaiser-Meyer-Olkin measure that resulted in 0.96, it was noted a good sampling adequacy. Moreover, the Cronbach's alpha equal

to 0.90 indicated reliability by the method of internal consistency. In turn, the Bartlett's test for eigenvalue significance, which p-value less than 0.001, confirmed that the correlation between variables is sufficient to perform the multivariate analysis. As a result from the PCA, FDA, and EFA the eigenvalues, cumulative variability explained by synthetic variables, and its respective projections based on the three first dimensions are shown in Figure 4.

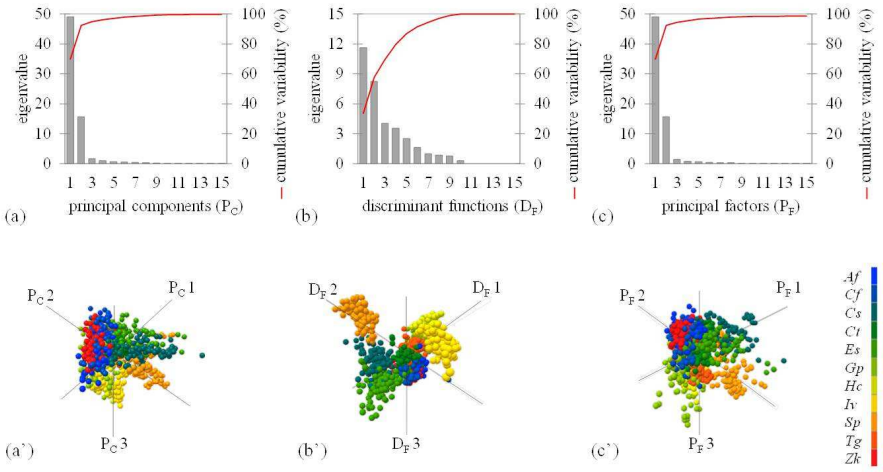

Figure 4: Cumulative variability explained by synthetic variables produced by Principal Component Analysis (a), Fischer Discriminant Analysis (b), and Exploratory Factor Analysis (c), even as the respective projections from the three first principal components (a'), discriminant functions (b'), and principal factors (c').

From the scree plots in Figure 4, it can be seen that the variables synthesized by PCA and EFA had quite similar eigenvalues and cumulative variability, respectively equal to 94.7% and 94.4% explained by three first dimensions. Nevertheless, based on distinct variance criteria, total (PCA) and common (EFA), these techniques projected different coordinate systems, which particular effects on its capability to separate the tree species samples. In contrast, the variability explained by the same dimensions projected by FDA accumulated only 69.3% of the variance in the data. In spite of this, taking into account only these three first dimensions, the feature space afforded by FDA seems to achieve the best outcomes. However, it is also need to consider the other

dimensions, in order to further the performance comparison.

Based on the best results of v-fold cross validation during the training process, the number of variables used as predictor in the k-NN classifier was 26 principal components (P_C) and 23 oblique components (O_C) from PCA, 10 discriminant functions (D_F) from FDA, and 29 principal factors (P_F) from EFA, in each case with cumulative variability equivalent to about 99.9%. Thus, the performance results of the evaluated alternatives are presented in Table 2.

Table 2: Performance from k-Nearest Neighbor (k-NN) based on original variables (z_i) and synthesized ones by principal components analysis (P_C), PCA-based oblique rotation (O_C), Fischer discriminant analysis (D_F), and Exploratory Factor Analysis (P_F).

	Training error				Testing accuracy (%)			
	1-NN	3-NN	5-NN	7-NN	1-NN	3-NN	5-NN	7-NN
70 z_i	0.19	0.23	0.25	0.28	91.8	90.1	90.1	89.2
26 P_C	0.20	0.23	0.25	0.29	91.8	89.9	89.5	89.2
23 O_C	0.10	0.15	0.19	0.23	98.0	96.6	94.5	95.5
10 D_F	0.07	0.07	0.08	0.08	98.3	96.9	96.6	95.5
29 P_F	0.09	0.15	0.18	0.22	98.6	96.6	95.5	94.4

As a reference for evaluating the performance improvement afforded by multivariate analysis, it can be seen in Table 2 that original variables had an error rate of 0.19 during the training (1-NN), achieving an overall accuracy of 91.8% based on hold-out validation with testing data set, decreasing to 90.1% and 89.2% for more stable settings, with 5 and 7-NN, respectively. These results can be considered a good performance by combining first and second order statistics as predictor variables. Notwithstanding, outcomes achieved by synthetic variables from multivariate analysis were even better.

Analyzing Table 2, it is noted that the principal components have not improved the initial performance, achieved by original variables. On the other hand, the performance (error and accuracy) was practically the same, but with a significant dimensionality reduction (- 62.9%), decreasing the number of predictors from 70 to 26 variables.

In turn, the oblique components, obtained by rotation from PCA, increased the accuracy in up to 7.2% (with 3-NN) over the original performance. Moreover, the error rate decrease has achieved 47.4% (for 1-NN), using an even smaller number of predictors (23 variables). The rotations are often used to

retrieve as far as possible the meaning of the variables, aiming to enhance their interpretative. Nevertheless, from such results, it is noted that the PCA-rotated data can also provide a better performance in classification tasks.

Table 3: Performance metrics afforded by the predicting models with the best overall accuracy based on 3-NN classifier, according to: precision (P), sensitivity (tp_{rate}), specificity (tn_{rate}), and area under the curve (AUC).

		Arboreal species and performance										
		A_f	G_p	C_f	S_p	C_s	H_c	I_v	E_s	C_t	T_g	Z_k
O_C	P	1	1	.94	.97	1	1	1	1	.86	.94	.94
	tp_{rate}	1	.94	.91	1	.91	1	1	.94	1	.97	.97
	tn_{rate}	1	1	.99	1	1	1	1	1	.98	.99	.99
	AUC	1	.97	.95	1	.96	1	1	.97	.99	.98	.98
D_F	P	1	1	.91	1	1	1	1	.94	.91	.97	.94
	tp_{rate}	.97	.94	.97	1	.97	1	1	.91	1	.94	.97
	tn_{rate}	1	1	.94	1	1	1	1	.99	1	1	.99
	AUC	.99	.97	.96	1	.98	1	1	.99	1	.97	.98
P_F	P	1	1	.93	1	1	1	1	1	.90	.91	.89
	tp_{rate}	1	.97	.84	1	1	1	1	.94	.88	1	1
	tn_{rate}	1	1	.99	1	1	1	1	1	.99	.99	.98
	AUC	1	.99	.92	1	1	1	1	.97	.94	1	.99

In general, the best performances were obtained by variables synthesized from FDA and EFA. The FDA provided the most expressive dimensionality reduction (-85.7%), decreasing from 70 to only 10 predictors. Hence, optimizing the computational effort during the machine learning, the FDA had the lower error rate, mainly for a larger k nearest neighbors.

In this sense, the reduction provided by FDA in the and 63.6% over EFA in the most stable setting (7-NN). In turn, the EFA provided the best accuracy (98.6%) among all evaluated settings and techniques, equivalent to an increasing of 7.4% over original variables performance.

On the order hand, in more stable settings (three or more nearest neighbors) the performance provided by FDA outperforms the EFA. Taking into account that the 1-NN classifier can be less stable, i.e., more sensitive to different dataset of learning and testing by considering less information, the predictors variables with best overall accuracies (O_C, D_F and P_F) were compared using the 3-NN results, according to the performance metrics presented in Table 3.

Analyzing Table 3 it is possible to calculate that the average precision achieved by the discriminant functions (97.0%) was slightly better than one provided from principal factors (96.7%) and oblique components (96.8%). In this sense, the D_F was the only one which provided precision superior than 91% for all species, while the O_C achieved 86.5% for the *Centrolobium tomentosum* (Ct), and the P_F got 88.9% to *Zanthoxylum kleinii* (Zk).

The same superiority was observed in relation to average sensitivity (tp_{rate}), equal to 96.9% for D_F, against 96.6% for both O_C and P_F. These results indicate that D_F had larger generalization capability to classify samples truly belonging to each species, making less omission errors. The biggest omission errors were made by PF, the only predictors set which had sensitivity lower than 88%, such as 84.4% for *Cedrela fissilis* (Cf).

On the other hand, the predictors based on the DF had the lowest average specificity, achieving 99.2%, while the O_C and P_F obtained 99.6%. Hence, the D_F caused the largest commission errors, but even so in the worst case the specificity was 93.8% for the *Gochnatia polymorpha* (Gp), which can be considered a very high refusal rate when the sample really does not belong to tree species.

By evaluating the area under the ROC curve (AUC), it is noted that the PF achieved a perfect performance (100%) for five tree species, while the O_C and D_F for only four ones. Nevertheless, the PF had also the lowest AUC (91.5%), associated with *Cedrela fissilis* (Cf). As a consequence of this balance among advantages in one or another aspect, all three predictor sets obtained the same average AUC (98.1%). Thus, based on an integrated analysis of the commission and omission errors, it is reasonable to consider that these three alternatives (O_C, D_F and P_F) achieved a quite similar ability in supporting the tree species identification.

4. Conclusion

By reviewing previous studies it was noted that the use of multivariate analysis represent a lack in the study of the trunk texture as indicator of the tree species. Then, the use of variables synthesized from multivariate analysis was compared to the performance of original variables based on trunk texture patterns, in order to support the arboreal identification using computational intelligence.

Regarding to the compress information, all assessed multivariate techniques provided expressive dimensionality reduction, achieving up to 85.7% of decrease in the number of predictor variables. Thus, by optimizing the computational effort, there was a fall in the error rate that achieved 71.4% during machine learning. As an expressive result, the best accuracy (98.6%) represented an increasing of 7.4% over the generalization capability of the original variables, during the validation test.

In conclusion, the use of variables synthesized from multivariate analysis can be considered a promising strategy. Nevertheless, a progressive inclusion of more tree species tends to make its identification more difficult. Therefore, in future studies an approach able to deal with a more expressive overlapping of feature values could be experienced, such as the use of patterns with soft boundaries, aiming at further improving the performance of the computer-aided tree identification.

Acknowledgments

Supported by the Coordination for the Improvement of Higher Education Personnel (CAPES).

References

Abdi, A. and Williams, L. J. (2010). Principal component analysis. *Computational statistics*, 2(4):433–459.

Bao, Y., Du, X., and Ishi, N. (2002). *Combining feature selection with feature weighting for k-NN classifier. A machine learning approach to detecting instantaneous cognitive states*, chapter 67, pages 461–468. Lecture Notes in Computer Science. Springer, Manchester.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., W., H. S., Skolits, G. J., and Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical assessment, research & evaluation*, 18(6):1–13.

Bressane, A., Roveda, J. A. F., and Martins, A. C. G. (2015). Statistical analysis of texture in trunk images for biometric identification of tree species. *Environmental monitoring and assessment*, 187:1–9.

Bro, R. and k. Smilde, A. (2014). Principal component analysis. *Analytical methods*, 6(9):2812–2831.

Fawcett, L. T. (2005). An introduction to roc analysis. *Pattern Recognition Letters*, 1:861–874.

Jolliffe, L. T. (2002). *Principal Component Analysis*. Springer.

Landgrebe, C. W. T. and Duin, R. P. W. (2007). Approximating the multiclass roc by pairwise analysis. *Pattern Recognition Letters*, 28:1747–1758.

Lovrek, I., Howlett, R. J., and Jain, L. C. (2008). *Knowledge-based intelligent information and engineering systems*. Springer.

Machado, B. B., Casanova, D., Gonalves, W. N., and Bruno, O. M. (2013). Partial differential equations and fractal analysis to plant leaf identification. *Journal of physics*, 410:1–4.

Martinez, A. M. and Kak, A. C. (2001). Pca versus lda. *Transactions on pattern analysis and machine intelligence*, 23(2):228–233.

Ramirez, R. and Puiggros, M. (2007). *A machine learning approach to detecting instantaneous cognitive states*, chapter 25, pages 248–259. Advances in knowledge discovery and data mining. Springer, Nanjiing.

Vaucher, H. (2010). *Tree bark: a color guide*. Timer press.

Wojtech, M. and Wessels, T. (2011). *Bark: a field guide to trees of the northeast*. UPNE.

Yanikoglu, B., Aptoulak, E., and Tirkaz, C. (2014). Automatic plant identification from photographs. *Machine vision and applications*, 25(6):1369–1383.