UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Sorocaba

PROGRAMA DE PÓS-GRADUAÇÃO em
Ciências Ambientais

**ADRIANO BRESSANE**

# IDENTIFICAÇÃO DE ESPÉCIES ARBÓREAS APOIADA POR RECONHECIMENTO DE PADRÕES DE TEXTURA NO TRONCO USANDO INTELIGÊNCIA COMPUTACIONAL

Sorocaba
2017

**ADRIANO BRESSANE**

# IDENTIFICAÇÃO DE ESPÉCIES ARBÓREAS APOIADA POR RECONHECIMENTO DE PADRÕES DE TEXTURA NO TRONCO USANDO INTELIGÊNCIA COMPUTACIONAL

Tese apresentada como requisito para a obtenção do título de Doutor em Ciências Ambientais da Universidade Estadual Paulista "Júlio de Mesquita Filho" na Área de Concentração Diagnóstico, Tratamento e Recuperação Ambiental

Orientador:

Prof. Dr. José Arnaldo Frutuoso Roveda

Coorientador:

Prof. Dr. Antonio Cesar Germano Martins



Sorocaba

2017

Ficha catalográfica elaborada pela Biblioteca da Unesp
Instituto de Ciência e Tecnologia – Câmpus de Sorocaba

Bressane, Adriano.
Identificação de espécies arbóreas apoiada por reconhecimento de padrões de textura no tronco usando inteligência computacional / Adriano Bressane, 2017.
112 f.: il.

Orientador: José Arnaldo Frutuoso Roveda
Coorientador: Antonio Cesar Germano Martins

Tese (Doutorado) – Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Ciência e Tecnologia (Câmpus de Sorocaba), 2017.

1. Bioinformática. 2. Processamento de imagens. 3. Aprendizado do computador. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Ciência e Tecnologia (Câmpus de Sorocaba). II. Título.

UNIVERSIDADE ESTADUAL PAULISTA

**Câmpus de Sorocaba**

## CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Identificação de espécies arbóreas apoiada por reconhecimento de padrões de textura no tronco usando inteligência computacional

**AUTOR: ADRIANO BRESSANE**
**ORIENTADOR: JOSE ARNALDO FRUTUOSO ROVEDA**
**COORIENTADOR: ANTONIO CESAR GERMANO MARTINS**

Aprovado como parte das exigências para obtenção do Título de Doutor em CIÊNCIAS AMBIENTAIS, área: Diagnóstico, Tratamento e Recuperação Ambiental pela Comissão Examinadora:

Prof. Dr. JOSE ARNALDO FRUTUOSO ROVEDA
Engenharia Ambiental / UNESP - Campus de Sorocaba / UNESP-Sorocaba

Prof. Dr. ADMILSON IRIO RIBEIRO
Engenharia Ambiental - ICTS/ UNESP / UNESP-Sorocaba

Prof. Dr. GERSON ARAUJO DE MEDEIROS
Engenharia Ambiental - ICTS/ UNESP / UNESP Sorocaba

Profa. Dra. NELI REGINA SIQUEIRA ORTEGA
Consultor Independente / Consultor Indepedente

Prof. Dr. MARCOS EDUARDO RIBEIRO DO VALLE MESQUITA
Matemática / Departamento de Matemática Aplicada- IMECC/ Unicamp

Sorocaba, 31 de março de 2017

## AGRADECIMENTOS

Há muitas pessoas a quem devo agradecimento e dedico essa conquista que, na realidade, é de todos nós, familiares, amigos e professores.

À *Universidade Estadual Paulista* e ao *Programa de Pós-Graduação em Ciências Ambientais* do Instituto de Ciência e Tecnologia, campus de Sorocaba.

À *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*, pelo amparo financeiro à pesquisa durante o desenvolvimento da tese.

Aos amigos e familiares, em especial aos meus pais *Reginaldo Bressane* e *Lucelena da Cruz Bressane*, pessoas de valor, exemplos de caráter, superação e bondade.

Aos professores, em particular ao *José Arnaldo F. Roveda* e a *Sandra R. M. M. Roveda*, por todo ensinamento, mas principalmente pelo apoio e amizade.

Agradeço a todos sem exceção, mas dedico especialmente à pessoa que dá sentido a minha vida, ***Patricia Satie Mochizuki***, por tudo que palavras não seriam capazes de expressar.

Bressane A. **Identificação de espécies arbóreas apoiada por reconhecimento de padrões de textura no tronco usando inteligência computacional**. 2017. 112f. Tese (Doutorado em Ciências Ambientais) - Campus de Sorocaba, UNESP - Univ Estadual Paulista, Sorocaba, 2017.

## RESUMO

Embora fundamental para diversas finalidades, a identificação de espécies arbóreas pode ser complexa e até mesmo inviável em determinadas condições, motivando o desenvolvimento de métodos assistidos por inteligência computacional. Nesse sentido, estudos têm se concentrado na avaliação de características extraídas a partir de imagens da folha e, apesar dos avanços, não são aplicáveis a espécies caducifólias em determinadas épocas do ano. Logo, o uso de características baseadas na textura em imagens do tronco poderia ser uma alternativa, mas ainda há poucos resultados reportados na literatura. Portanto, a partir da revisão de trabalhos anteriores, foram realizados experimentos para avaliar o uso de métodos de inteligência computacional no reconhecimento de padrões de textura em imagens do tronco arbóreo. Para tanto, foram consideradas espécies arbóreas caducifólias nativas da flora brasileira. As primeiras análises experimentais focaram na avaliação de padrões. Como resultado, verificou-se que a melhor capacidade de generalização é alcançada combinando o uso de estatísticas de primeira e segunda ordem. Contudo, o aumento de variáveis preditoras demandou uma abordagem capaz de lidar com informação redundante. Entre as técnicas avaliadas para essa finalidade, a análise fatorial exploratória proporcionou redução na taxa de erros durante o aprendizado de máquina e aumento da acurácia durante a validação com dados de teste. Por fim, constatando que a variabilidade natural da textura no tronco arbóreo causa uma ambiguidade no reconhecimento de padrões, o uso da modelagem fuzzy foi avaliado. Em comparação com outros algoritmos de aprendizagem de máquina, a abordagem fuzzy proporcionou resultados competitivos e, assim, pode ser considerada uma alternativa promissora para novos avanços no apoio a identificação de espécies arbóreas usando inteligência computacional.

Palavras-chave: bioinformática, processamento de imagens, aprendizagem de máquina.

Bressane A. **Arboreal species identification supported by texture pattern recognition in trunk using computational intelligence**. 2017. 112f. Thesis (Doctoral's degree in Environmental Sciences) - Campus de Sorocaba, UNESP - Univ Estadual Paulista, Sorocaba, 2017.

## ABSTRACT

Although the arboreal identification is mandatory for several purposes, it can be complex and infeasible under certain conditions, motivating the development of computer-aided methods. In this sense, studies have focused on the assessment of features extracted from leaf images and, despite advancements, they are not applicable for deciduous species in some periods of year. Therefore, the usage of features based on texture in trunk images could be an alternative, but there are still few outcomes reported in the literature. Thus, from the review on previous studies, experiments have been performed for evaluating the use of computational intelligence methods for texture patterns recognition in trunk images. For that, native species from the deciduous Brazilian forest were considered. Firstly, the experimental analyzes focused on the evaluation of patterns. As a result, it was noted that the best generalization ability is reached using the first-order statistics in combination with second-order descriptors. Nevertheless, the increase of predictor variables required an approach capable of dealing with redundant information. Among the techniques assessed for this purpose, the exploratory factor analysis provided an error rate reduction during the machine learning, and an accuracy improvement in the validation over testing dataset. Finally, taking into account that the natural variability of texture in arboreal trunk causes an ambiguity in the pattern recognition, the usage of fuzzy modeling has been evaluated. In comparison with other machine learning algorithms, the fuzzy approach afforded competitive results, and hence it can be a promising alternative for further progress in the arboreal identification supported by computational intelligence.

Key words: bioinformatics, image processing, machine learning.

## LISTA DE FIGURAS

INTRODUÇÃO

Figura 1. Estrutura organizacional na composição da tese ..... 18

CAPÍTULO 1

Figure 1. Overview of the approach based on computational intelligence for supporting the arboreal species identification ..... 22

Figure 2. Bark features with influence on texture in arboreal trunk images: (a) smooth, (b) striated, (c) fissured, (d) cancerous, (e) with protrusions, (f) with lenticels, (g) spines or aculeus, (h) powdery, detaching themselves (i) as fine pieces, (j) as coriaceous pieces, and (k) thick plaques ..... 28

CAPÍTULO 2

Figure 1. Location of Biquinha Municipal Natural Park, in the city of Sorocaba, São Paulo, Brazil. ..... 38

Figure 2. Trunk images with 512 x 512 pixels from: (a) *Chorisia speciosa*, (b) *Schizolobiun parahyba*, (c) *Gochnatia polymorpha*, (d) *Cedrela fissilis*, (e) *Anadenanthera falcata* ..... 38

Figure 3. Histogram for uniformity values from the trunk images ..... 42

Figure 4. Histogram for smoothness values from the trunk images ..... 43

Figure 5. Histogram for third moment (asymmetry) values from the trunk images..... 43

Figure 6. Histogram for entropy values from the trunk images ..... 44

Figure 7. Representation of the identification system for *Anadenanthera falcata* (*Af*); *Cedrela fissilis* (*Cf*); *Gochnatia polymorpha* (*Gp*); *Schizolobiun parahyba* (*Sp*); *Chorisia speciosa* (*Cs*), with threshold selection for: *U - Uniformity*; *e - Entropy*; *R - Smoothness*; and $\mu_3$ – *Asymmetry* ..... 45

CAPÍTULO 3

Figure 1. Outer bark images of the tree trunk from: (a) *Chorisia speciosa*, (b) *Schizolobiun parahyba*, (c) *Gochnatia polymorpha*, (d) *Cedrela fissilis*, (e) *Anadenanthera falcata*, (f) *Hymenaea courbaril*, and (g) *Inga vera*. ..... 54

Figure 2. Decision tree built for species *Anadenanthera falcata* (*Af*), *Cedrela fissilis* (*Cf*), *Chorisia speciosa* (*Cs*), *Gochnatia polymorpha* (*Gp*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), and *Schizolobiun parahyba* (*Sp*), based on both first-order statistics and co-occurrence descriptors ($DT_{S+C}$). ..... 60

Figure 3. Area under the ROC curve for the decision trees (DT) based on statistical parameters (S), co-occurrence descriptors (G) and both (S+G), in supporting the identification

of species: *Anadenanthera falcata* (*Af*), *Cedrela fissilis* (*Cf*), *Chorisia speciosa* (*Cs*), *Gochnatia polymorpha* (*Gp*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), and *Schizolobiun parahyba* (*Sp*) ..... 63
CAPÍTULO 4
Figure 1. Outer bark images (512 x 512 pixels) of the tree trunk from: (a) *Anadenanthera falcata,* (b) *Cedrela fissilis,* (c) *Ceiba speciosa,* (d) *Centrolobium tomentosum*, (e) *Erythrina speciosa*, (f) *Gochnatia polymorpha,* (g) *Hymenaea courbaril,* (h) *Inga vera*, (i) *Schizolobiun parahyba*, (j) *Tibouchina granulosa*, and (k) *Zanthoxylum kleinii* (*Zk*)..... 68
Figure 2. Synthetic variables from linear combination of the original variavles ($z_1$ 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 $F(\boldsymbol{w})$ given by FDA (b') ..... 71
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)..... 71
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')..... 74
CAPÍTULO 5
Figure 1. Tree trunk images (512x512 pixels) from: *Anadenanthera falcata* (*Af*), *Anadenanthera macrocarpa* (*Am*), *Bauhinia forficate* (*Bf*), *Caesalpinia peltophoroides* (*Ca*), *Caesalpinia echinata* (*Ce*), *Cedrela fissilis* (*Cf*), *Caesalpinia peltophoroides* (*Cp*), *Ceiba speciosa* (*Cs*), *Centrolobium tomentosum* (*Ct*), *Enterolobium contortisiliquum* (*Ec*), *Erythrina speciosa* (*Es*), *Gochnatia polymorpha* (*Gp*), *Guazuma ulmifolia* (*Gu*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), *Piptadenia gonoacantha* (*Pg*), *Schizolobiun parahyba* (*Sp*), *Tibouchina granulosa* (*Tg*), *Tabebuia roseoalba* (*Tr*), and *Zanthoxylum kleinii* (*Zk*).....87
Figure 2. Grid-type fuzzy partition: (a) partitioning of the predictor variable - $x_i$ into regions correspondent to the antecedents terms - $a_{ij}$ using trapezoidal-shaped membership functions; (b) intervals of certainty and uncertainty that comprises the fuzzy region of the antecedent term. ..... 91
Figure 3. Split of database for the learning process and to assessing the generalization ability based on testing dataset. ..... 92
Figure 4. Eigenvalues and cumulative variability explained by the first 20 latent variables

(principal factors) produced from the Exploratory Factor Analysis..... 94
Figure 5. Performance of different setings of the fuzzy rule-based classification model, from the variations of antecedent terms number in combination with minimum and product t-norm ..... 97
Figure 6. Aggregation process of predictor variables ($x_i$) in the rules 1 ($R_1$) and 2 ($R_2$), using minimum and product t-norm..... 98
Figure 7. Increase in decision areas formed by the fuzzy if-then rules as consequence of the increment of the antecedent terms numbers. ..... 99

## LISTA DE TABELAS

CAPÍTULO 2

Table 1. Correct hit estimates for the classification system based on: $S_j$- System output; $S_D^j$- Dominant output; $p_A^{S_j}$- Sample coefficient for species; $P^{S_j}$- Probability for a sample belongs to species; $H_{rate}$ - Hit rate for each species; $\overline{H}_{rate}$ - Average hit rate..... 47
Table 2. Identified class as the dominant output ..... 47
Table 3. Confusion matrix for the testing image classification outcomes, with measures: $V_{sp_i}$- Total number of samples actually belonging to species; $I_{sp_i}$ - Total number of samples identified as belonging to species; $T_{sp_i}$- Ratio of correctly classified samples; $F_{sp_i}$- Ratio of samples miss classified..... 47
Table 4. Performance assessment for the identification system, using: $P_{rate}$- Precision rate; $E_{rate}$ - Error rate; $S_{rate}$ - Sensitivity or hit rate; $\theta_1$ - Accuracy or rate of overall accuracy; $K$- Kappa or agreement index..... 48

CAPÍTULO 3

Table 1. Performance from decision trees (DT) based on: first-order statistics (S), co-occurrence descriptors (C), and both (S+C) ..... 59
Table 2. Texture pattern importance as indicator in the $DT_{S+C}$ ..... 61
Table 3. Confusion matrix for the classification results based on testing dataset achieved by $DT_{S+C}$ ..... 62
Table 4. Performance metrics based on testing dataset achieved by $DT_{S+C}$..... 62

CAPÍTULO 4

Table 1. Original variables based on first and second order statistics, considering: grey levels number ($L$), pixel intensity ($\varphi_i$), image histogram ($p(\varphi_i)$), matrix dimension ($\delta$), relative

position ($\emptyset$), probability of satisfying $\emptyset$ ($p_{ij}$), mean of rows ($m_r$) and columns ($m_c$)..... 69
Table 2. Performance 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 ($D_F$)..... 75
Table 3. Performance metrics afforded by the predicting models with the best overall accuracies based on 3-NN classifier, according to: precision ($P$), sensitivity ($tp_{rate}$), specificity ($tn_{rate}$), and area under the curve (AUC) ..... 77
CAPÍTULO 5
Table 1. Machine learning algorithms considered for performance comparison and control parameters settings adjusted during the learning process, which provide the best results in the cross-validation over the checking dataset..... 93
Table 2. Performance of the machine learning algorithms in the benchmarking experiments, based on the settings that reach the best accuracy over checking dataset during the learning process, using the first 20 principal factors as predictor variables..... 94

## LISTA DE SIGLAS E ABREVIATURAS

ACH - angle code histogram
AUC - area under the curve
ANN - artificial neural network
ACM - auto-correlation method
BPN - back propagation neural network
BDT - binary decision tree
C5- boosted rule-based model
CDA - canonical discriminant analysis
CNN - cascade-correlation neural network
CNN - cellular neural network
CCD - centroid-contour distance
CR - centroid-radii
CI - computational intelligence
CM - contour moment
COMM - cooccurrence matrices method
CSS - curvature scale space
DT - decision trees

EFA - exploratory factor analysis

FDA - Fisher discriminant analysis

FMT - Fourier moment technique

FRBCS - fuzzy rule-based classification system

GF - geometrical features

GLCM - grey level co-occurrence matrix

HM - Hu moment invariants

HSV - hue-saturation-value

IVM - import vector machine

SIFT - invariant feature transform

KMO - Kaiser-Meyer-Olkin

*k*-NN - *k*-nearest neighbor

LSH - level-saturation-hue

LDA - linear discriminant analysis

MDP - modified dynamic programming

MFD - modified Fourier descriptor

MMC - move median centers

MCH - moving center hypersphere

MLP - multi-layer perceptron network

MWM - multi-resolution wavelet method

PDE - partial differential equations

PFT - polar Fourier transform

PCA - principal component analysis

PNN - probabilistic neural network

RBF - radial basis function neural network

ROC - receiver operating characteristic

RGB - red-green-blue

ROI - regions of interest

SDT - single decision tree

SVM - support vector machine

VFD - volumetric fractal dimension

**SUMÁRIO**

**INTRODUÇÃO** ..... 16

**1 Objetivos** ..... 17

1.1 Geral ..... 17

1.2 Específicos ..... 17

**2 Estrutura da tese** ..... 18

**CAPÍTULO 1**

**ARBOREAL SPECIES IDENTIFICATION USING COMPUTATIONAL INTELLIGENCE** ..... 20

**Abstract** ..... 20

**1 Introduction** ..... 21

**2 Morphological based identification and computer-aided approach** ..... 21

**3 Advances and limits of using leaf-based approach** ..... 23

**4 Texture patterns recognition from the arboreal trunk images** ..... 27

**5 Conclusion** ..... 29

**References** ..... 30

**CAPÍTULO 2**

**STATISTICAL ANALYSIS OF TEXTURE IN TRUNK IMAGES FOR BIOMETRIC IDENTIFICATION OF TREE SPECIES** ..... 36

**Abstract** ..... 36

**1 Introduction** ..... 37

**2 Materials and methods** ..... 37

2.1 Construction of the system ..... 39

2.2 Classification performance of the constructed system ..... 40

**3 Results and discussion** ..... 42

3.1 Statistical properties of texture from trunk images ..... 42

3.2 Construction of the classification system ..... 44

3.3 Validation of the classification system ..... 46
**4 Conclusions** ..... 49
**References** ..... 49
**CAPÍTULO 3**
**CO-OCCURRENCE PATTERNS ANALYSIS ON THE TRUNK TEXTURE AS INDICATOR FEATURES FOR COMPUTER–AIDED TREE IDENTIFICATION** ..... 52
**Abstract** ..... 52
**1 Introduction** ..... 53
**2 Methods** ..... 54
2.1 Data sampling and collection ..... 54
2.2 Bark texture patterns in tree trunk images ..... 55
2.3 Predictive modeling procedure ..... 57
2.4 Recognition performance assessment ..... 57
**3 Results and discussion** ..... 59
**4 Conclusions** ..... 64
**References** ..... 64
**CAPÍTULO 4**
**MULTIVARIATE ANALYSES OF TRUNK TEXTURE PATTERNS FOR SUPPORTING TREE SPECIES IDENTIFICATION USING COMPUTATIONAL INTELLIGENCE** ..... 66
**Abstract** ..... 66
**1 Introduction** ..... 67
**2 Methods** ..... 68
2.1 Data collection for the experimental analysis ..... 68
2.2 Original variables extraction based on trunk texture patterns ..... 69
2.3 Synthetic variables generation from multivariate analyses ..... 70
2.4 Predicitive modeling and performance assessment ..... 72
**3 Results and discussion** ..... 74
**4 Conclusions** ..... 78
**References** ..... 78

**CAPÍTULO 5**

**ARBOREAL IDENTIFICATION SUPPORTED BY FUZZY MODELING FOR TRUNK TEXTURE RECOGNITION** ..... 84

**Abstract** ..... 84

**1 Introduction** ..... 85

**2 Methods** ..... 86

2.1 Data collection and feature extraction ..... 86

2.2 Fuzzy modeling for the pattern recognition ..... 90

2.3 Benchmarking experiment ..... 92

**3 Results and discussion** ..... 94

**4 Conclusions** ..... 99

**References** ..... 100

**CONSIDERAÇÕES FINAIS** ..... 104

**REFERÊNCIAS** ..... 107

## INTRODUÇÃO

As espécies arbóreas possuem características inatas que lhes atribuem vocações funcionais próprias no ambiente. Logo, particularidades que tornam uma espécie apta para certo fim podem ser prejudiciais à outras finalidades. Algumas espécies possuem madeira pesada e resistente, outras são moles e com baixa durabilidade (LORENZI, 1992). Há espécies com aspectos ornamentais e porte adequado para arborização urbana, já outras podem ser tóxicas, ter raízes que danificam edificações, ou altura e copa que interferem com o sistema viário e de iluminação pública (SOUZA et al., 2011, SILVA, 2009; SANTOS; TEIXEIRA, 2001).

Na dinâmica de sucessão primária, assim como nos processos de regeneração natural ou assistida pelo homem, as espécies arbóreas podem compor grupos ecológicos com funções e comportamentos distintos (MACIEL et al., 2003; GANDOLFI et al., 1995; BUDOWSKI, 1970). Espécies pioneiras e secundárias iniciais são mais tolerantes a certas condições, possuem crescimento mais rápido e ciclo de vida curto. Assim, as pioneiras atuam como colonizadoras, enquanto as secundárias iniciais criam condições para proliferação de espécies tardias e climácicas. Existem espécies que atraem e sustentam a fauna, fornecendo abrigo e alimento, mas também espécies exóticas invasoras que desequilibram o ecossistema e ameaçam a biodiversidade, além de causar impactos econômicos severos e, portanto, precisam ser reconhecidas e controladas (SAKAI et al., 2001; WIT; CROOKES; WILGEN, 2001).

Pelo exposto nesses exemplos, fica evidente que a identificação de espécies arbóreas é fundamental para diversas finalidades, tanto para o aproveitamento econômico, que inclui a produção madeireira, de alimentos e extração de produtos medicinais, quanto para fins ecológicos, como a conservação da biodiversidade, a recuperação de áreas degradadas, e o manejo da arborização urbana. Contudo, em certos casos essa identificação pode ser complexa, morosa, imprecisa e até mesmo impraticável (BACKES; CASANOVA; BRUNO, 2011; GOUVEIA et al., 1997).

Nesse contexto, o estudo de métodos computacionais para apoiar a identificação de espécies arbóreas vem se desenvolvendo nos últimos anos. Contudo, dada a complexidade biológica ainda há questões a serem superadas (YANIKOGLU; APTOULA; TIRKAZ, 2014; MACHADO et al., 2013; PRIYA; THANAMANI, 2012; KAUR; MONGA, 2012; BACKES; CASANOVA; BRUNO, 2011).

As técnicas atuais têm focado no reconhecimento de características extraídas de

imagens da folha (GOUVEIA et al., 1997; IM et al., 1998; IM et al., 1999; FU e CHI, 2003; QI e YANG, 2003; WANG et al., 2003; YE et al., 2004; LI et al., 2005; PLOTZE et al., 2005; DU et al., 2005; GU et al., 2005; LEE; CHEN, 2006; DU et al., 2006; DU et al., 2007; WU et al., 2007; NAM et al., 2008; WANG et al., 2008; BRUNO et al., 2008; CASANOVA et al., 2009; BACKES et al., 2009; SINGH et al., 2010; BACKES; BRUNO, 2010; BACKES et al., 2011; KADIR et al., 2011a, 2011b; ROSSATTO et al., 2011; CHAKI; PAREKH, 2011; MACHADO et al., 2013; YANIKOGLU et al., 2014).

Apesar dos avanços, a abordagem baseada em imagens da folha não atende a determinadas demandas, como no caso das espécies caducifólias em certos períodos do ano. Além das espécies caducifólias, que perdem suas folhas sazonalmente, há ainda os casos de árvores que foram cortadas e estruturas como as folhas não se preservaram. Logo, o reconhecimento de padrões baseados na textura em imagens do tronco arbóreo pode ser uma alternativa para apoiar a identificação de espécies usando inteligência computacional, mas ainda há poucos resultados reportados na literatura (CHI et al., 2003; WAN et al., 2004, SONG et al., 2004; HUANG et al., 2006, HUANG, 2006, POREBSKI et al., 2007, FIEL; SABLATNIG, 2011, KIM et al., 2011, BOMAN (2013).

# 1. Objetivos

## 1.1 Geral

Avaliar o uso de métodos de inteligência computacional no reconhecimento de padrões de textura em imagens do tronco arbóreo de espécies caducifólias nativas da flora brasileira.

## 1.2 Específicos

1.2.1 Revisar estudos sobre o uso da inteligência computacional no apoio à identificação de espécies arbóreas;

1.2.2 Avaliar o uso de propriedades estatísticas para reconhecimento de padrões de textura em imagens do tronco arbóreo;

1.2.3 Analisar o uso de padrões de coocorrência, individualmente e em conjunto com estatísticas de primeira ordem;

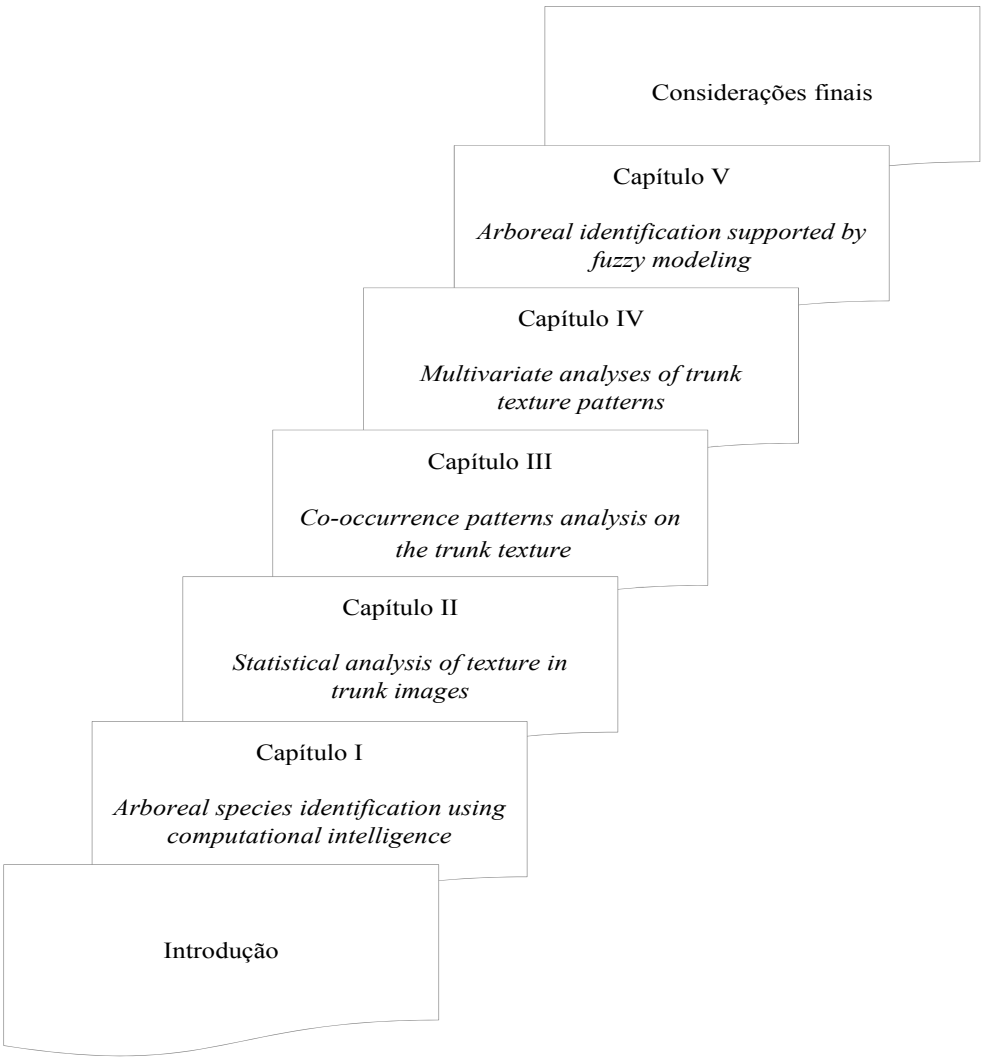
1.2.4 Estudar o uso de análises multivariadas para reforçar o desempenho de padrões de textura no tronco como características indicadoras de espécies arbóreas;

1.2.5 Analisar o uso da modelagem fuzzy para o reconhecimento da textura em imagens do tronco, em comparação com outros algoritmos de aprendizagem de máquina.

## 2. Estrutura da tese

A estrutura da tese foi organizada em sete seções. A primeira seção corresponde à introdução, na qual é apresentada uma contextualização, os objetivos geral e específicos, além da estrutura em si. As cinco seções intermediárias são compostas por capítulos, entre os quais artigos publicados ou submetidos para publicação em periódicos científicos. Por fim, a sétima seção apresenta as considerações finais (Figura 1).

Figura 1. Estrutura organizacional na composição da tese.



O capítulo 1 apresenta uma visão geral sobre identificação arbórea baseada em caracteres morfológicos e assistida por computador. Em seguida, são discutidos os avanços e limites do uso da abordagem computacional baseada em características foliares. Então, o reconhecimento de padrões de textura em imagens de tronco arbóreo é abordado como uma alternativa para superar limitações no uso de técnicas atuais (Objetivo 1.2.1). A partir dessa revisão foram identificados estudos anteriores, norteando a etapa experimental da pesquisa que buscou contribuir com análises originais, reportadas nos capítulos 2 a 5.

No capítulo 2 é desenvolvida a primeira análise experimental da pesquisa, dedicada a avaliar o uso de propriedades estatísticas no reconhecimento de padrões de textura em imagens do tronco (Objetivo 1.2.2). Para tanto, foram usadas 540 amostras de cinco espécies caducifólias nativas da flora brasileira.

O capítulo 3 traz uma análise comparativa do desempenho proporcionado por descritores de coocorrência e estatísticas de primeira ordem (Objetivo 1.2.3), por meio de uma análise experimental com 756 amostras de sete espécies arbóreas.

No capítulo 4, considerou-se que o uso de um número maior de características requer uma abordagem capaz de tratar informação redundante e, para isso, foi avaliado o uso de técnicas de análise multivariada (Objetivo 1.2.4). Para os procedimentos experimentais foram usadas 1188 amostras de onze espécies arbóreas.

A partir dos resultados dos experimentos anteriores, o capítulo 5 apresenta o estudo da modelagem fuzzy como uma alternativa para lidar com a incerteza no reconhecimento da textura em imagens do tronco, em comparação com outros algoritmos de aprendizagem (Objetivo 1.2.5). Para tanto, foram utilizadas 2160 amostras pertencentes a vinte espécies arbóreas.

As referências bibliográficas de cada capítulo são apresentadas ao final de cada seção, exceto aquelas relacionadas a essa seção introdutória, que se encontram após as considerações finais.

# CAPÍTULO 1

# ARBOREAL SPECIES IDENTIFICATION USING COMPUTATIONAL INTELLIGENCE

Adriano Bressane[1], José Arnaldo Frutuoso Roveda[2], Antonio Cesar Germano Martins[3], Maurício Tavares da Mota[4], Minoru Iwakami Beltrão[5]

[1] Environmental engineer, São Paulo State University (UNESP), Brazil

[2] Mathematician, University of Brasília (UnB), Brazil

[3] Physicist, University of Campinas (Unicamp), Brazil

[4] Biologist, Pontifical Catholic University of São Paulo (PUC), Brazil

[5] Agronomist, University of São Paulo (USP), Brazil

## Abstract

The computational intelligence has been used for dealing with complex issues in several fields of application. However, the use of computer-aided methods in some areas, as the biometric identification of arboreal species, still requires studies for achieving greater performance and acceptance. In this context, the present study aims to review the use of computational intelligence for that proposal, i.e., for supporting the arboreal species identification. After an overview of morphological and computer-aided identification, we discuss the advances and limits of using leaf-based approach. Then, the texture pattern recognition in arboreal trunk images is addressed as an alternative to overcome limitations in the use of current techniques. Finally, we conclude pointing out possibilities for approaching in future studies, as the analysis of more features, of multivariate analysis techniques, and of soft boundaries for further improve the arboreal species identification using computational intelligence.

Keywords: pattern recognition; image processing; taxonomy; computing techniques; bioinformatics.

## 1 Introduction

Over the years, the advancement of computational intelligence has provided new approaches to face old challenges. From performing a simple task up to supporting complex decision-making, the computer-aided methods have an ever-increasing number of applications in several fields, including precision crop management, industrial automation, medical procedures, and environmental assessment.

The computational intelligence (CI) comprises soft computing methods able to deal with complexity issues, common in the most practical applications. Thus, CI is considered a promising approach that may outperform methods of classical artificial intelligence, based on rigid inference mechanisms or hard computing techniques (Bittermann, 2011a, 2011b).

The digital image processing to extract features and the machine learning for pattern recognition are among the areas related to the usage of CI in the bioinformatics, i.e, in analysis of biological data using computationally intelligent systems (Saeys et al., 2007). Notwithstanding, although the computer intelligence is more consolidated in some fields of application, with specialized methods for solving well-defined tasks, other areas still require studies to achieve greater performance and acceptance, as in the case of the biometric identification of arboreal species (Machado et al., 2013; Aptoula and Yanikog, 2013; Priya and Thanamani, 2012).

In this context, this paper aims to review the use of computational intelligence for supporting the arboreal species identification. For that, an initial set of studies has been obtained through manual search on Google Scholar. Then, additional studies were identified considering the referenced publications and also studies that were citing the papers already identified. Thus, in section 2 we start from a brief overview of morphological characters and computer-aided identification. Section 3 discusses the advances and limits of using leaf image processing. Then, in section 4 we approach the texture recognition in arboreal trunk images as an alternative to overcome limitations in the use of current techniques. Finally, the last section presents our conclusions and perspective for future studies.

## 2 Morphological based identification and computer-aided approach

The arboreal identification process can be carried out by means of similarities comparison with specimens classified and stored in a herbarium (Bridson and Forman, 1998). Nevertheless, experts often uses an identification key, which takes into account similarities

and differentiations in terms of the form and structure of the arboreal features (Urbanetz et al., 2010). It is not unusual to combine both approaches, starting with an identification key to reach a broader classification and then to conclude at a herbarium.

The usage of reproductive morphological structures is more common because they suffer less alteration with habitat changes (Marchiori, 1995). However, in the absence of these reproductive characters the vegetative ones can be determinant, with the advantage of being used whatever time of the year (Batalha et al., 1998).

Therefore, identifying arboreal species may be a complex task and time-consuming, by requiring analysis of fertile branches, seeds, structure of the flower and fruit, leaf type and shape, bark features, shape and size of treetop, among other morphological characters, and also consider environmental conditions in the area of occurrence (Backes et al., 2011; Rossatto et al., 2011; Gouveia et al., 1997).

In this context, it is worth to highlight that the computational intelligence can be useful for supporting arboreal species identification. Nevertheless, it should not be intended to a self-sufficient system or completely autonomous for replacing the specialist's experience and interpretation. Instead, the usage of computational intelligence methods aims to extract and classify features which can be analyzed together with the morphological characters, and hence to support the identification by experts.

The computer-aided identification may be described by means of five main steps: data collection, digital image processing, feature extraction, pattern recognition, and machine learning. In turn, the last one is composed by the machine learning process (training and checking), and validation test. Thus, this approach starts with a data collection, obtaining pictures of the arboreal structure (Figure 1).

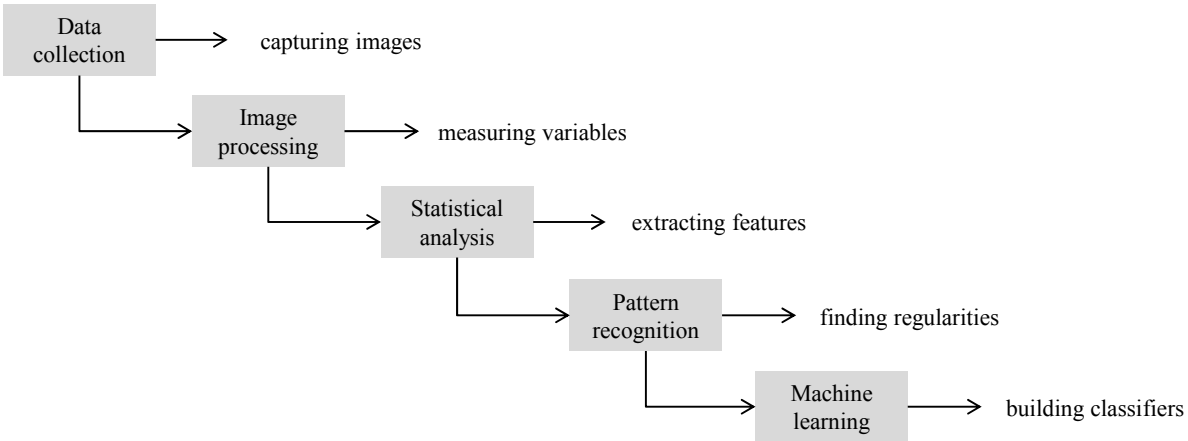



Figure 1. General steps of the computer aided arboreal species identification.

Then, a digital image processing is used to perform mathematical operations that produce data related to variables measured over the pictures. To improve preciseness, such variables can be statistically treated to find new dimensions generated by the features extraction with higher discriminant power. From that, the patterns recognition finds regularities in the dataset used for training and checking, allowing a classification into different categories. Finally, this classification is usually performed through a predictive model built with use of machine learning algorithms (Bishop, 2006).

## 3 Advances and limits of using leaf-based approach

The study of the leaf-based approach can be found in findings reported in the literature from the late 1990s, among which the study of Gouveia et al. (1997), which analyzed measures of area, dimensions of the enclosing rectangle, number of teeth and secondary veins, in order to use them for distinguishing different varieties of species.

Im et al. (1998) also studied leaves structural properties, from a polygonal approximation of their contour. Then, taking into account that the leaves are subject to undesirable deformations, in a subsequent survey the authors found that techniques of normalization to reduce variations allowed improving the recognition of the species in some cases (Im et al., 1999).

Fu and Chi (2003, 2006) used a segmentation approach based on histogram of pixels intensity over the leaf image, to extracted geometric parameters related to the leaves vascular system, then used them as predictor variables in an artificial neural network (ANN).

Qi and Yang (2003) focused on the feature extraction present in the edge of the leaf. For that, the authors applied a machine learning algorithm based on support vector machine (SVM) to classify sawtooth and nonsawtooth samples, obtained by a rectangular windows sliding along the leaf edge.

Wang et al. (2003) evaluated shape characterization functions for leaf image retrieval. Thus, using shape features referred to as centroid-contour distance (CCD), object eccentricity, and angle code histogram (ACH), the authors achieved better results than ones provided by modified Fourier descriptor (MFD) and Curvature Scale Space (CSS) methods.

Ye et al. (2004) presented a computerized plant species recognition system composed by two retrieval methods, the text-based information and features extracted from leaf image. In the last on the CCD method has been used for obtaining the leaf contour. Then, the leaf

apex, base and width-height ratio were calculated and applied to equate a similarity metric used for leaf retrieval.

Li et al. (2005) studied a method of segmentation known as snakes technique in combination with cellular neural networks (CNN) for improving preciseness and robustness in the extraction of vascular system and outlines, to subsequent leaf modeling and recognition.

In Plotze et al. (2005), measures of leaf vein and outline was also evaluated. Notwithstanding, in this study the authors applied a multiscale function of fractal dimension based on the Minkowski method for extracting theses morphometric characteristics from leaves image.

Du et al. (2005) compared the performance from different types of artificial neural networks in the classification of features based on the leaf shape, descripts by using a modified Fourier method. Thus, the authors found the probabilistic neural network (PNN) provided better results than radial basis function neural network (RBF), back propagation neural network (BPN), and multi-layer perceptron network (MLP).

Gu et al. (2005) proposed a combination of wavelet transform with gaussian interpolation for leaves retrieval based on run-length features extracted from the leaf skeleton, using as classifiers the *k*-nearest neighbor (*k*-NN), and radial basis probabilistic neural network (RBPNN).

Lee and Chen (2006) analyzed the use of region-based features extracted from leaf image, which included aspect ratio, horizontal and vertical projections, compactness, and centroid. By extracting features from the regions of interest (ROI), the authors achieved better results than ones provided by contour-based methods.

Du et al. (2006) adopted an accelerated Douglas-Peucker algorithm for a leaf shape polygonal approximation. Then, the authors assessed the use of modified dynamic programming (MDP) method for leaf recognition based on shape matching, in comparison with other methods, as modified Fourier descriptors (MFD), Hu moment invariants (HM), contour moment (CM), curvature scale space (CSS), and geometrical features (GF).

In Du et al. (2007) a method described as move median centers (MMC) hypersphere classifier is introduced for recognizing leaf using contour-based approach. By considering Hu moment invariants and geometrical features, as rectangularity, convexity, circularity, eccentricity, and form factor, the authors reached some improvement in more complex cases.

Wu et al. (2007) also used a probabilistic neural network (PNN) for leaf retrieval based on shape information, already assessed in comparison with other algorithms by previous studies. Nevertheless, before the machine learning process, the authors performed a principal component analysis (PCA), in order to reduce the data dimensionality.

Nam et al. (2008) studied the use of morphological information (external shape) in combination with (nervure) vascular system features for leaf retrieval, using a scheme based on the similarity degree between images. For that, the authors implemented an adaptive grid-based matching algorithm that outperformed other existing methods.

Wang et al. (2008) analyzed the use of marker-controlled watershed method in combination with pre-segmentation and morphological operation to segment leaf images with complicated background. After image segmentation, shape features based on Hu geometric moments and sixteen Zernike moments has been extracted and then used for leaf retrieval with a moving center hypersphere (MCH) classifier.

Bruno et al. (2008) compared the box-counting and multiscale Minkowski–Sausage methods for estimating fractal dimensions in analyzes of the leaf complexity. Thus, by taking into account internal and external morphological features, the authors discuss the best approach for supporting the species identification from the leaf shape.

Casanova et al. (2009) assessed the usage of Gabor wavelet filters for extracting and discriminating texture patterns in the foliar surface, in order to improve the species identification accuracy by adding these texture features to other leaf morphological attributes.

Backes et al. (2009) presented the usage of the volumetric fractal dimension (VFD) method for analyzing, describing, and characterizing the complexity related to the leaf texture patterns that presents a huge variation. By using this approach the authors produced a texture signature able to improve traditional techniques as Gabor filters and Fourier descriptors.

Singh et al. (2010) evaluated the performance of Support Vector Machine utilizing Binary Decision Tree (SVM-BDT) in comparison with Probabilistic Neural Network (PNN) and Fourier moment technique (FMT). By using leaf morphological features, the authors found that the SVM-BDT classifier provided the best accuracy.

Backes and Bruno (2010) proposed an approach based on color texture analysis for leaf classification using fractal dimension. For that, the authors modeled each color channel from the foliar image as a surface. Then, the complexity of the surfaces has been analyzed using Bouligand-Minkowski and multiscale fractal dimension, overcoming other methods as chromaticity moments and Gabor EEE descriptors.

In subsequent study, Backes et al. (2011) also assessed the approach based on multiscale fractal dimension, using the box counting in combination with the Otsu method, in comparison with other techniques. As a result, the fractal analysis and the co-occurrence matrices method achieved higher performances than Gabor filters and Fourier descriptors.

Kadir et al. (2011a) performed a comparative experiment of methods for recognizing species using morphological features extracted from leaf images. Thus, the authors verified that polar Fourier transform (PFT) method outperformed approaches based on geometric features, Zernike orthogonal moments, and Hu moment invariants. By using a PNN as classifier, these authors also studied the use of shape features captured by PFT method in combination with color moments, vein and texture features (Kadir et al., 2011b).

Rossatto et al. (2011) also studied the volumetric fractal dimension (VFD) method for recognizing species based on leaf-texture properties. Nevertheless, for that the authors used a naive Bayes classifier that assumes a conditional independence hypothesis. Then, a canonical discriminant analysis (CDA) has been performed for removing the correlations among features and maximizing the separation among classes.

By using artificial neural networks as classifiers for recognizing leaf images, Chaki and Parekh (2011) discuss the use of the centroid-radii (CR) model for estimating shape-based features, in comparison with Hu moments invariant (MI) method, already assessed in studies aforementioned. In addition, the combined use of features from both methods (CR and MI) has been explored in order to find the best performance.

Before using the Bouligand-Minkowski method to estimate fractal dimensions over the leaf image, Machado et al. (2013) analyzed the application of non-linear partial differential equations (PDE) of Perona-Malik for enhancing the texture components. Thus, based on classification experiments with usage of linear discriminant analysis (LDA), the authors found that the proposed approach allows improving the performance in the leaf identification.

Yanikoglu et al. (2014) implemented an approach based on a large set of features. As color-based features were used color moments, even as RGB histogram, LSH histogram, and the saturation-weighted hue histogram. The texture futures consisted of orientation histogram and Gabor wavelets. In turn, shape-based features included Fourier descriptors, perimeter and area convexity, compactness, elongation, basic shape statistics, area width factor, regional moments of inertia, angle code and contour point distribution histogram. These features have been assessed individually and in group using a support vector machine (SVM) as classifier.

Thus, the authors achieved promising result in recognizing isolated leaves images, but for unconstrained photographs, with complex background, the performance was considered unsatisfactory.

Although the species identification supported by computer methods is considered a recent area with outcomes still insufficient to completely solve the involved issues, we verified a significant number of studies focused on leaf properties analyzes, such as color and texture, but mainly shape-based features, both external and internal.

By analyzing these studies, we can found the analysis of several techniques for extracting features and recognizing patterns, in order for improving the leaf information retrieval and, consequently, supporting the species identification.

In this sense, the results reported in the literature presents important advancements to face the biological complexity that requires, among other challenges, to deal with a huge variation in the morphological features. Moreover, the leaf features are also quite sensible to foliar maturity level, sun exposure, soil properties, and other environmental influences, as weather, pollution and diseases.

Despite this, it was noted that the achievement of higher performances demands a leaves collect and preparation, for an image acquisition by scanning or photographing of the isolated sample. Thus, the samples are subject to manipulation and decomposition process that starts after collection and may impair the conservation of characteristics, and hence make hardier their recognizing.

## 4 Texture patterns recognition from the arboreal trunk images

In general, texture patterns in arboreal trunk images are related to the presence, arrangement and dimension of bark features. According to Martins-da-Silva (2002) the arboreal bark can be smooth (without salience or depression), striated (with small grooves), fissured (with deep grooves), cancerous (with small craters kind of rounded), powdery (covered with dust), with protrusions (saliences kind of rounded but without openings), with lenticels, spines or aculeus, and detaching themselves as fine pieces, coriaceous pieces, and thick plaques (Figure 2).

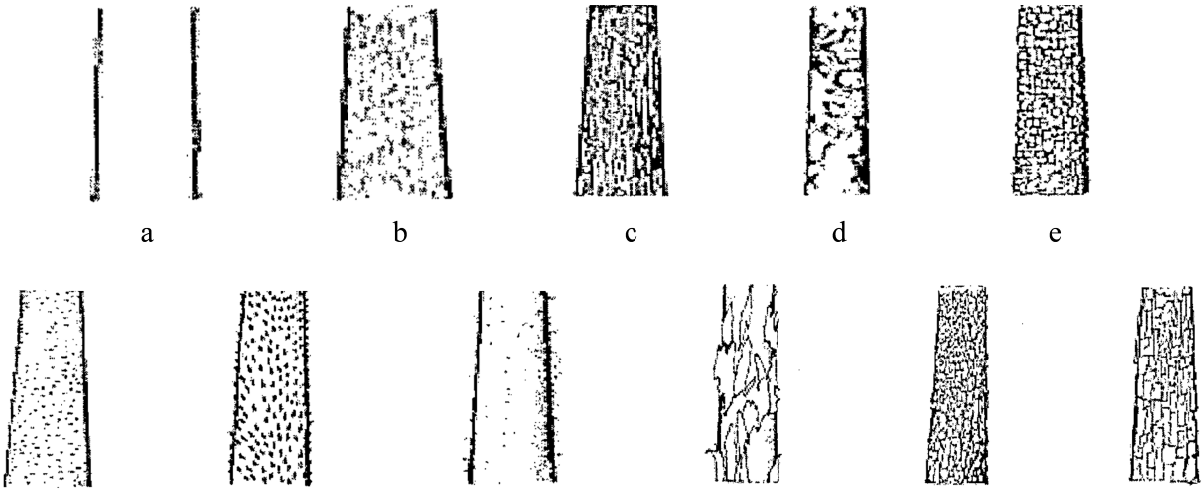



Figure 2. Bark features with influence on texture in arboreal trunk images: (a) smooth, (b) striated, (c) fissured, (d) cancerous, (e) with protrusions, (f) with lenticels, (g) spines or aculeus, (h) powdery, detaching themselves (i) as fine pieces, (j) as coriaceous pieces, and (k) thick plaques. Source: adapted from Martins-da-Silva, R.C.V. 2002. Coleta e identificação de espécimes botânicos (p. 28). Belém-PA: Embrapa (Série Documentos 143).

Taking into account that the bark features are relatively uniform by species (Wojtech and Wessels, 2011; Vaucher, 2010), the arboreal trunk texture recognition can be an alternative for further improving the computer-aided identification, mainly when the foliar structures are not available or are insufficient, as the case of deciduous species in certain seasons of the year.

Chi et al. (2003) evaluated the use of Gabor filter banks to characterize different texture patterns based on central frequencies and normalized rations of amplitudes, extracted by discrete Fourier Transform method, and modeled as multiple narrowband signals, achieving promising results.

Wan et al. (2004) compared the texture features performance extracted by run-length, auto-correlation (ACM), histogram (HM), and co-occurrence matrices method (COMM). Using the $k$-nearest neighbor ($k$-NN) and moving median centers hypersphere classifiers, the authors conclude that COMM features were superior to the ones afforded by the other methods.

By combining grayscale features extracted by COMM and binary texture patterns, Song et al. (2004) achieved better results than that when each features set was used individually. In turn, Huang et al. (2006) also used features obtained by COMM in combination with fractal dimension descriptors, in order to compare the performance of different artificial neural networks topologies.

In Huang (2006), texture features were studied in combination with color information, both extracted using multi-resolution wavelet method and classified by radial basis probabilistic neural network (RBPNN) and support vector machine (SVM). As a result, the author highlights that by combining color and texture features the RBPNN was better than past performances using only features extracted by COMM, ACM and HM.

In this context, the use of color information seemed another interesting research direction, but the total number of candidate features also became very high. Then, Porebski et al. (2007) proposed an iterative procedure for selecting the most discriminating texture features extracted by COMM in different color spaces, and classified using $k$-NN method.

Fiel and Sablatnig (2011), even as Kim et al. (2011), experienced the combination of features (shape, color and texture) of different tree parts, including leaf, needles, flower and bark texture in the tree trunk. In Kim et al. (2011) the trunk texture was characterized by grayscale and binary features recognized using multi-resolution wavelet method (MWM). Besides the MWM, Fiel and Sablatnig (2011) also used COMM for extracting features and SVM as classifying model. In both studies the features combination afforded better results than when they were used individually.

Boman (2013) performed a comparison between SVM and the import vector machine (IVM) classifiers, in the pattern recognition based on features extracted by grey level co-occurrence matrix (GLCM), scale invariant feature transform (SIFT), wavelet with GLCM (WGLCM), and wavelet co-occurrence histogram (WCH) methods. As a result, the best performance was obtained by SVM with GLCM.

## 5 Conclusions

The usage of computational intelligence for identifying arboreal species is considered relatively recent, but it has been developed over the last 20 years. By reviewing previous studies, it was found that the outcomes reported in the literature have focused on the leaf image processing.

Indeed, in the absence of reproductive morphological structures, the leaf features are among the most important vegetative characters used by experts for characterizing species. Besides that, the shape approximately two-dimensional of leaves is other factor that encourages the digital image processing for subsequent features extraction and retrieval.

On the other hand, when there is no physical collection of leaves for photographing or digital scanning, i.e, in the cases where the images are captured in field, its segmentation for removing overleaping or background elements in the image, make its use more complex and less efficient. Moreover, deciduous species lose their leaves seasonally, making the use of leaf-based approach impractical certain periods of year. However, analyzes of texture in arboreal trunk images are still understudied, with fewer outcomes reported in the literature.

From the foregoing, in future studies we intended to perform experiments for analyzing more texture features in trunk images, as the use of first-order statistics, individually, and in combination with co-occurrence descriptors. Furthermore, multivariate analysis techniques will be also evaluated for optimizing the computational effort and improving preciseness during the machine learning. After that, the usage of the soft boundaries, by means of fuzzy modeling, will be experienced for dealing with ambiguity in the pattern matching based on texture in trunk images.

**References**

Aptoula, E., Yanikoglu, B. 2013. Morphological features for leaf based plant recognition. *IEEE International Conference on Image Processing*, p. 1496–1499.

Backes, A.R., Casanova, D., Bruno, O.M. 2011. Identificação de plantas por análise de textura foliar. *Anais do VI Workshop de Visão Computacional*, Presidente Prudente.

Backes, A.R., Bruno, O.M. 2010. Plant leaf identification using color and multi-scale fractal dimension. *Lecture notes on computer science*. 6134: 463-470.

Backes, A.R., Casanova, D., Bruno, O.M. 2009. Plant leaf identification based on volumetric fractal dimension. International. *Journal of pattern recognition and artificial intelligence*. 23(6): 1145-1160.

Batalha, M.A., Aragaki, S., Mantovani, W. 1998. Chave de identificação das espécies vasculares do cerrado em Emas (Pirassununga, SP), baseada em caracteres vegetativos. *Boletim de Botânica da Universidade de São Paulo*. 17, 85-108.

Bishop, C.M. 2006. *Pattern recognition and machine learning*. Cambridge: Springer.

Bittermann, M.S. 2011a. A computational design system with cognitive features based on multi-objective evolutionary search with fuzzy information processing. In John S. Gero (ed), *Design Computing and Cognition '10*, p. 505-524. Netherlands: Springer.

Bittermann, M.S. 2011b. Artificial intelligence versus computational intelligence for treatment of complexity in design. In: *Proceedings of Workshop Assessing the Impact of Complexity Science in Design at Design Computing and Cognition '10* – DCC'10, Stuttgart 1-8.

Boman, J. 2013. *Tree species classification using terrestrial photogrammetry*. Umeå: Umeå University.

Bridson, D., Forman, L. 1998. *The herbarium handbook*. Royal Botanic Gardens.

Bruno, O. M., Plotze, R., O., Falvo, M., Castro, M. 2008. Fractal Dimension applied to plant identification. *Information Sciences*, 178: 2722-2733.

Casanova, D., Bruno, O.M. 2009. Plant leaf identification using Gabor wavelets. *International journal of imaging systems and technology*, 19(3): 236-243.

Chaki, J., Parekh, R. 2011. Plant leaf recognition using shape based features and neural network classifiers. *International journal of advanced computer science and applications*, 2(10): 41-47.

Chi, Z., Houqiang, L., Chao, W. 2003. Plant species recognition based on bark patterns using novel Gabor filter banks. In: *Proceedings of the 2003 International conference on neural networks and signal processing*.

Du, J., Huang, D., Wang, X., Gu, X. 2005. Shape recognition based on radial basis probabilistic neural network and application to plant species identification. *Lecture Notes in Computer Science*, 3497: 281-285.

Du, J. X., Wang, X. F., Gu, X. 2006. Computer-aided plant species identification (CAPSI) based on leaf shape matching technique, *Transactions of the Institute of Measurement and Control*, 28 (3): 275-284.

Du, J.X., Wang, X.F., Zhang, G.J. 2007. Leaf shape based plant species recognition. *Applied mathematics and computation*, 185: 883-893.

Fiel, S., Sablatnig, R. 2011. Automated identification of tree species from images of the bark, leaves and needles. In: *Proceedings of the 2011 Computer vision winter workshop*.

Fu, H., Chi, Z. 2003. A two-stage approach for leaf vein extraction. In: IEEE International Conference on Neural Networks and Signal Processing, *Proceedings*…, Nanjing, 208 - 211, 2003.

Gouveia, F., Filipe, V., Reis, M., Couto, C., Bulas-Cruz, J. 1997. Biometry: the characterisation of chestnut-tree leaves using computer vision. In: IEEE International Symposium on Industrial Electronics, *Proceedings*…, Guimarães, p. 757-760.

Gu, X., Du, J.X., Wang, X.F. 2005. Leaf recognition based on the combination of wavelet transform and gaussian interpolation. *Lecture notes in computer science*. 3644: 253-262.

Huang, Z.K., Huang, D.S., Du, J., Quan, Z.H., Guo, S.B. 2006. Bark classification based on textural features using Artificial Neural Networks. *Lecture notes in computer science*, 3972: 355-360.

Huang, Z.K. 2006. Bark classification using RBPNN based on both color and texture feature. *International journal of computer science and network security*. 6(10):100-103.

Im, C., Nishida, H., Kunii, T.L. 1998. Recognizing plant species by leaf shapes-a case study of the Acer family. *Pattern recognition*, 2:1171-1173.

_____. 1999. Recognizing plant species by normalized leaf shapes. *Vision Interface*, 19(21): 397-404.

Kadir, A., Nugroho, L.E., Susanto, A., Santosa, P.I. 2011a. A comparative experiment of several shape methods in recognizing plants. *International Journal of Computer Science & Information Technology*, 3(3): 256-263.

_____.2011b. Leaf classification using shape, color, and texture. *International Journal of Computer Trends and Technology*, 2(1): 225-230.

Kim, S.J., Kim, B.W., Kim, D.P. 2011. Tree recognition for landscape using by combination of features of its leaf, flower and bark. In: *Proceedings of the 2011 Society of Instrument and Control Engineers Annual Conference*.

Lee, C.L., Chen, S.Y. 2006. Classification of leaf images. *International journal of imaging systems and technology*. 16(1): 15-23.

Li, Y., Zhu, Q., Cao, Y., Wang, C. 2005. A leaf vein extraction method based on snakes technique. In: IEEE International Conference on Neural Networks and Brain, *Proceedings*…. Beijing, p. 885-888.

Machado, B.B., Casanova, D., Gonçalves, W.N., Bruno, O.M. 2013. Partial differential equations and fractal analysis to plant leaf identification. *Journal of Physics* 410, 1-4. doi:10.1088/1742-6596/410/1/012066

Marchiori, J.N.C. 1995. *Elementos de Dendrologia.* Santa Maria: Editora UFSM.

Nam, Y., Hwang, E.J., Kim, D.Y. 2008. A similarity-based leaf image retrieval scheme: joining shape and venation features. *Computer vision image understanding*. 110:245-259.

Plotze, R.O., Falvo, M., Pádua, J.G., Bernacci, L.C., Vieira, M.L.C., Oliveira, G.C.X., Bruno, O.M. 2005. Leaf shape analysis by the multiscale minkowski fractal dimension, a new morphometric method: a study in *Passiflora L*. (Passifloraceae). *Canadian journal of botany*.

83(3): 287-301.

Porebski, A., Vandenbroucke, N., Macaire, L. 2007. Iterative feature selection for color texture classification. In: *Proceedings of the 2007 IEEE International conference on image processing.*

Priya, A.C., Thanamani, A.S. 2012. A survey on species recognition system for plant classification. *International Journal Computer Technology & Applications*, 3(3): 1132-1136.

Qi, H., Yang, J.G. 2003. Sawtooth feature extraction of leaf edge based on support vector machine. *Machine learning and cybernetics*. 5:3039-3044.

Rossatto, D.R., Casanova, D., Kolb, R.M., Bruno, O.M. 2011. Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with species from Neotropical Melastomataceae. *Plant Systematics and Evolution* 291, 103-116. doi:10.1007/s00606-010-0366-2.

Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23 (19): 2507-2517.

Martins-da-Silva, R.C.V. 2002. *Coleta e identificação de espécimes botânicos*. Belém-PA: Embrapa (Série Documentos 143).

Singh, K., Gupta, I., Gupta, S. 2010. SVM-BDT PNN and Fourier Moment Technique for classification of leaf shape. *International journal of signal processing, image processing and pattern recognition*. 3(4):67-78.

Song, J., Chi, Z., Liu, J., Fu, H. 2004. Bark classification by combining grayscale and binary texture features. In: *Proceedings of the 2004 Intelligent multimedia, video and speech processing*.

Urbanetz, C.; Tamashiro, J.Y.; Kinoshita, L. S. 2010. Chave de identificação de espécies lenhosas de um trecho de Floresta Ombrófila Densa Atlântica, no Sudeste do Brasil, baseada em caracteres vegetativos. *Biota Neotropica*. 10: 349-398.

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

Wan, Y.Y., Xiang, D.J., Huang, D.S., Chi, Z., Cheung, Y., Wang, X.F., Zhang, G.J. 2004. Bark texture feature extraction based on statistical texture analysis. In: *Proceedings of the 2004 Intelligent multimedia, video and speech processing.*

Wang, X., Huang, D.S., Du, J.X., Xu, H., Heutte, L. 2008. Classification of plant leaf images with complicated background. *Applied mathematics and computation*. 205:916-926.

Wang, Z., Chi, Z., Feng, D., Wang, Q. 2003. Leaf image retrieval with shape features. *Lecture notes in computer science*. 1929 (2000):477-487.

Wojtech, M., Wessels, T. 2011. *Bark: a field guide to trees of the northeast*. New England: UPNE.

Wu, S.G., Bao, F.S., Xu, E.Y., Wang, YX., Chang, Y.F., Xiang, Q.L. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. *The computing research repository*. 1:11-16.

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

Ye, Y., Chen, C., Li, C.-T., Fu, H., Chi, Z. 2004. A computerized plant species recognition system. In: International Symposium on Intelligent Multimedia, Machine vision and applications, *Proceedings*…, p. 723 - 726, Hong Kong.

## CAPÍTULO 2

# STATISTICAL ANALYSIS OF TEXTURE IN TRUNK IMAGES FOR BIOMETRIC IDENTIFICATION OF TREE SPECIES [a]

Adriano Bressane[1], José Arnaldo Frutuoso Roveda[2], Antonio Cesar Germano Martins[3]

[1] Environmental engineer, São Paulo State University (UNESP), Brazil

[2] Mathematician, University of Brasília (UnB), Brazil

[3] Physicist, University of Campinas (Unicamp), Brazil

## Abstract

The identification of tree species is a key step for sustainable management plans of forest resources, as well as for several other applications that are based on such surveys. However, the present available techniques are dependent on the presence of tree structures, such as flowers, fruits and leaves, limiting the identification process to certain periods of the year. Therefore, this article introduces a study on the application of statistical parameters for texture classification of tree trunk images. For that, 540 samples from 5 Brazilian native deciduous species were acquired and measures of entropy, uniformity, smoothness, asymmetry (third moment), mean and standard deviation were obtained from the presented textures. Using a decision tree, a biometric species identification system was constructed and resulted a 0.84 average precision rate for species classification with 0.83 accuracy and 0.79 agreement. Thus, it can be considered that the use of texture presented in trunk images can represent an important advance in tree identification, since the limitations of the current techniques can be overcome.

Key words: Image processing; Statistical parameters; Image texture, Tree identification.

[a] Published in Environmental Monitoring and Assessment. 2015, 187(4): 212.
DOI 10.1007/s10661-015-4400-2.

## 1 Introduction

In constant development, image processing has been widely used in various areas, with several applications, including multidisciplinary studies, such as vegetation (Zehm, Nobis and Schwabe, 2003) and agricultural analysis (Vibhute and Bodhe, 2012) or for conservation and environmental management purposes (Yemshanov, McKenney and Pedlar, 2012; Pu, 2011;Ge et al., 2006; Weber and Glenn, 2001).

However, the study of computer vision methods to identify species of plants is still a new area, with significant growth potential (Machado et al., 2013). For instance, with texture fractal analysis, Casanova, Florindo and Bruno (2011) achieved promising results with almost 50% accuracy in identifying plants from European countries, highlighted by the authors as a rate much higher than that of related works. Also based on the analysis of images from leaves, there are several other works in the literature Silva et al. (2014), Sá Júnior et al. (2013), Rossato et al. (2011), Sá Júnior et al. (2011), Oliveira and Bruno (2009), Casanova and Bruno (2009), Backes, Casanova and Bruno (2009).

In general, species identification based on analysis of images from leaves is considered an advance over conventional morphological techniques which are limited to certain times of year, as they are commonly done from flowers and fruits. However, in the case of deciduous species, that lose their leaves during cold and dry seasons and may not present flowers or fruits, analyzes based on images of leaves are not suitable. The requirement on the presence of leaves also restricts the application of such techniques in the case of tree identification when they were cut and those morphological structures had been removed.

Therefore, this paper introduces a study on the application of statistical properties for classification of texture patterns in images from the trunk that can lead to a biometric tree species identification system.

## 2 Materials and methods

For the present study, images of 5 tree species from the Brazilian native deciduous forest were acquired at Biquinha Municipal Natural Park, a nature conservation unit composed of forest remnants and isolated arboreal individuals, located in the city of Sorocaba, São Paulo, Brazil (see Figure 1).

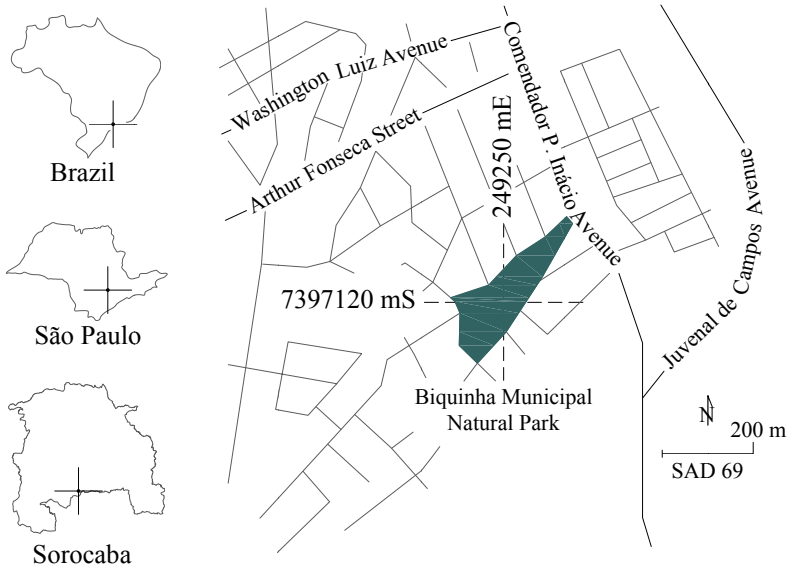



Figure 1. Location of Biquinha Municipal Natural Park, in the city of Sorocaba, São Paulo, Brazil.

The sample set was obtained from isolated trees (open grown) using a conventional digital camera, recording 12 images with 2560 x 1920 pixels, for each of the following species: *Anadenanthera falcata*, *Gochnatia polymorpha*, *Cedrela fissilis, Chorisia speciosa* and *Schizolobiun parahyba*. The images from each species were taken at different heights of the trunk, all around the trees, discarding shaded areas or with other interference, such as the presence of insects.

From each image a central area of 1024 x 1024 pixels was cut, and, using a moving mask of 512 x 512 pixels displaced by 128 pixels in the horizontal and vertical directions, nine images were further obtained, generating a total of 540 images, with 108 for each species, from which 80 were used for the construction of the classification system and 28 for performance tests. Figure 2 shows examples of images for each of the 5 tree species.


Figure 2. Trunk images with 512 x 512pixels from: (a) *Chorisia speciosa,* (b) *Schizolobiun parahyba,* (c) *Gochnatia polymorpha*, (d) *Cedrela fissilis*, (e) *Anadenanthera falcata*.

For the present investigation, the images were transformed from the RGB to the HSV space (Gonzales and Woods, 2008), and the V channel was used in the study.

2.1 Construction of the system

Texture from the 400 images (80 for each species) was analyzed in the spatial domain based on four statistical parameters: uniformity, entropy, asymmetry, and smoothness.

Uniformity (*U*) that measures how close gray level intensity are in the image (Gonzales, Woods and Eddins, 2009), is obtained by:

$$U = \sum_{i=0}^{L-1} p^2(z_i) \tag{1}$$

where *L* is the number of gray levels, $z_i$is the intensity value of pixel *i*, and $p(z_i)$is the image histogram.

Based on the randomness of the gray levels in the image that brings the information of the structure organization presented (Jain, Rangachar and Schunck, 1995), entropy (e) can be calculated by:

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \tag{2}$$

Third moment, also known as Asymmetry ($\mu_3$), and smoothness (*R*), that takes in to account the transition between the shades of gray in the image, are respectively obtained by:

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - \mu_1)^3 p(z_i) \tag{3}$$

$$R = 1 - \frac{1}{1 + {\mu_2}^2} \tag{4}$$

where, $\mu_1$ is the average intensity (or first moment), and $\mu_2$ is the standard deviation (or second moment), obtained by:

$$\mu_1 = \sum_{i=0}^{L-1} z_i p(z_i) \tag{5}$$

$$\mu_2 = \sqrt{\frac{\sum_{i=1}^{n} (z_i - \mu_1)^2}{n-1}} \tag{6}$$

where *n* is the number of pixels in the image.

The output from each of the four statistical parameters were normalized to the range [0;1]. Histograms with those values were constructed, from which, thresholds limits between distributions associated with different texture patterns were obtained in an attempt to separate species.

With those limits, the classification system was built in the form of a decision tree in which a given statistical property permitted the implementation of a logical operation that separates the samples according to a binary rule based on relevant ranges and thresholds. Those operations are sequentially integrated forming the branches of the decision tree. When the addition of a new operation did not provide significant gains in the classifying ability of the system, the growth of the tree in that direction was terminated and an output $S_j$ was obtained.

2.2 Classification performance of the constructed system

The classifying ability of the system was evaluated through a hit rate for each species ($H_{\text{rate}}$) and an average hit rate ($\overline{H}_{\text{rate}}$), given by:

$$H_{\text{rate}}(sp_i) = \sum_{j=1}^{k} (P^{S_j} . p_A^{S_j} . S_D^j) \quad (7)$$

$$\overline{H}_{\text{rate}} = \sum_{i=1}^{l} (H_{rate}^{sp_i} . p_{sp_i}) \quad (8)$$

where $k$ is the number of outputs for the classifying system, $P^{S_j}$ is the probability that a sample from species ($sp_i$) is in the output $S_j$, $p_A^{S_j}$ is the sample coefficient for $sp_i$ in $S_j$, $S_D^j$ is the dominant output for $S_j$, and $p_{sp_i}$ is the ratio of samples from species $i$ by the total number of samples used in the construction of the system.

The probability that a sample from $sp_i$ is part of the output $S_j$ was estimated by the ratio of the number of samples from $sp_i$ in $S_j$ $\left( n_{sp_i}^{S_j} \right)$ by the number of output samples in $S_j$ $\left( n_T^{S_j} \right)$:

$$P^{S_j}(sp_i) = n_{sp_i}^{S_j} / n_T^{S_j} \quad (9)$$

To obtain the coefficient of species samples in a given output, it was used the ratio of the number of samples of that species ($sp_i$) contained in the output set ($S_j$) by the total initial samples for the same specie $\left( n_T^{sp_i} \right)$:

$$p_A^{S_j}(sp_i) = n_{sp_i}^{S_j} / n_T^{sp_i} \quad (10)$$

To study the dominant output for each species $sp_i$, a binary value defined as a function of the probability that the sample belongs to $sp_i$ was obtained in such a way that:

$$S_D^j(spi) = \begin{cases} 1, & \text{if } P^{S_j}(sp_i) \text{ is maximum for all the } sp_i \text{ in the output } S_j \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

It is emphasized that the classification system is probabilistic, once it estimates the probability that a sample from a give species is in each output ($S_j$).

To evaluate the system performance (validation), the results of the classification with

the test samples were analyzed using metrics derived from the method of Confusion Matrix (Kohavi and Provost, 1998), including: Precision ($P_{rate}$), Error ($E_{rate}$), Sensitivity ($S_{rate}$), Accuracy ($\theta_1$) and *Kappa.*

Precision rate ($P_{rate}$) for each species $sp_i$ was estimated based on the ratio of correctly classified samples ($T_{sp_i}$) by the total number of samples identified as belonging to species $sp_i$ ($I_{sp_i}$):

$$P_{rate} = \frac{T_{sp_i}}{I_{sp_i}} \tag{12}$$

Error rate ($E_{rate}$) for each species $sp_i$ was estimated using the ratio of samples miss classified ($F_{sp_i}$) by the total number of samples identified as belonging to species $sp_i$ ($I_{sp_i}$):

$$E_{rate} = \frac{F_{sp_i}}{I_{sp_i}} \tag{13}$$

Sensitivity or hit rate ($S_{rate}$) for each species $sp_i$ was estimated as the ratio of correctly classified samples ($T_{sp_i}$) over the total number of samples actually belonging to species $sp_i$ ($V_{sp_i}$):

$$S_{rate} = \frac{T_{sp_i}}{V_{sp_i}} \tag{14}$$

Accuracy ($\theta_1$), or rate of overall accuracy of the classifier was estimated by the ratio of correctly classified samples in all species evaluated by the total number of samples ($n_T$):

$$\theta_1 = \frac{1}{n_T} \sum_{i=1}^{n_{sp}} T_{sp_i} \tag{15}$$

where $n_{sp}$ is the number of species.

Finally, to further evaluate the system, comparing the output and the true species classification (Carletta, 1996), Kappa index (*K*), defined by:

$$K = \frac{\theta_1 - \theta_2}{1 - \theta_2} \tag{16}$$

with:

$$\theta_2 = \frac{1}{{n_T}^2} \sum_{i=1}^{n_{sp}} (V_{sp_i} . I_{sp_i}) \tag{17}$$

was used.

## 3 Results and discussion

### 3.1 Statistical properties of texture from trunk images

By analyzing uniformity, it can be seen from Figure 3 that this parameter enables the separation of 100% samples from *Chorisia speciosa* using a threshold value of 0.36.

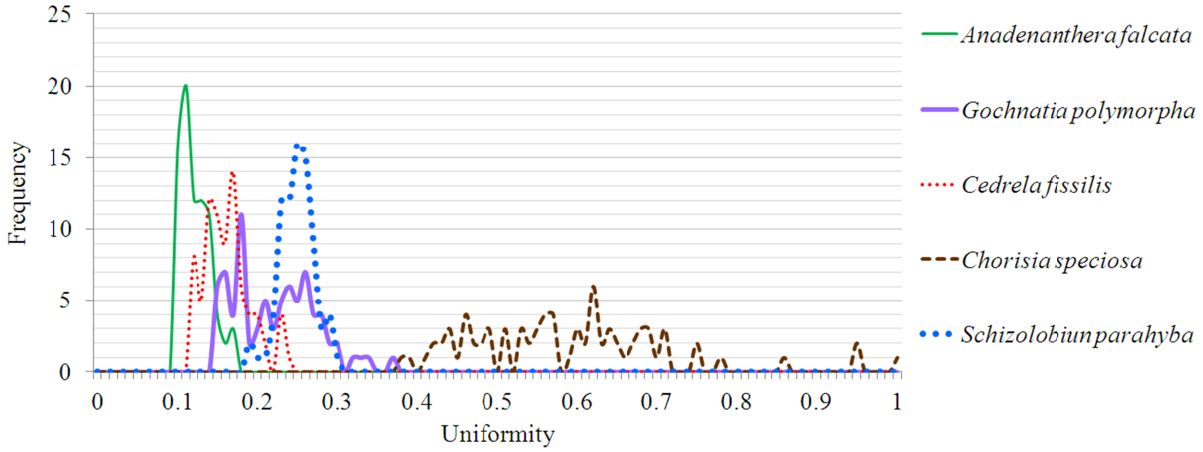



Figure 3. Histogram for uniformity values from the trunk images.

Although this distribution separation does not occur for the other species, it can be noted that there are intervals with predominance of certain ones, such as for uniformity values less than or equal to 0.14, where *Anadenantera falcata* is dominant, values in the range]0.14; 0.18] and ]0.19; 0.36] with higher frequency of *Cedrela fissilis* and *Schizolobiun parahyba,* respectively, indicating that the use of other statistical measurements in such an interval could lead to its correct classification.

Analyzing smoothness parameter, with zero being assigned to constant intensity and values close to 1 when there are abrupt changes (Gonzales and Woods, 2010), 93% separation of *Schizolobiun parahyba* images was obtained using a threshold limit of 0.2, although samples from *Gochnatia polymorpha* and *Chorisia speciosa* are mixed within (see Figure 4).Nevertheless, as shown ahead, using uniformity as a previous operation, *Chorisia speciosa* presence in the interval [0.0;0.2[ can be identified, and *Gochnatia polymorpha* can be further segregated using moment metrics (asymmetry) as following operations.

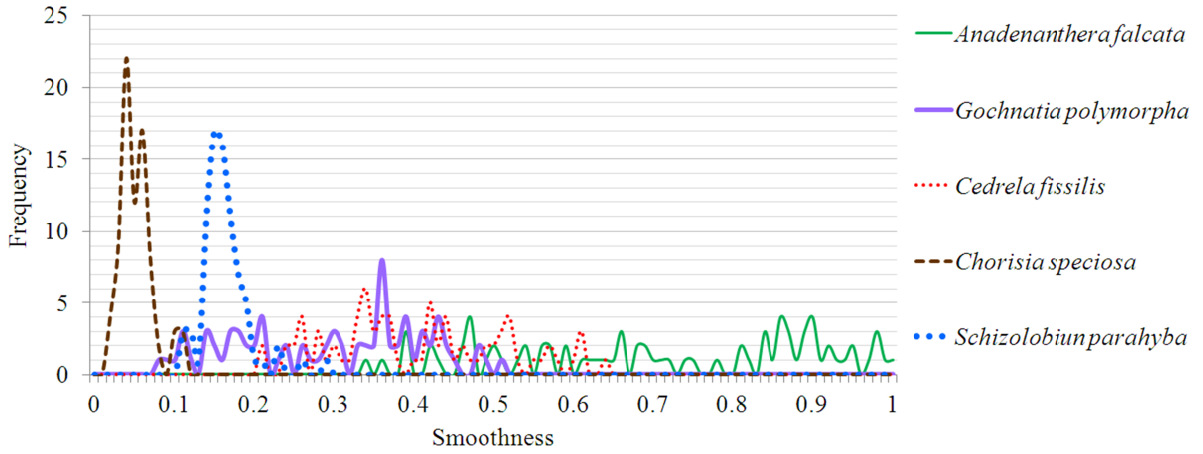



Figure 4. Histogram for smoothness values from the trunk images.

Analyzing the third moment as a measure of the histogram asymmetry, with zero meaning symmetric distributions, and positive or negative values associated to distributions that are shifted to the right and left, respectively (Harlick, Shanmugam and Dinstein, 1973), from Figure 5, it appears that a -0.10 threshold can be applied to separate *Gochnatia polymorpha* samples from *Schizolobiun parahyba*, although some mixing remains. Nevertheless, using uniformity and smoothness thresholds, there is a reduction in the number of samples from those species that reaches up to 18.69% for *Gochnatia polymorpha* and 93.75% for *Schizolobiun parahyba* leading to the separation between them and the others.

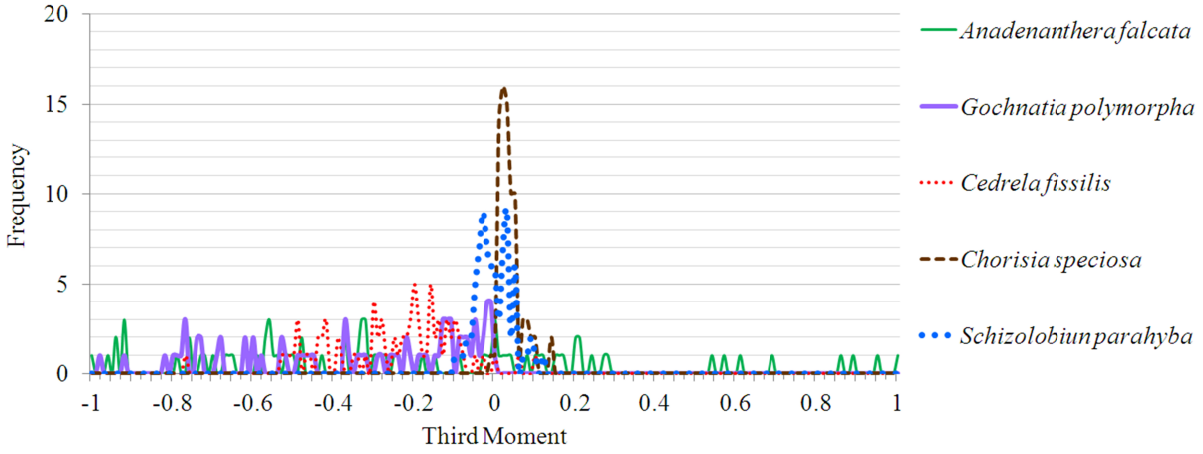



Figure 5. Histogram for third moment (asymmetry) values from the trunk images.

Considering uniformity as previous operation, for samples of *Anadenanthera falcata* and *Cedrela fissilis* a separation can be obtained with a threshold value of 0.14, for *Gochnatia polymorpha*, in the interval ]0.14; 018], entropy (shown in the Figure 6) can be applied to reduce the mixing, using threshold limits of 0.91 and 0.94.

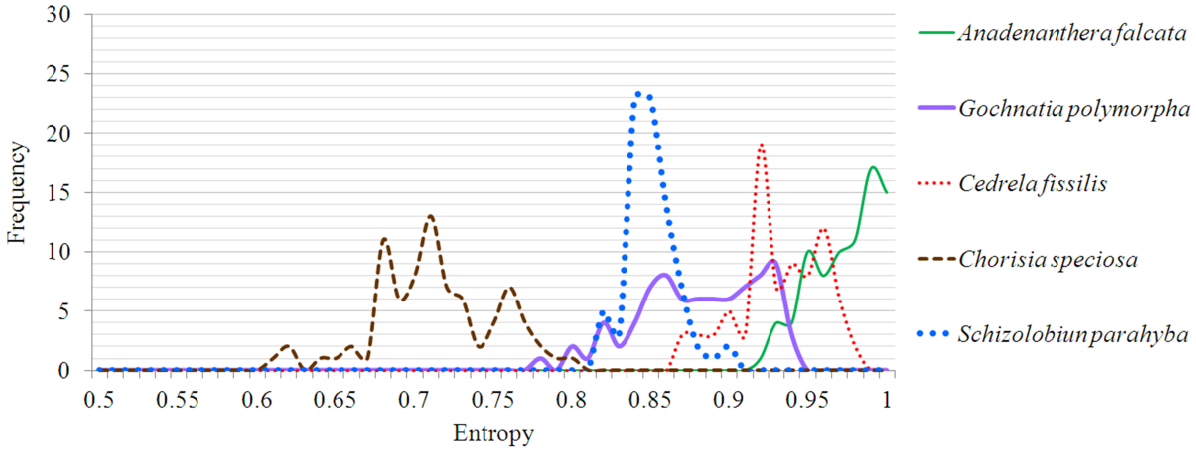



Figure 6. Histogram for entropy values from the trunk images.

Similarly, samples of *Gochnatia polymorpha* and *Cedrela fissilis can be separated with a sequence of threshold applications:* uniformity values in the range ]0.18; 0.36], 0.2 for smoothness and 0.94 for entropy. Thus, entropy is analyzed after other thresholds had been implemented.

## 3.2 Construction of the classification system

Taking in to account the study on the statistical properties obtained from the 400 images (80 for each species), ranges and thresholds for the parameters were defined in such a way to lead to the maximization of $H_{\text{rate}}$ values, resulting the configuration shown in Figure 7, where the internal and external parentheses values express $p_A^{S_j}$ and $P^{S_j}$, respectively. It is noted that the coefficient of samples $p_A$is also indicated at the output of each operation.

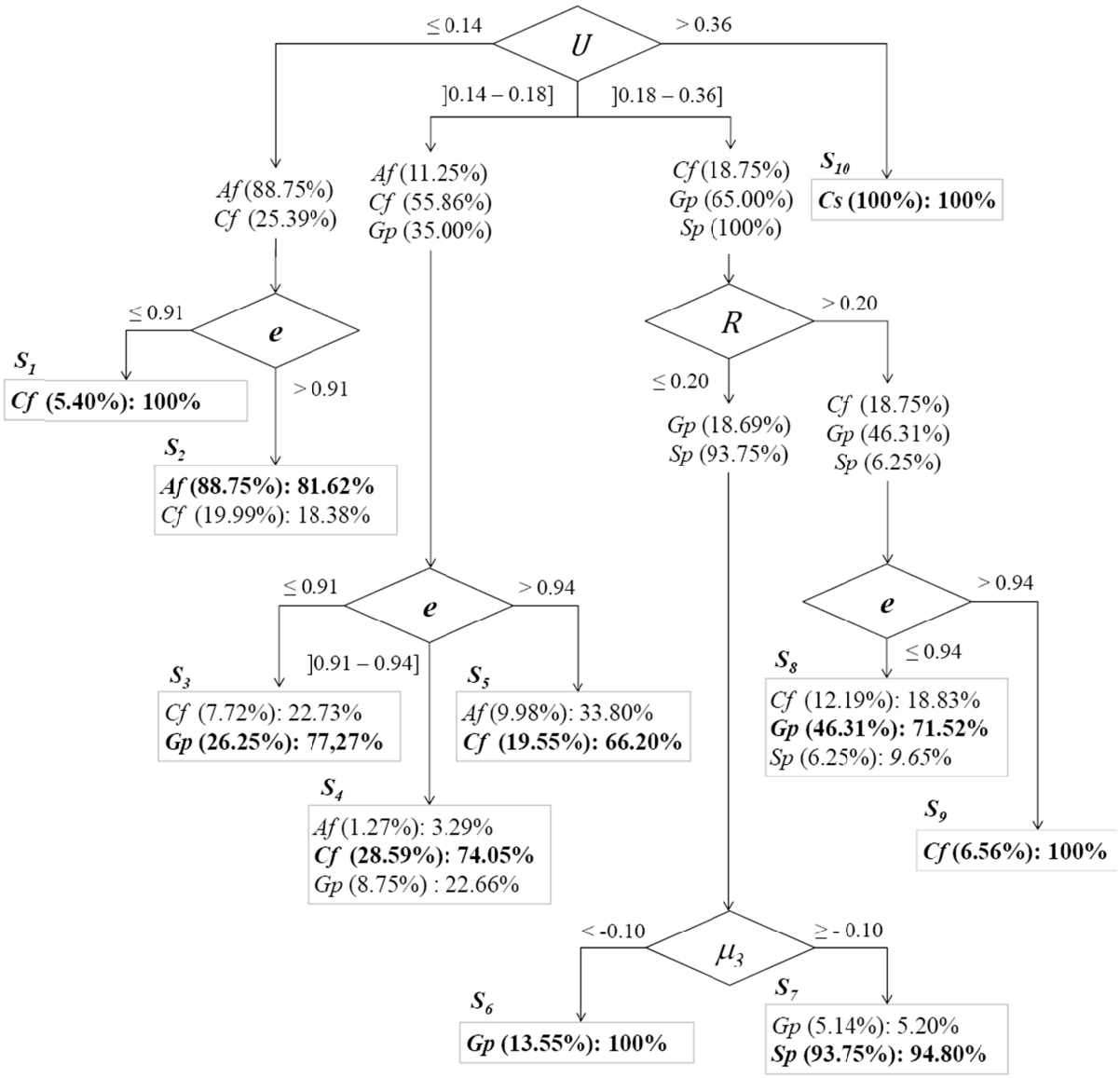



Figure 7. Representation of the identification system for *Anadenanthera falcata* (*Af*); *Cedrela fissilis* (*Cf*); *Gochnatia polymorpha* (*Gp*); *Schizolobiun parahyba* (*Sp*); *Chorisia speciosa* (*Cs*), with threshold selection for: *U - Uniformity*; *e - Entropy*; *R - Smoothness*; and $\mu_3$ - *Asymmetry*.

The proposed classification system can be considered robust since, in addition to estimating the probability that each output sample belongs to a particular class, also estimates the contribution of such class in the overall output given by $p_A^{S_j}$. As an example, it can be observed from Figure 7 that in the case of *Schizolobiun parahyba*, after the application of thresholds in uniformity, smoothness, and asymmetry values, there is a 94.80% probability of correct classification ($P^{S_7}$) in 93.75% of the total number of samples for that species ($p_A^{S_7}$).

The relative hit rate for each species and the average hit rate of the classifier are shown in Table 1, where it can be seen that the proposed method yields a quite substantial accuracy rate for *Chorisia speciosa* (100%) and *Schizolobiun parahyba* (88.87%) samples, leading to a 74.86% average rate for the classifier ($\overline{H}_{\text{rate}}$). However, samples from *Gochnatia polymorpha* and especially from *Cedrela fissilis* showed lower success rates, 66.95% and 46.07% respectively.

Table 1. Correct hit estimates for the classification system based on: $S_j$- System output; $S_D^j$- Dominant output; $p_A^{S_j}$- Sample coefficient for species; $P^{S_j}$- Probability for a sample belongs to species; $H_{rate}$ - Hit rate for each species; $\overline{H}_{rate}$ - Average hit rate.

| *Species* ($sp_i$) | $S_j$ | $S_D^j$ | $p_A^{S_j}$ | $P^{S_j}$ | $H_{rate}$ | $\overline{H}_{rate}$ |
|---|---|---|---|---|---|---|
| *Cedrela fissilis* | $S_1$ | 1 | 5.40% | 100% | 46.07% | 74.86% |
| | $S_2$ | 0 | 19.99 % | 18.38 % | | |
| | $S_3$ | 0 | 7.72 % | 22.73 % | | |
| | $S_4$ | 1 | 28.59 % | 74.05 % | | |
| | $S_5$ | 1 | 19.55 % | 66.20 % | | |
| | $S_8$ | 0 | 12.19 % | 18.83% | | |
| | $S_9$ | 1 | 6.56 % | 100% | | |
| *Anadenanthera falcata* | $S_2$ | 1 | 88.75 % | 81.62 % | 72.44% | |
| | $S_4$ | 0 | 1.27 % | 3.29 % | | |
| | $S_5$ | 0 | 9.98 % | 33.80 % | | |
| *Gochnatia polymorpha* | $S_3$ | 1 | 26.25% | 77,27% | 66.95% | |
| | $S_4$ | 0 | 8.75% | 22.66% | | |
| | $S_6$ | 1 | 13.55% | 100% | | |
| | $S_7$ | 0 | 5.14% | 5.20% | | |
| | $S_8$ | 1 | 46.31% | 71.52% | | |
| *Schizolobiun parahyba* | $S_7$ | 1 | 93.75% | 94.80% | 88.87% | |
| | $S_8$ | 0 | 6.25% | *9.65*% | | |
| *Chorisia speciosa* | $S_{10}$ | 1 | 100% | 100% | 100% | |

3.3 Validation of the classification system

As the classification system is not discrete, for performance analysis (validation) purposes using the testing samples (140 images), the class considered to be the identified species at each output was adopted to be the dominant one (see equation 11), as shown in Table 2.

Table 2. Identified class as the dominant output.

| *Output* ($S_j$) | *Identified class* |
|---|---|
| $S_1$ | *Cedrela fissilis* |
| $S_2$ | *Anadenanthera falcata* |
| $S_3$ | *Gochnatia polymorpha* |
| $S_4$ | *Cedrela fissilis* |
| $S_5$ | *Cedrela fissilis* |
| $S_6$ | *Gochnatia polymorpha* |
| $S_7$ | *Schizolobiun parahyba* |
| $S_8$ | *Gochnatia polymorpha* |
| $S_9$ | *Cedrela fissilis* |
| $S_{10}$ | *Chorisia speciosa* |

Thus, the results for the classification of the testing images are summarized and shown as a confusion matrix in Table 3, and from that, performance values were calculated and shown in Table 4.

Table 3. Confusion matrix for the testing image classification outcomes, with measures: $V_{sp_i}$- Total number of samples actually belonging to species; $I_{sp_i}$ - Total number of samples identified as belonging to species; $T_{sp_i}$- Ratio of correctly classified samples; $F_{sp_i}$- Ratio of samples miss classified.

| *True species* / *Identifying species* | *Chorisia speciosa* | *Schizolobium parahyba* | *Gochnatia polymorpha* | *Cedrela fissilis* | *Anadenanthera falcata* | $I_{sp_i}$ |
|---|---|---|---|---|---|---|
| *Chorisia speciosa* | 28 | 0 | 0 | 0 | 0 | 28 |
| *Schizolobiun parahyba* | 0 | 24 | 2 | 0 | 0 | 26 |
| *Gochnatia polymorpha* | 0 | 4 | 24 | 9 | 0 | 37 |
| *Cedrela fissilis* | 0 | 0 | 2 | 15 | 3 | 20 |
| *Anadenanthera falcata* | 0 | 0 | 0 | 4 | 25 | 29 |
| $V_{sp_i}$ | 28 | 28 | 28 | 28 | 28 | 140 |

$T_{sp_i}$ $F_{sp_i}$

Table 4. Performance assessment for the identification system, using: $P_{rate}$- Precision rate; $E_{rate}$ - Error rate; $S_{rate}$ - Sensitivity or hit rate; $\theta_1$ - Accuracy or rate of overall accuracy; $K$- Kappa or agreement index.

| *Species* | *Performance Metrics* | | | | |
|---|---|---|---|---|---|
| | $P_{rate}$ | $E_{rate}$ | $S_{rate}$ | $\theta_1$ | $K$ |
| *Chorisia speciosa* | 1.0 | 0.0 | 1.0 | 0.83 | 0.79 |
| *Schizolobiun parahyba* | 0.92 | 0.08 | 0.86 | | |
| *Gochnatia polymorpha* | 0.65 | 0.35 | 0.86 | | |
| *Cedrela fissilis* | 0.75 | 0.25 | 0.54 | | |
| *Anadenanthera falcata* | 0.86 | 0.12 | 0.89 | | |

Considering the rates for all species, it can be seen that the model had a mean precision of 0.84, meaning that from the total number of samples identified as belonging to a particular species, 84% were corrected classified, especially *Chorisia speciosa* (100%) and *Schizolobiun parahyba* (92%), in accordance with the relative rates previously calculated.

In turn, $S_{rate}$ values indicates that the classification system was able to correctly identify with substantial rates species *Anadenanthera falcata* (89%) and *Gochnatia polymorpha*(86%), maintaining full performance for *Chorisia speciosa* (100%).

However, considering the samples identified by the classifier as belonging to each species, there were significant error rates for those assigned to *Gochnatia polymorpha* (0.35) and *Cedrela fissilis* (0.25).

In general, the system had a mean score that can be considered reasonable, achieving an accuracy ($\theta_1$) of 0.83, i.e., 83% of the samples from all 5 species were correctly identified.

Considering that every classifier is likely to include samples per class belonging to others, known as inclusion error (or commission), as well as not computing samples that belonging to that class (omission error), accuracy may overestimate the system performance (Cohen, 1960) and Kappa can be considered a better indicator. However, in the present study, Kappa presented rates closed to accuracy, indicating little interference from those errors.

According to the interpretation proposed by Landis and Koch (1977), the calculated Kappa for the classification system (0.79) represents a substantial agreement between the classification provided by the model and the correct ones.

## 4 Conclusions

From the obtained results, it can be concluded that the proposed system was able to classify the studied species with performance that can be considered substantial, achieving high rates of accuracy for some classes.

However, two species showed significant errors, which indicates the complexity involved, in agreement with the related investigations in the literature that also show overall low rate accuracy .

It must be stressed that the proposed classification system is not intended to replace conventional techniques or the expertise of specialists, but to provide a support tool for forest inventories and surveys that can be applied in conjunction with other methods.

Nevertheless, to improve the classification ability of the system new features must be analyzed. For future investigations, it is indicated the use of properties that can better identify texture patterns from tree trunk images, as well as the increase of the database with the inclusion of new tree species.

It is important to notice that the use of images from trunks represents an important advance, once it overcomes the limitations of the approaches based on leaves, flowers and fruits.

## References

Backes, A. R., Casanova, D., & Bruno, O. M. (2009). Plant leaf identification based on volumetric fractal dimension. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(6), 1145-1160.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*. 22(2), 249-254.

Casanova, D., & Bruno, O. M. (2009). Plant leaf identification using Gabor wavelets. *International Journal of Imaging Systems and Technology*, 19(3), 236-243.

Casanova, D., Florindo, J. B., & Bruno, O. M. (2011). IFSC/USP at ImageCLEF 2011: Plant identification task. In: Conference and Labs of the Evaluation Forum (CLEF), 2011, Amsterdam. CLEF 2011 Evaluation Labs and Workshop: Online Working Notes, 2011.

Cohen, J. A. Coefficient of Agreement for Nominal Scales. (1960). *Educational and Measurement*. 20(1), 37-46.

Ge, S. Carruthers, R., Gong, P., & Herrera, A. Texture Analysis for Mapping *Tamarix parviflora* Using Aerial Photographs along the Cache Creek, California. (2006). *Environ Monit Assess.*, 114(1-3), 65-83.

Gonzales, R. C., & Woods, R. E. (2008). *Digital image processing*. 3ed. New Jersey: Pearson Prentice Hall.

Gonzales, R. C., Woods, R. E., & Eddins, S. L. (2009). *Digital image processing using MATLAB*. 2ed. Gatesmark Publishing.

Harlick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6), 610–621.

Jain, R., Rangachar, K., & Schunck, B. G. (1995). *Machine vision*. New York: McGraw-Hill.

Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. *Machine Learning - Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*. 30(2-3), 127-132.

Landis J. R., & Koch, G. G. (1977).The measurement of observer agreement for categorical data. *Biometrics,* 33(1), 159-174.

Machado, B. B., Casanova, D., Gonçalves, W. N., & Bruno, O. M. (2013). Partial differential equations and fractal analysis to plant leaf identification. *J. Phys.: Conf. Ser. 410 012066*.

Oliveira, P. R., & Bruno, O. M. (2009). Automatic leaf structure biometry: computer vision techniques and their applications in the plant taxonomy. *International Journal of Pattern Recognition and Artificial Intelligence,* 23(2), 247-262.

Pu, R. (2011). Mapping urban forest tree species using IKONOS imagery: preliminary results. *Environ Monit Assess.*, 172(1-4), 199-214.

Rossato, D. R., Casanova, D., Kolb, R. M., & Bruno, O. M. (2011). Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with species from Neotropical Melastomataceae (Miconieae tribe). *Plant Systematics and Evolution,* 291(1-2), 103-116.

Sá Júnior, J. J. M., Backes, A. R., Rossato, D. R., Kolb, R. M., & Bruno, O. M. (2011). Measuring and analyzing color and texture information in anatomical leaf cross sections: an approach using computer vision to aid plant species identification. *Botany,* 89(7), 467-479.

Sá Júnior, J. J. M., Rossato, D. R., Kolb, R. M., & Bruno, O. M. (2013).A computer vision approach to quantify leaf anatomical plasticity: a case study on Gochnatia polymorpha (Less.) Cabrera. *Ecological Informatics,* 15, 34-43.

Silva, N. R., Florindo, J. B., Gómez, M. C., Kolb, R. M., & Bruno, O. M. (2014). Fractal descriptors for discrimination of microscopy images of plant leaves. *J. Phys.: Conf. Ser. 490 012085*

Vibhute, A., & Bodhe, S. K. (2012). Applications of Image Processing in Agriculture: A Survey**.** *International Journal of Computer Applications,* 52(2), 34-40.

Weber, R. M., & Glenn, A. D. (2001). Riparian Vegetation Mapping and Image Processing Techniques, Hopi Indian Reservation, Arizona. *Photogrammetric Engineering and Remote Sensing.* 67(2), 179-186.

Yemshanov, D., McKenney, D. W., & Pedlar, J. H. (2012). Mapping forest composition from the Canadian National Forest Inventory and land cover classification maps. *Environ Monit Assess.*, 184(8), 55-69.

Zehm, A., Nobis, M., & Schwabe, A. (2003). Multiparameter analysis of vertical vegetation structure based on digital image processing. *Flora*, 198(2), 142-160.

## CAPÍTULO 3

# CO-OCCURRENCE PATTERNS ANALYSIS ON THE TRUNK TEXTURE AS INDICATOR FEATURES FOR COMPUTER–AIDED TREE IDENTIFICATION [b]

Adriano Bressane[1], José Arnaldo Frutuoso Roveda[2],
Sandra Regina Monteiro Masalskiene Roveda[2], Antonio Cesar Germano Martins[3]

[1] Environmental engineer, São Paulo State University (UNESP), Brazil
[2] Mathematician, University of Brasília (UnB), Brazil
[3] Physicist, University of Campinas (Unicamp), Brazil

## Abstract

The identification of arboreal species is often a hard task, fostering the development of computer-aided methods, which have focused on the tree leaf features. However, the use of leaf-based approach has limitations in the cases that leaves are not available, such as occurs for deciduous species. Therefore, the purpose of this study was to analyze co-occurrence patterns, in comparison to the first-order statistics, as indicator features for supporting the identification of tree species from the trunk texture. For that, 756 samples from 7 deciduous tree species, native from the Brazilian flora, were used in the predictive modeling procedure. As a result, the best generalization capability by using all the texture features reached 91.1% overall accuracy, in the hold-out validation with testing dataset. Moreover, despite the co-occurrence descriptors have presented better differentiation among the tree species, measures on importance of trunk texture features highlight the role of first-order statistics on the reduction of commission and omission errors, providing an average area under the ROC curve of 94.4%. In conclusion, the integrated use of co-occurrence descriptors and first-order statistics as indicator features represents a promising alternative to the advancement in the computer-aided tree recognition from trunk texture.

Keywords: pattern matching; predictive assay; environmental informatics.

---

[b] Under review in the journal ***Ecological Informatics***.

## 1 Introduction

The plant recognition is essential for several applications, such as development of medicinal products, forestry and food-producing in the agriculture, logging and forest resources management, even as urban afforestation planting and control, but particularly important for studies and research in environmental sciences, in order to support ecological purposes, as ecosystem conservation and disturbed-land reclamation.

However, the plant taxonomy can be complex, time consuming, and even impractical in the absence of fertile branches, flowers, and fruits. Thus, techniques to support the tree species identification using computational intelligence have been developed, mainly based on pattern recognition in leaves images (Silva et al., 2014; Sá Júnior et al., 2013; Kadir et al., 2011; Rossatto et al., 2011).

In spite of the advancement of leaves image-based approach, the current techniques cannot be applied for deciduous species, neither during the periods in which they lose their leaves, nor for identifying after tree felling, when its leaves and other morphological structures (buds, flowers, etc.) had already been removed.

In this context, pattern recognition in tree bark images can be an alternative for the advancement in the computer-aided identification, but still there are few outcomes reported in the scientific literature. For instance, Bressane, Roveda and Martins (2015) analyzed the bark texture in tree trunk images using first-order statistical parameters, achieving promising results, with 0.83 accuracy in the 540 samples recognition from 5 tree species.

Nevertheless, considering that those first-order statistical parameters provide texture measures without evaluating the relationship among pixels in the image (Rao et al., 2013; Srinivasan and Shobha, 2008), the use of texture descriptors based on co-occurrence matrices could improve the tree identification preciseness.

Therefore, this study aims to analyze co-occurrence descriptors on outer bark texture from trunk images, in order to evaluate its performance in the tree recognition, the prospective improvement in relation to the use of first-order statistics (comparative assessment), as well as when used in combination with them (integrated assessment).

## 2 Methods

### 2.1 Data sampling and collection

Data sampling and analyses were performed using outer bark images of 7 deciduous tree species, native from the Brazilian flora: *Anadenanthera falcata* (Af), *Cedrela fissilis* (Cf), *Chorisia speciosa* (Cs), *Gochnatia polymorpha* (Gp), *Hymenaea courbaril* (Hc), *Inga vera* (Iv), and *Schizolobiun parahyba* (Sp).

The images were obtained using a digital camera, twelve per species with 2560 x 1920 pixels, taken at different heights of the trunk, all around the trees (with 50 mm of distance from the camera to the object). Then, a central area was cut from each image, and using a moving mask (512 x 512 pixels), displaced by 256 pixels, nine samples were further obtained (Figure 1).

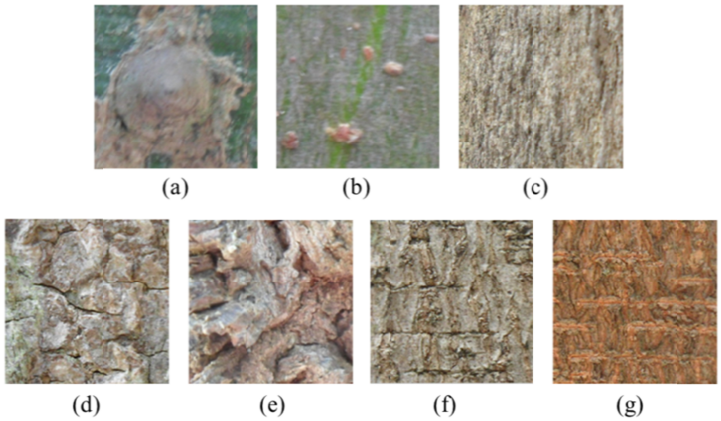



Figure 1. Outer bark images of the tree trunk from: (a) *Chorisia speciosa*, (b) *Schizolobiun parahyba*, (c) *Gochnatia polymorpha*, (d) *Cedrela fissilis*, (e) *Anadenanthera falcata*, (f) *Hymenaea courbaril*, and (g) *Inga vera*.

From the foregoing, 756 samples were generated for the analyses, 108 per species, so that 70% were used during the learning process (training dataset), and 30% for the performance assessment (testing dataset), by means of hold-out validation.

2.2 Bark texture patterns in tree trunk images

Before starting the feature extraction, image samples were transformed from the RGB (red-green-blue) system to the HSV (hue-saturation-value) space. Then, the value (V channel) was used to extract texture descriptors based on gray-level co-occurrence matrix (GLCM).

GLCM corresponds to tabulation on the frequency of intensity combinations in the image, whose calculations are performed according to the distance between pixels and different directions (Rao et al., 2013). For the analyses in the present study, the texture descriptors extracted from GLCM were contrast, correlation, energy, and homogeneity, being all values measured at directions (∅) equivalent to 0, 45, 90, and 135 degrees.

Contrast ($c$) measures the local variations, comparing the intensity of neighboring pixels over the entire image, given by:

$$c_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij} \qquad (1)$$

where $k$ is the row (or column) dimension of square co-occurrence matrix, $ij$ is an element of the GLMC, and $p_{ij}$ is an estimate of the probability that the relative position of two pixels is satisfied.

Correlation ($r$) infers the joint probability occurrence of the specified pixel pairs, considering a mean computed along rows ($m_r$) and columns ($m_c$), as in:

$$r_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{(i-m_r)(j-m_c)}{\sigma_r - \sigma_c} p_{ij} \qquad (2)$$

where $\sigma_r$ and $\sigma_c$ are respectively the standard deviation measured along rows and columns, with:

$$m_r = \sum_{i=1}^{k} iP(i) \quad \text{and} \quad m_c = \sum_{j=1}^{k} jP(j) \qquad (3)$$

$$\sigma_r^2 = \sum_{i=1}^{k}(i-m_r)^2 P(i) \quad \text{and} \quad \sigma_c^2 = \sum_{j=1}^{k}(i-m_c)^2 P(j) \qquad (4)$$

$$P(i) = \sum_{j=1}^{k} p_{ij} \quad \text{and} \quad P(j) = \sum_{i=1}^{k} p_{ij} \qquad (5)$$

The energy ($\varepsilon$) returns sum of squared elements in GLMC and homogeneity ($H$) measures the closeness of gray levels in the spatial distribution over image, which are respectively obtained as in:

$$\varepsilon_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k} {p_{ij}}^2 \quad (6)$$

$$H_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{p_{ij}}{1+|i-j|} \quad (7)$$

For comparative and integrated assessment, the first-order statistical parameters analyzed were: uniformity, entropy, asymmetry, smoothness, intensity, and standard deviation.

Based on histogram, uniformity ($U$) measures how close gray levels are in the image, as in:

$$U = \sum_{i=0}^{L-1} p^2(z_i) \quad (8)$$

where $L$ is the number of gray levels, $z_i$ is the intensity of pixel $i$, and $p(z_i)$ is the image histogram.

Entropy ($e$) as a first-order statistic measures of randomness in the image, calculated by:

$$e = -\sum_{i=0}^{L-1} p(z_i)\log_2 p(z_i) \quad (9)$$

Asymmetry ($\mu_3$) and smoothness ($R$), which takes in to account the transition of gray shades in the image, are respectively obtained by:

$$\mu_3 = \sum_{i=0}^{L-1} (z_i-\mu_1)^2 p(z_i) \quad (10)$$

$$R = 1 - \frac{1}{1+{\mu_2}^2} \quad (11)$$

where $\mu_1$ is the average intensity, and $\mu_2$ is the standard deviation, obtained by:

$$\mu_1 = \sum_{i=0}^{L-1} z_i p(z_i) \quad (12)$$

$$\mu_2 = \frac{\sum_{i=1}^{n}(z_i-\mu_1)^2}{n-1} \quad (13)$$

where $n$ is the number of pixels in the image.

## 2.3 Predictive modeling procedure

Due to the exploratory nature of this study, the prediction model interpretability is quite important for evaluating the discriminant capability of features related to the bark texture patterns. Therefore, the predictive modeling procedure was based on a single-decision tree.

Tree building process occurred from training dataset (supervised learning), which provided samples (input-output) on how the dependent variables (tree species) are related to the values of bark texture patterns (indicator features). Thus, a rules-based logical model was developed for modeling this relationship, as a binary decision tree (two-way split), using the DTREG® software.

Decision tree fitting were performed using Gini impurity ($I_G$), a measure of misclassification that evaluates the split quality, guiding selection of indicator features and boundary values with the best discriminant capacity among different species samples, i.e., that minimizes the misclassification, given by:

$$I_G(X_m) = \sum_k p_{mk}(1 - p_{mk}) \qquad (14)$$

where $X_m$ is candidate splitter for the node $m$, $p_{mk}$ is the proportion of class $k$ observations in node $m$.

To improve preciseness during the learning process, there is a progressive inclusion of indicator features that promotes the growth of the branches in the tree, to fit the model to the learning dataset with extreme accuracy. However, an overly-large tree can cause overfitting, reducing its generalization capability. Then, for avoiding overfitting problems, the tree optimal size was statistically determined by post-pruning control using 10-fold cross-validation as a performance measure over training data (backward pruning). That way, the decision tree was pruned to minimal cross-validated error, before assessment with testing dataset.

## 2.4 Recognition performance assessment

Besides the k-fold cross validation aforementioned, a hold-out validation was performed for evaluating the tree recognition performance from a testing dataset, i.e., using data that were not used for training (30 percent of all samples, selected randomly from the full dataset), in order to verify the generalization accuracy. For this purpose, the metrics measured were overall accuracy, precision, sensitivity, and specificity.

Overall accuracy ($\theta$), or hit rate of the model considering all species, was estimated by the ratio of samples correctly classified by the total number of samples ($n_T$), given by:

$$\theta = \frac{1}{n_T} \sum_{i=1}^{n_{sp}} TP_{sp_i} \qquad (15)$$

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

Precision measures the hit rate for each species ($sp_i$), estimated by ratio of true positive samples by the total number of samples identified as belonging to $sp_i$ ($I_{sp_i}$), as in:

$$Precision\ (sp_i) = \frac{TP_{sp_i}}{I_{sp_i}} = \frac{TP_{sp_i}}{TP_{sp_i} + FP_{sp_i}} \qquad (16)$$

where and $FP_{sp_i}$ is the total number of false positive samples.

Sensitivity, or true positive rate ($tp_{rate}$), measures the proportion of positives samples that are correctly identified as such, estimated by the ratio of true positive samples ($TP_{sp_i}$) over the total number of samples actually belonging to $sp_i$ ($V_{sp_i}$), as in:

$$Sensitivity\ (sp_i) = \frac{TP_{sp_i}}{V_{sp_i}} = \frac{TP_{sp_i}}{TP_{sp_i} + FN_{sp_i}} \qquad (17)$$

where $FN_{sp_i}$ is the total number of false negative samples.

Specificity measures the proportion of negatives samples that are correctly identified as such, computed by the ratio of total number identified as belonging to other species by the total number of samples actually belonging to others species, given by:

$$Specificity\ (sp_i) = \frac{TN_{sp_i}}{TN_{sp_i} + FP_{sp_i}} = 1 - fp_{rate} \qquad (18)$$

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

In addition, the receiver operating characteristic (ROC), which allows an integrated measure of false and true positive rates, was used for further comparative evaluation among models built with first-order statistical parameters and co-occurrence descriptors.

By ROC method, a two-dimensional space is formed with the dimensions $fp_{rate}$ and $tp_{rate}$ on the horizontal and vertical axes, respectively. Nevertheless, as the use of ROC graphics is limited to only two classes, in the present study was adopted the one-against-all strategy (Landgrebe and Duin, 2007). Then, for each species ($sp_i$) the area under the curve (AUC) passing through the points [(0, 0); ($fp_{rate}$, $tp_{rate}$); (1, 1)] was measured, so that the best performance corresponded to one with AUC closest to 1 (Fawcett, 2005).

## 3 Results and discussion

As can be seen in Table 1, the results of the assessment based on overall accuracy confirmed the prospective improvement in the tree recognition provided by co-occurrence descriptors, in comparison to the performance of the first-order statistics.

Table 1. Performance from decision trees (DT) based on:
first-order statistics (S), co-occurrence descriptors (C), and both (S+C).

| Recognition Model | Overall accuracy (%) | | |
|---|---|---|---|
| | Training | Validation | Testing |
| $DT_S$ | 85.90 | 77.26 | 77.23 |
| $DT_C$ | 94.17 | 88.16 | 87.50 |
| $DT_{S+C}$ | 95.11 | 88.91 | 91.07 |

Texture parameters based on co-occurrence matrices took into account the distribution and spatial relationship of the bark features in the tree trunk images. Accordingly, co-occurrence descriptors reached 94.2% accuracy in the training process, an increase of almost 10 percent over the performance of first-order statistical parameters (85.9%), even as in the evaluation with the testing dataset from 77.2 to 87.5 %.

On the other hand, even though with a similar performance in the training, the recognition model built using both first-order statistics and co-occurrence descriptors got an improvement even greater during testing, providing an overall hit rate of 91.1%. In this sense, it is further noted that the recognition model with integrated use of such texture patterns ($DT_{S+C}$) was the one that had the smallest performance reduction on the tests compared to the train. Thus, as a result of pruning-post control the decision tree showed in Figure 2 was the one that produced the best generalization capability.

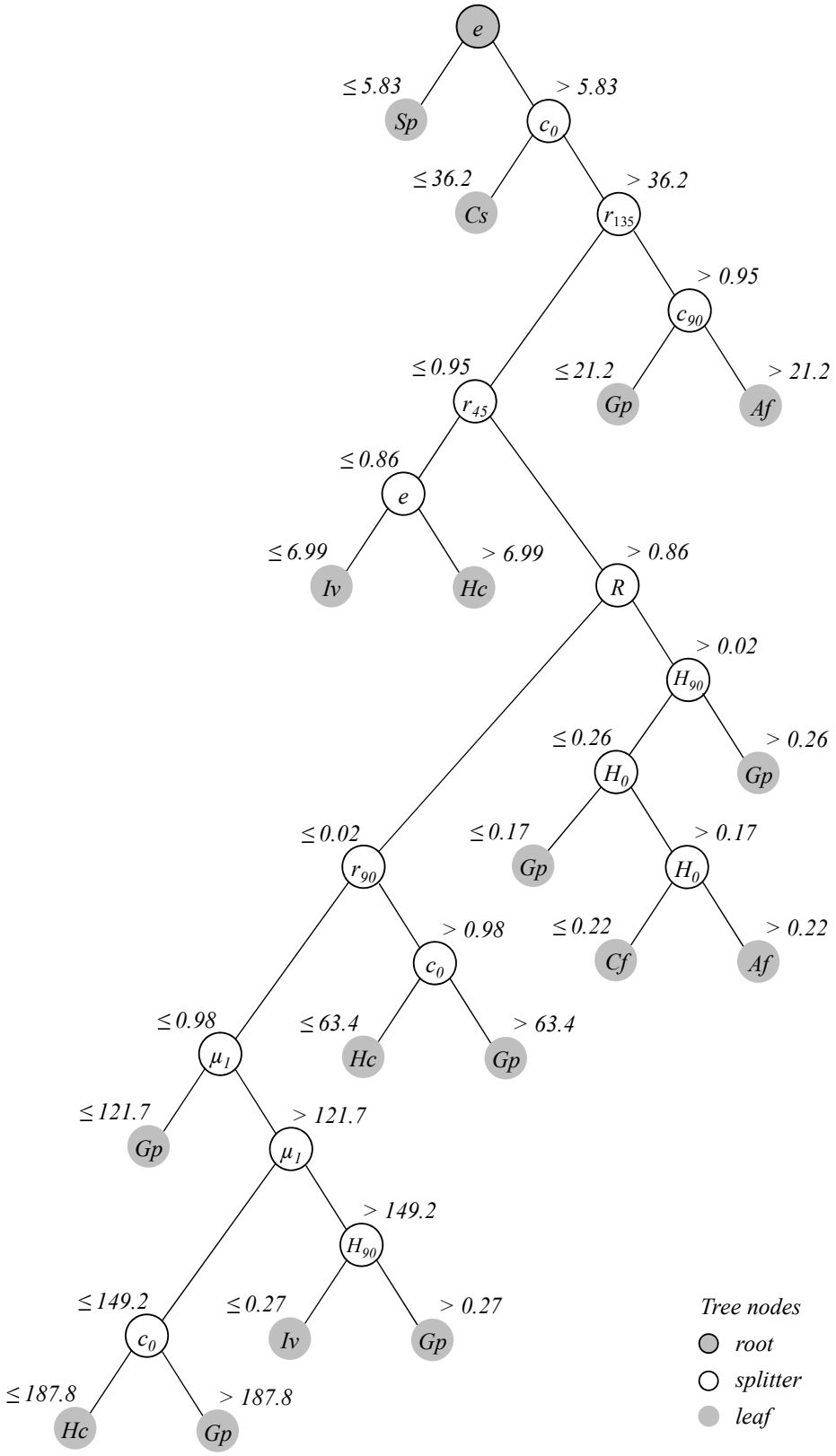



Figure 2. Decision tree built for species *Anadenanthera falcata* (*Af*), *Cedrela fissilis* (*Cf*), *Chorisia speciosa* (*Cs*), *Gochnatia polymorpha* (*Gp*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), and *Schizolobiun parahyba* (*Sp*), based on both first-order statistics and co-occurrence descriptors ($DT_{S+C}$).

As aforementioned, the ease of interpretation of the single-decision tree is one great advantage for understanding the importance of bark texture patterns as indicator features in the species identification. By analyzing Figure 2, it can be seen that the entropy (*e*) was selected as the best splitter to form the root of tree decision. As such, this first-order statistical parameter had the highest relative importance among the assessed patterns, including the co-occurrence descriptors, because on this specific position (initial node) it composes all the rules of recognition model.

On this regard, the overall importance of the main texture patterns, calculated by adding up the improvement in tree species recognition gained by each split that used the pattern as indicator feature, is showed in Table 2.

Table 2. Texture pattern importance as indicator in the $DT_{S+C}$.

| Primary splitters | | Surrogate splitters | |
|---|---|---|---|
| Feature | Importance (%) | Feature | Importance (%) |
| *e* | 100.0 | *R* | 100.0 |
| $c_0$ | 99.03 | $\mu_2$ | 100.0 |
| $r_{135}$ | 78.88 | $r_0$ | 89.57 |
| $r_{45}$ | 74.97 | *e* | 82.67 |
| *R* | 57.31 | $r_{45}$ | 76.07 |
| $H_{90}$ | 26.79 | *U* | 59.79 |
| $\mu_1$ | 26.11 | $r_{135}$ | 56.95 |
| $r_{90}$ | 11.87 | $c_{90}$ | 44.21 |
| $c_{90}$ | 11.04 | $r_{90}$ | 39.95 |
| $H_0$ | 9.22 | $c_0$ | 39.83 |

From the Table 2, it is noted that others first-order statistics also had great importance as indicator features in the recognition model. For instance, the smoothness (*R*) has got around 57% of overall importance, constituting 11 from all the 17 rules. In turn, by considering surrogate splits, the uniformity (*U*) also overcame the importance of some co-occurrence descriptors.

Regarding the others performance metrics based on testing dataset, the classification results achieved by recognition model with the greatest overall accuracy are summarized in Table 3 as a confusion matrix, and in Table 4 in which the precision, sensitivity, and specificity can be seen.

Table 3. Confusion matrix for the classification results based on testing dataset achieved by $DT_{S+C}$.

| Actual species | Predicted species | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Af* | *Cf* | *Cs* | *Gp* | *Hc* | *Iv* | *Sp* | $V_{sp_i}$ |
| *Af* | 31 | 0 | 0 | 1 | 0 | 0 | 0 | 32 |
| *Cf* | 1 | 30 | 0 | 0 | 1 | 0 | 0 | 32 |
| *Cs* | 0 | 0 | 32 | 0 | 0 | 0 | 0 | 32 |
| *Gp* | 2 | 3 | 1 | 19 | 6 | 1 | 0 | 22 |
| *Hc* | 1 | 0 | 0 | 1 | 28 | 2 | 0 | 32 |
| *Iv* | 0 | 0 | 0 | 0 | 0 | 32 | 0 | 32 |
| *Sp* | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 32 |
| $I_{sp_i}$ | 35 | 33 | 33 | 21 | 35 | 35 | 32 | 224 |

Table 4. Performance metrics based on testing dataset achieved by $DT_{S+C}$.

| Tree Species | Performance metrics (%) | | |
|---|---|---|---|
| | Precision | Sensitivity | Specificity |
| *Af* | 88.57 | 96.88 | 97.92 |
| *Cf* | 90.91 | 93.75 | 98.44 |
| *Cs* | 96.97 | 100.0 | 99.48 |
| *Gp* | 90.48 | 59.38 | 98.96 |
| *Hc* | 80.00 | 87.50 | 96.35 |
| *Iv* | 91.43 | 100.0 | 98.44 |
| *Sp* | 100.0 | 100.0 | 100.0 |

Analyzing Table 4, it is noted that the integrated use of first-order statistics and co-occurrence descriptors ($DT_{S+C}$) provided high precision, with a hit rate over 90% of samples identified as belonging to the most of the tree species, especially *Schizolobiun parahyba* (100%) and *Chorisia speciosa* (97%).

By sensitivity, it finds that there was only one tree species, *Gochnatia polymorpha* (59.4%), with low proportion of its samples correctly identified, due to the more significant omission error (positives samples incorrectly rejected). Notwithstanding, for the others evaluated species the ratio of testing positive among their samples was quite high, mostly for *Schizolobiun parahyba*, *Inga vera* and *Chorisia speciosa*, with 100% sensitivity.

In turn, the specificity results indicated that capability of the $DT_{S+C}$ to correctly reject those samples that not belonging to each species was extremely high, only the *Hymenaea courbaril* had a commission error (by computing samples belonging to others species) slightly higher, but still with a false positive rate lower than 5%.

The discriminant capacity of recognition models in supporting the identification of tree species also can be comparatively analyzed through the area under the ROC curve (Figure 3).

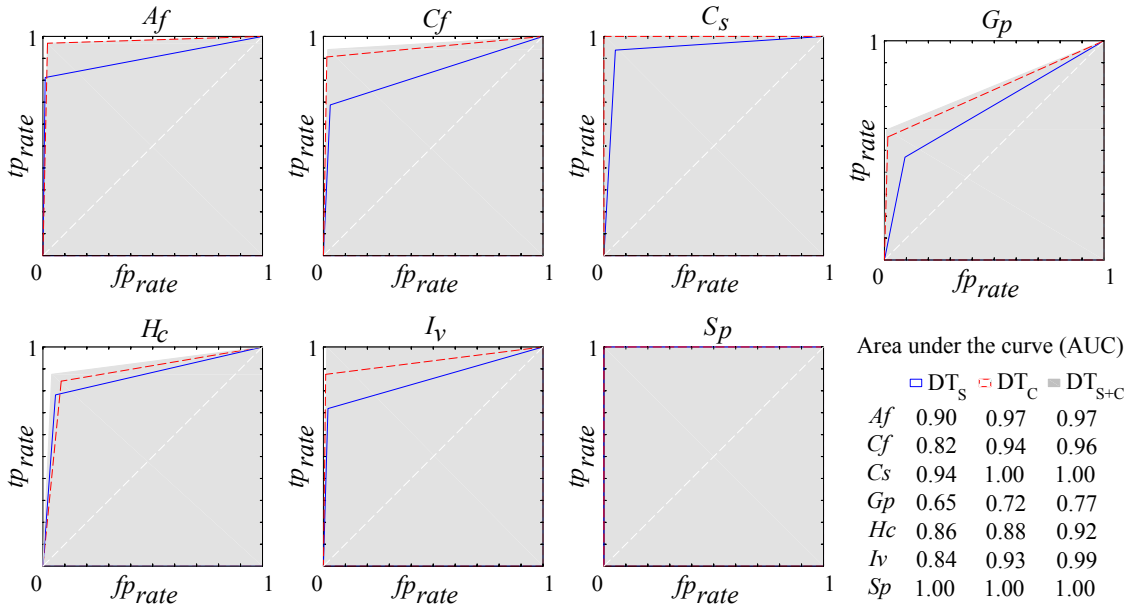



Figure 3. Area under the ROC curve for the decision trees (DT) based on statistical parameters (S), co-occurrence descriptors (G) and both (S+G), in supporting the identification of species: *Anadenanthera falcata* (*Af*), *Cedrela fissilis* (*Cf*), *Chorisia speciosa* (*Cs*), *Gochnatia polymorpha* (*Gp*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), and *Schizolobiun parahyba* (*Sp*).

Using the ROC curves allows comparing the performance of models as a measure of ability in distinguishing samples of the different tree species, taking into account in an integrated manner the commission and omission errors. By analyzing Figure 3, it is notable that the isolated use of first-order statistics ($DT_S$) obtained the lowest values, with average AUC of 0.86.

In turn, the integrated use of the first-order statistical parameters with co-occurrence descriptors ($DT_{S+C}$) always had the ROC curve closer of the upper left corner, achieving an average AUC of 0.94, an increase of 9 % over the $DT_S$. Nevertheless, when compared with the model built using only co-occurrence descriptors ($DT_C$), which obtained an average AUC of 0.92, the improvement provided by $DT_{S+C}$ was less significant.

Despite this, there was no worsening for any species by using all patterns (S+C). Contrariwise, the performance gain by $DT_{S+C}$ was substantial for some species, as *Hymenaea courbaril* (from 88 to 92%) mainly due to a lower commission error ($<fp_{rate}$), and *Inga vera* (from 93 to 99%) owing to a minor omission error ($>tp_{rate}$).

## 4 Conclusions

In this study, the use of co-occurrence descriptors in the tree recognition was analyzed in comparison to the performance of first-order statistics. From the overall accuracy assessment, the intended improvement was provided by co-occurrence descriptors, but the best performance and generalization capability was actually achieved through its integrated use with the first-order statistical parameters, overcoming the expected outcomes. The prediction model interpretability, using a single decision tree, allowed understanding the potential of the bark texture as indicator feature of tree species, highlighting the role of the first-order statistical parameters by means of a measure of the relative importance of assessed patterns. In addition, by analyzing the ROC curves was further highlighted such importance on the reduction of commission and omission errors from this alternative.

Although the co-occurrence descriptors have presented better capacity than first-order statistics to distinguish samples of different tree species, the integrated use of both texture properties is an advantageous alternative to the advancement in the computer-aided tree recognition from bark texture.

Thereby, the texture pattern recognition in the tree bark has been evaluated as a promising alternative. On the other hand, this approach requires advances as such the analysis of more features. There are still issues to overcome, such as reducing the confusion among tree species with greater overlap in the distributions of its sample values. Thus, in future studies, techniques able to transform the coordinates space of original variables, providing uncorrelated metrics and with the greatest variance, could be experienced.

## References

Bressane, A., Roveda, J.A.F., 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. doi:10.1007/s10661-015-4400-2

Fawcett, T. 2005. An introduction to ROC analysis. *Pattern Recognition Letters* 1, 861-874. doi:10.1016/j.patrec.2005.10.010

Kadir, A., Nugroho, L.E., Susanto, A., Santosa, P.I. 2011. Leaf classification using shape, color, and texture. *International Journal of Computer Trends and Technology* 2, 225-230.

Landgrebe, C.W.T., Duin, R.P.W. 2007. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognition Letters* 28, 1747-1758. doi:10.1016/j.patrec.2007.05.001

Machado, B.B., Casanova, D., Gonçalves, W.N., Bruno, O.M. 2013. Partial differential equations and fractal analysis to plant leaf identification. *Journal of Physics* 410, 1-4. doi:10.1088/1742-6596/410/1/012066

Rao, C.N., Sastry, S.S., Mallika, K., Tiong, H.S., Mahalakshmi, K.B. 2013. Co-occurrence matrix and its statistical features as an approach for identification of phase transitions of mesogens. *International Journal of Innovative Research in Science, Engineering and Technology* 2, 4531-4538.

Rossatto, D.R., Casanova, D., Kolb, R.M., Bruno, O.M. 2011. Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with species from Neotropical Melastomataceae. *Plant Systematics and Evolution* 291, 103-116. doi:10.1007/s00606-010-0366-2

Sá Júnior, J.J.M., Rossato, D.R., Kolb, R.M., Bruno, O.M. 2013. A computer vision approach to quantify leaf anatomical plasticity: a case study on Gochnatia polymorpha (Less.). *Ecological Informatics* 15, 34-43. doi:10.1016/j.ecoinf.2013.02.007

Silva, N.R., Florindo, J.B., Gómez, M.C., Kolb, R.M., Bruno, O.M. 2014. Fractal descriptors for discrimination of microscopy images of plant leaves. *Journal of Physics* 490, 1-4. doi:10.1088/1742-6596/490/1/012085

Srinivasan, G., Shobha, G. Statistical texture analysis. 2008. *Proceedings of world academy of science, engineering and technology* 36, 1264-1269.

## CAPÍTULO 4

# MULTIVARIATE ANALYSES OF TRUNK TEXTURE PATTERNS FOR SUPPORTING TREE SPECIES IDENTIFICATION USING COMPUTATIONAL INTELLIGENCE [c]

Adriano Bressane[1], Felipe Hashimoto Fengler[1], Sandra Regina Monteiro Masalskiene Roveda[2], José Arnaldo Frutuoso Roveda[2], Antonio Cesar Germano Martins[3]

[1] Environmental engineer, São Paulo State University (UNESP), Brazil

[2] Mathematician, University of Brasília (UnB), Brazil

[3] Physicist, University of Campinas (Unicamp), Brazil

## Abstract

The texture patterns recognition in the tree trunk has been evaluated as an alternative to support species identification. However, to deal with the variability within species and the similarity between some of them, the growing demand for extracting more patterns requires an approach able to treat redundant information, owing to the possibility of these new patterns are correlated. Therefore, the present study aims to evaluate the use of multivariate analyses for improving the performance of trunk texture patterns as tree species identification features. For the experimental procedures, 1188 samples were obtained from 11 arboreal species, taken at 50 mm of distance, in different heights of the trunk, all around the trees. By processing on gray-level digital images, 70 texture patterns were extracted based on first and second order statistics. Then, synthetic variables were obtained by transformations of the original measured variables, and used as input in a predictive modeling process. As a result, the multivariate analyses 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, the use of the multivariate analyses 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.

Keywords: predictive modeling; pattern matching; machine learning; computational ecology.

---

[c] Under review in the journal ***Environmental Monitoring and Assessment***.

## 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, Roveda and Martins, 2015; Yanikoglu, Aptoula and Tirkaz, 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 or of its latent 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 Smilde, 2014; Abdi and Williams, 2010; Jolliffe, 2002).

Hence, it is considered that the use of the synthetic variables as indicators could provide better results than the original variables. Therefore, the present study aims to evaluate the use of multivariate analyses 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 analyses were performed using outer bark images of 11 deciduous tree species, native from the Brazilian flora: *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*). 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 (Figure 1).

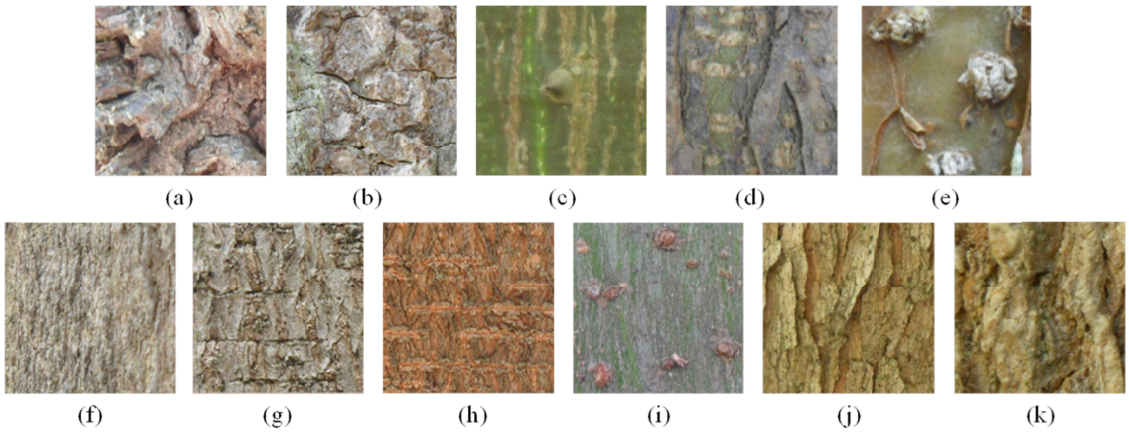



Figure 1. Outer bark images (512 x 512 pixels) of the tree trunk from: (a) *Anadenanthera falcata,* (b) *Cedrela fissilis,* (c) *Ceiba speciosa,* (d) *Centrolobium tomentosum*, (e) *Erythrina speciosa*, (f) *Gochnatia polymorpha,* (g) *Hymenaea courbaril,* (h) *Inga vera*, (i) *Schizolobiun parahyba*, (j) *Tibouchina granulosa*, and (k) *Zanthoxylum kleinii* (*Zk*).

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

2.2 Original variables extraction based on trunk texture patterns

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 (Table 1).

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

| Original variables ($z_i$) | | Texture patterns description | |
|---|---|---|---|
| | | Feature | Function |
| First-order statistics | Uniformity | measures how close gray levels are in the image | $u = \sum_{i=0}^{L-1} p^2(\varphi_i)$ |
| | Entropy | as a first-order statistic, measures the randomness in the image | $e = -\sum_{i=0}^{L-1} p(\varphi_i) \log_2 p(\varphi_i)$ |
| | Smoothness | measures the transition of gray shades in the image | $s = 1 - (1 + {\mu_2}^2)^{-1}$ |
| | Intensity | measures the average gray level in the entire image | $\mu_1 = \sum_{i=0}^{L-1} \varphi_i p(\varphi_i)$ |
| | Standard deviation | returns a measure of standard deviation in the entire image | $\mu_2 = \sum_{i=1}^{n} (\varphi_i - \mu_1)^2 (n-1)^{-1}$ |
| | Skewness | returns the measure of the image asymmetry | $\mu_3 = \sum_{i=0}^{L-1} (\varphi_i - \mu_1)^2 p(\varphi_i)$ |
| Second-order statistics | Contrast | compares the intensity of neighboring pixels | $c_\emptyset = \sum_{i=1}^{\delta} \sum_{j=1}^{\delta} (i-j)^2 p_{ij}$ |
| | Correlation | infers the joint probability occurrence of specified pixel pairs | $r_\emptyset = \sum_{i=1}^{\delta} \sum_{j=1}^{\delta} (i-m_r)(j-m_c)(\sigma_r - \sigma_c)^{-1} p_{ij}$ |
| | Energy | returns sum of squared elements in gray-level co-occurrence matrices | $\varepsilon_\emptyset = \sum_{i=1}^{\delta} \sum_{j=1}^{\delta} {p_{ij}}^2$ |
| | Homogeneity | measures the closeness of gray levels in the spatial distribution | $h_\emptyset = \sum_{i=1}^{\delta} \sum_{j=1}^{\delta} p_{ij} (1 + \lvert i-j \rvert)^{-1}$ |

In the second-order statistics extraction, the values of each of the four parameters were measured at 16 relative positions (∅), 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 generation from multivariate analyses

From the multivariate analysis based on PCA, the synthetic variables ($z_i'$) called principal components ($P_C$) 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 ($\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_p$), by solving the characteristic equation of the correlation matrix (R), given by (Bro and Smilde, 2014):

$$det(R - \lambda I) = 0,$$

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

$$z_i' = w_{i1}z_1 + w_{i2}z_2 + \ldots + w_{ip}z_p, \quad (i = 1, \ldots, p'),$$

where $p$ is the number of original variables.

In addition, synthetic variables based on oblique components ($O_C$) was also extracted by means of rotation after PCA, using *oblimin* method ($\tau$ equal to 0) by allowing orthogonal dimensions (when existing) and at the same time does not require independent dimensions.

In the FDA the $w_i$ is also known as weight vector (or weighting coefficients) of the discriminant functions ($D_F$), similarly considered as synthetic variables ($z_i'$), but with $p$ limited a condition ($p'$) as in (Russell et al., 2000):

$$p' = \min(g - 1, p),$$

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 (Figure 2), finding the eigenvector ($\boldsymbol{w}$) associated with the largest eigenvalue ($\boldsymbol{\lambda}$) that maximizes the Fischer's objective function ($F$), given by:

$$F(\boldsymbol{w}) = \boldsymbol{w^T} S_B \boldsymbol{w}\, (\boldsymbol{w^T} S_W \boldsymbol{w})^{-\mathbf{1}}, \quad S_T = S_B + S_W.$$

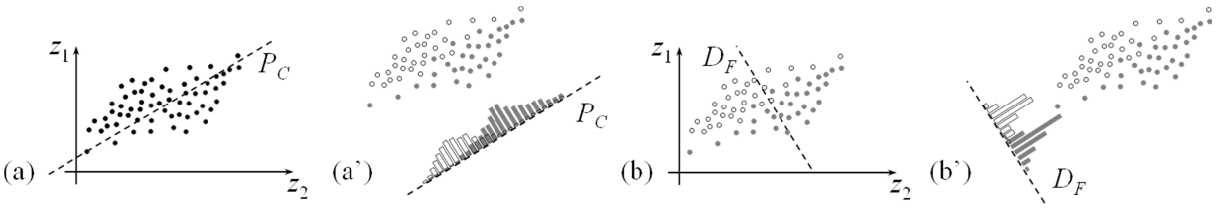



Figure 2. Synthetic variables from linear combination of the original variables ($z_1$ and $z_2$), correspondent to principal components - $P_C$ (a) and discriminant functions - $D_F$ (b), even as their directions with the largest total scatter ($S_T$) projected by PCA (a'), and maximum $F(\boldsymbol{w})$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

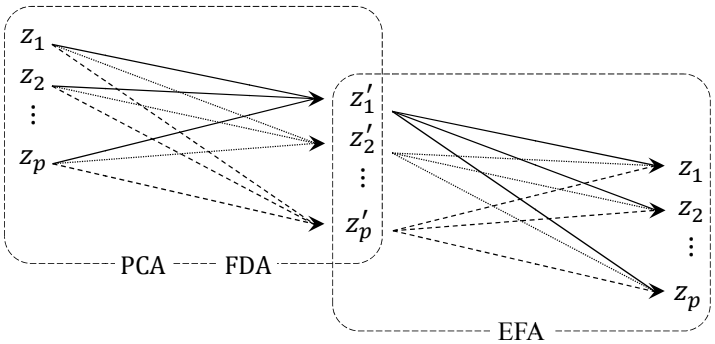



Figure 3. Causal relationships between synthetic variables ($z_i'$) and original ones ($z_i$) in Principal Component Analysis (PCA), Fischer Discriminant Analysis (FDA), and Exploratory Factor Analysis (EFA).

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 ($P_F$), and the communalities ($h_i$) were measured by squared multiple correlations, in order to find eigenvector ($\boldsymbol{w}$) associated with the largest eigenvalue ($\boldsymbol{\lambda}$) that maximizes the total communality, given by:

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

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

$$h_i = \sum_{j=1}^{m} l^2{}_{ij},$$

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 analyses the dataset ($x$) was standardized, converting all texture patterns to a common scale with an average ($\bar{x}$) of zero and standard deviation ($\sigma$) of one, given as in:

$$z = (x - \bar{x})\sigma^{-1}.$$

Furthermore, as these multivariate analyses 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 analyses 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 analyses 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 $P_C$, $O_C$, $D_F$ 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 analyses. Therefore, the predictive modeling procedure was based on a $k$-Nearest Neighbor ($k$-NN) classifier, once it is quite sensitive to features relevance (Lovrek, Howlett and Jain, 2008; Ramirez and Puiggros, 2007; Bao, Du and Ishii, 2002).

The $k$-NN is a non-linear and non-parametric supervised machine learning method, requiring for the training process a learning dataset ($L$) composed by pre-classified samples ($l_i$) in their respective arboreal species (A):

$$L = \{(l_1, sp(l_1)), \ldots, (l_N, sp(l_N))\},$$

where $f(l_i)$ denotes the class (or arboreal species) of the learning sample $l_i$, so that the $f \in A = (\alpha_1, \ldots, \alpha_{n_{sp}})$, 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:

$$f(t_q) = \text{argmax}_{\alpha \in A}\left(\sum_{i=1}^{k} \delta(\alpha, f(l_i))\right),$$

where $\delta(\alpha, f(l_i))$ is equal to 1 if $\alpha$ 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 ($d_{\mathrm{E}}$), given by:

$$d_{\mathrm{E}}(x_i, x_j) = \sqrt{\sum_{f=1}^{n}(x_i - x_j)^2}\,.$$

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 dataset, 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 dataset 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 ($\theta$) measures the ratio of samples correctly classified by the total number of samples ($n_T$), given by:

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

where $TP_{sp_i}$ 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 ($sp_i$), take into account the total number of samples identified as belonging to $sp_i$ ($I_{sp_i}$), as in:

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

where and $FP_{sp_i}$ 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 $sp_i$ ($V_{sp_i}$), as in:

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

where $FN_{sp_i}$ is the total number of false negative 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:

$$tn_{rate}(sp_i) = 1 - fp_{rate} = TN_{sp_i}\left(TN_{sp_i} + FP_{sp_i}\right)^{-1},$$

where $TN_{sp_i}$ is total number of true negative 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 accuracies.

## 3 Results and discussion

By analyzing the Kaiser-Meyer-Olkin (KMO) 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 analyses.

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.

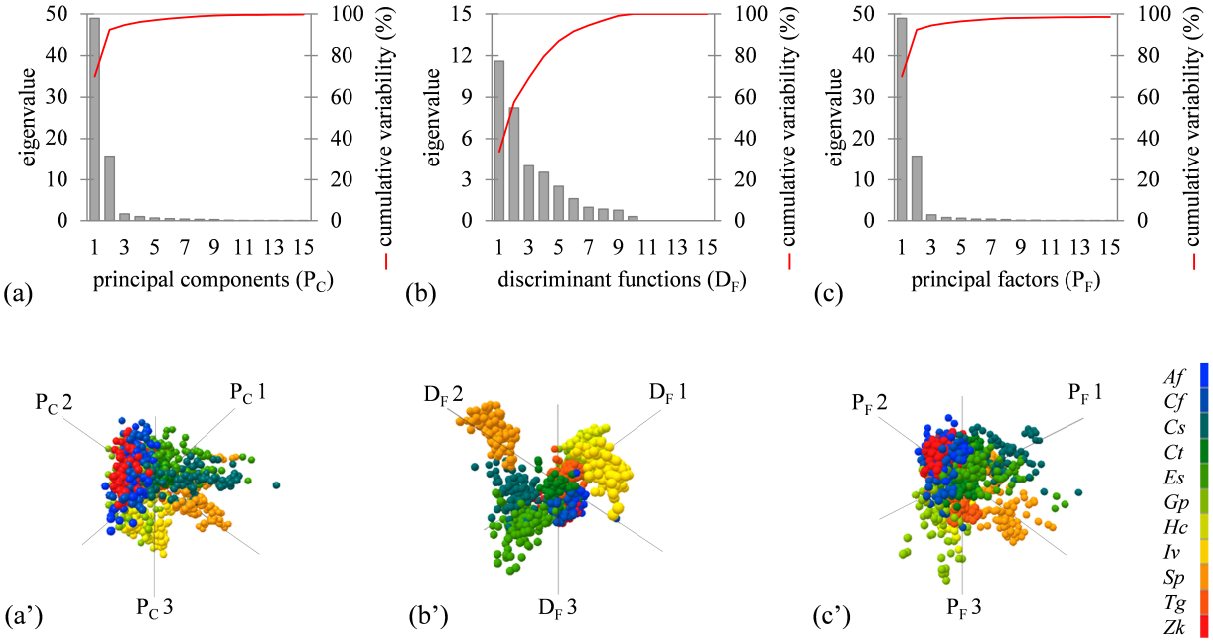



Figure 4. Cumulative variability explained by synthetic variables produced by Principal Component Analysis (a), Fischer Discriminant Analysis (b), and Exploratory Factor Analysis (c), even as the respective projections from the three first principal components (a'), discriminant functions (b'), and principal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

| Dataset | | Training (error) v-fold cross validation | | | | Testing (accuracy) hold-out validation (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k$-Nearest Neighbor ($k$-NN) | | 1-NN | 3-NN | 5-NN | 7-NN | 1-NN | 3-NN | 5-NN | 7-NN |
| Predictors | 70 original variables ($z_i$) | 0.19 | 0.23 | 0.25 | 0.28 | 91.8 | 90.1 | 90.1 | 89.2 |
| | 26 principal components ($P_C$) | 0.20 | 0.23 | 0.25 | 0.29 | 91.8 | 89.9 | 89.5 | 89.2 |
| | 23 oblique components ($O_C$) | 0.10 | 0.15 | 0.19 | 0.23 | 98.0 | 96.6 | 94.5 | 95.5 |
| | 10 discriminant functions ($D_F$) | 0.07 | 0.07 | 0.08 | 0.08 | 98.3 | 96.9 | 96.6 | 95.5 |
| | 29 principal factors ($P_F$) | 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 analyses, 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 dataset, 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 analyses 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 interpretability. Nevertheless, from such results, it is noted that the PCA-rotated data can also provide a better performance in classification tasks.

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 $P_F$, 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 $D_F$ 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.

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

| Predictors and performance metrics (%) | | Arboreal species | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Af* | *Gp* | *Cf* | *Sp* | *Cs* | *Hc* | *Iv* | *Es* | *Ct* | *Tg* | *Zk* |
| Oblique components ($O_C$) | *P* | 100 | 100 | 93.5 | 97 | 100 | 100 | 100 | 100 | 86.5 | 93.9 | 93.9 |
| | $tp_{rate}$ | 100 | 93.8 | 90.6 | 100 | 90.6 | 100 | 100 | 93.8 | 100 | 96.9 | 96.9 |
| | $tn_{rate}$ | 100 | 100 | 99.2 | 99.6 | 100 | 100 | 100 | 100 | 98.0 | 99.2 | 99.2 |
| | AUC | 100 | 97.0 | 95.0 | 100 | 95.5 | 100 | 100 | 97.0 | 99.0 | 98.0 | 98.0 |
| Discriminant functions ($D_F$) | *P* | 100 | 100 | 91.2 | 100 | 100 | 100 | 100 | 93.5 | 91.4 | 96.8 | 93.9 |
| | $tp_{rate}$ | 96.9 | 93.8 | 96.9 | 100 | 96.9 | 100 | 100 | 90.6 | 100 | 93.8 | 96.9 |
| | $tn_{rate}$ | 100 | 100 | 93.8 | 100 | 100 | 100 | 100 | 99.2 | 99.8 | 99.6 | 99.2 |
| | AUC | 98.5 | 97.0 | 95.5 | 100 | 98.5 | 100 | 100 | 95.0 | 100 | 97.0 | 98.0 |
| Principal factors ($P_F$) | *P* | 100 | 100 | 93.1 | 100 | 100 | 100 | 100 | 100 | 90.3 | 91.4 | 88.9 |
| | $tp_{rate}$ | 100 | 96.9 | 84.4 | 100 | 100 | 100 | 100 | 93.8 | 87.5 | 100 | 100 |
| | $tn_{rate}$ | 100 | 100 | 99.2 | 100 | 100 | 100 | 100 | 100 | 98.8 | 98.8 | 98.4 |
| | AUC | 100 | 98.5 | 91.5 | 100 | 100 | 100 | 100 | 97.0 | 93.5 | 99.5 | 99.0 |

By evaluating the area under the ROC curve (AUC), it is noted that the $P_F$ achieved a perfect performance (100%) for five tree species, while the $O_C$ and $D_F$ for only four ones. Nevertheless, the $P_F$ 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 Conclusions

By reviewing previous studies it was noted that the use of multivariate analyses represent a lack in the study of the trunk texture as indicator of the tree species. Then, the use of variables synthesized from multivariate analyses 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 analyses 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.

## References

Abdi, H., Williams, L.J. 2010. Principal component analysis. *Computational statistics*. 2(4), 433-459.

Backes, A.R., Bruno, O.M. 2010. Plant leaf identification using color and multi-scale fractal dimension. *Lecture notes on computer science*. 6134: 463-470.

Backes, A.R., Casanova, D., Bruno, O.M. 2009. Plant leaf identification based on volumetric fractal dimension. International. *Journal of pattern recognition and artificial intelligence*.

23(6): 1145-1160.

Bao, Y., Du, X., Ishii, N. 2002. Combining feature selection with feature weighting for k-NN classifier. A machine learning approach to detecting instantaneous cognitive states. In: Hujun Yin, Nigel Allinson, Richard Freeman, John Keane, Simon Hubbard. (eds.). *Intelligent data engineering and automated learning*. Lecture Notes in Computer Science. Springer: Manchester.

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

Boman, J. 2013. Tree species classification using terrestrial photogrammetry. Umeå: Umeå University.

Bressane, A., Roveda, J.A.F., 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., Smilde, A.K. 2014. Principal component analysis. *Analytical methods*. 6(9): 2812-2831.

Casanova, D., Bruno, O.M. 2009. Plant leaf identification using Gabor wavelets. *International journal of imaging systems and technology*, 19(3): 236-243.

Chaki, J., Parekh, R. 2011. Plant leaf recognition using shape based features and neural network classifiers. *International journal of advanced computer science and applications*, 2(10): 41-47.

Chi, Z., Houqiang, L., Chao, W. 2003. Plant species recognition based on bark patterns using novel Gabor filter banks. In: *Proceedings of the 2003 International conference on neural networks and signal processing*.

Crisci, C., Ghattas, B., Perera, G. 2012. A review of supervised machine learning algorithms and their applications to ecological data. *Ecological modelling*, 240: 113–122.

Du, J.X., Wang, X.F., Zhang, G.J. 2007. Leaf shape based plant species recognition. *Applied mathematics and computation*, 185: 883-893.

Fawcett, T. 2005. An introduction to ROC analysis. *Pattern Recognition Letters*. 1: 861-874.

Fiel, S., Sablatnig, R. 2011. Automated identification of tree species from images of the bark, leaves and needles. In: *Proceedings of the 2011 Computer vision winter workshop*.

Gu, X., Du, J.X., Wang, X.F. 2005. Leaf recognition based on the combination of wavelet transform and gaussian interpolation. *Lecture notes in computer science*. 3644: 253-262.

Huang, Z.K., Huang, D.S., Du, J., Quan, Z.H., Guo, S.B. 2006. Bark classification based on textural features using Artificial Neural Networks. *Lecture notes in computer science*, 3972: 355-360.

Huang, Z.K. 2006. Bark classification using RBPNN based on both color and texture feature. *International journal of computer science and network security*. 6(10):100-103.

Im, C., Nishida, H., Kunii, T.L. 1998. Recognizing plant species by leaf shapes-a case study of the Acer family. *Pattern recognition*, 2:1171-1173.

Jolliffe, I.T. 2002. *Principal Component Analysis*. New York: Springer.

Kadir, A., Nugroho, L.E., Susanto, A., Santosa, P.I. 2011. Leaf classification using shape, color, and texture. *International journal of computer trends & information technology*. 2: 225-230.

Kim, S.J., Kim, B.W., Kim, D.P. 2011. Tree recognition for landscape using by combination of features of its leaf, flower and bark. In: *Proceedings of the 2011 Society of Instrument and Control Engineers Annual Conference*.

Landgrebe, C.W.T., Duin, R.P.W. 2007. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognition Letters*. 28: 1747-1758.

Lee, C.L., Chen, S.Y. 2006. Classification of leaf images. *International journal of imaging systems and technology*. 16(1): 15-23.

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

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

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

Nam, Y., Hwang, E.J., Kim, D.Y. 2008. A similarity-based leaf image retrieval scheme: joining shape and venation features. *Computer vision image understanding*. 110:245-259.

Plotze, R.O., Bruno, O.M. 2009. Automatic leaf structure biometry: computer vision techniques and their applications in the plant taxonomy. *International journal of pattern recognition and artificial intelligence*. 23(2):247-262.

Plotze, R.O., Falvo, M., Pádua, J.G., Bernacci, L.C., Vieira, M.L.C., Oliveira, G.C.X., Bruno, O.M. 2005. Leaf shape analysis by the multiscale minkowski fractal dimension, a new morphometric method: a study in *Passiflora L*. (Passifloraceae). *Canadian journal of botany*. 83(3): 287-301.

Porebski, A., Vandenbroucke, N., Macaire, L. 2007. Iterative feature selection for color texture classification. In: *Proceedings of the 2007 IEEE International conference on image processing.*

Qi, H., Yang, J.G. 2003. Sawtooth feature extraction of leaf edge based on support vector machine. *Machine learning and cybernetics*. 5:3039-3044.

Ramirez, R., Puiggros, M. 2007. A machine learning approach to detecting instantaneous cognitive states. In: Zhou, Z.H., Li, H., Yang, Q. (eds.). *Advances in knowledge discovery and data mining*. Springer: Nanjiing.

Rossato, D.R., Casanova, D., Kolb, R.M., Bruno, O.M. 2011. Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with species from Neotropical Melastomataceae (Miconieae tribe). *Plant systematics and evolution*. 291(1-2): 103-116.

Russell, E.L., Chiang, L.H., Braatz, R.D. 2000. Fisher Discriminant Analysis. In: Evan L. Russell PhD, Leo H. Chiang MS, Richard D. Braatz. (eds.). *Data-driven methods for fault detection and diagnosis in chemical processes*. Springer: London.

Sá Júnior, J.J.M., Rossato, D.R., Kolb, R.M., Bruno, O.M. 2013. A computer vision approach to quantify leaf anatomical plasticity: a case study on *Gochnatia polymorpha* (Less.). *Ecological Informatics*. 15:34-43.

Singh, K., Gupta, I., Gupta, S. 2010. SVM-BDT PNN and Fourier Moment Technique for classification of leaf shape. *International journal of signal processing, image processing and pattern recognition*. 3(4):67-78.

Song, J., Chi, Z., Liu, J., Fu, H. 2004. Bark classification by combining grayscale and binary texture features. In: *Proceedings of the 2004 Intelligent multimedia, video and speech processing*.

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

Wan, Y.Y., Xiang, D.J., Huang, D.S., Chi, Z., Cheung, Y., Wang, X.F., Zhang, G.J. 2004. Bark texture feature extraction based on statistical texture analysis. In: *Proceedings of the 2004 Intelligent multimedia, video and speech processing.*

Wang, X., Huang, D.S., Du, J.X., Xu, H., Heutte, L. 2008. Classification of plant leaf images with complicated background. *Applied mathematics and computation*. 205:916-926.

Wang, X.F., Du, J.X., Zhang, G.J. 2005. Recognition of leaf images based on shape features using a hypersphere classifier. *Lecture notes in computer science*. 3644:87-96.

Wang, Z., Chi, Z., Feng, D., Wang, Q. 2003. Leaf image retrieval with shape features. *Lecture notes in computer science*. 1929 (2000):477-487.

Wojtech, M., Wessels, T. 2011. *Bark: a field guide to trees of the northeast*. New England: UPNE.

Wu, S.G., Bao, F.S., Xu, E.Y., Wang, YX., Chang, Y.F., Xiang, Q.L. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. *The computing research repository*. 1:11-16.

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

# CAPÍTULO 5

## ARBOREAL IDENTIFICATION SUPPORTED BY FUZZY MODELING FOR TRUNK TEXTURE RECOGNITION [d]

Adriano Bressane[1], Felipe Hashimoto Fengler[1], Sandra Regina Monteiro Masalskiene Roveda[2], José Arnaldo Frutuoso Roveda[2], Antonio Cesar Germano Martins[3]

[1] Environmental engineer, São Paulo State University (UNESP), Brazil

[2] Mathematician, University of Brasília (UnB), Brazil

[3] Physicist, University of Campinas (Unicamp), Brazil

### Abstract

Due to the natural variability of the arboreal bark there are texture patterns in trunk images with values belonging to more than one species. Thus, the present study analyzed the usage of fuzzy modeling as an alternative to handle the uncertainty in the trunk texture recognition, in comparison with other machine learning algorithms. A total of 2160 samples, belonging to 20 tree species from the Brazilian native deciduous forest, were used in the experimental analyzes. After transforming the images from RGB to HSV, 70 texture patterns have been extracted based on first and second order statistics. Secondly, an exploratory factor analysis was performed for dealing with redundant information and optimizing the computational effort. Then, only the first dimensions with higher cumulative variability were selected as input variables in the predictive modeling. As a result, fuzzy modeling reached a generalization ability that outperformed algorithms widely used in classification tasks, besides of obtaining an almost perfect agreement with the classifier with the best accuracy in the validation tests. Therefore, the fuzzy modeling can be considered as a competitive approach, with reliable performance in arboreal trunk texture recognition.

Keywords: soft computing; image processing; pattern matching; computer vision; bioinformatics.

[d] Under review in the journal ***Trends in Applied and Computational Mathematics***.

## 1 Introduction

The usage of computational intelligence in the feature extraction and pattern recognition from biological data has been increasingly studied for supporting the arboreal identification. However, as the studies carried out have focused on the leaves image processing, its techniques are not applicable when the leaf structure is not available, as occurs with deciduous species at certain times of the year.

As an alternative, the texture recognition in tree trunk images still has few outcomes reported in the literature, in which the predictive modeling has been performed using machine learning algorithms based on *k*-Nearest Neighbors (Porebski et al., 2007; Wan et al., 2004), Artificial Neural Networks (Huang et al., 2006), Support Vector Machine (Boman, 2013; Fiel and Sablatnig, 2011; Huang, 2006), and Decision Tree (Bressane, Roveda and Martins, 2015).

By analyzing statistical properties in tree trunk images, Bressane, Roveda and Martins (2015) found that, due to the natural variability of the arboreal bark, commonly its texture patterns have some values belonging to more than one species, i.e, there is an overlap between neighboring subspaces. As a consequence, this overlapping in the pattern matching can lead to an ambiguity during predictive modeling.

In these cases, there is some uncertainty in relation to what species the sample belongs to, undermining the texture discriminant analysis by means predictor variables with a sharply defined boundary. Therefore, the present study aims to analyze the usage of fuzzy modeling as an approach to deal with the uncertainty in the trunk texture recognition, in comparison with other machine learning algorithms.

In the mid-1960s, the fuzzy set theory has been developed by Zadeh (1965) as an extension of the classical set theory to provide a mathematical treatment for complex phenomena, becoming it popular after 1980s (Zadeh, 2008; Pedrycz and Gomide, 2007). For that, the fuzzy modeling is a soft-computing method capable of processing uncertain knowledge or data.

Thus, by affording a convenient formalism for integrating different kinds of variables, by means of an user-friendly structure with transparency and interpretability, the usage of fuzzy modeling is becoming more and more common, with several applications in the environmental sciences over the years (e.g. Bressane et al., 2016; Liu and Zou, 2012; Liu et al., 2010; Lermontov et al., 2009; Ascough et al., 2008; Adriaenssens et al., 2004; Silvert, 2000).

According to Ishibuchi and Nakashima (2001), the main applications of the fuzzy modeling used to be optimization and control problems. Nevertheless, nowadays many other areas can be highlighted, such as the development of intelligent systems for supporting the decision making, data mining, signal processing, diagnosis, forecasting, regression, and classification from numerical data using pattern recognition based on the graded membership (Singh et al., 2013).

Thus, the fuzzy modeling can achieve a competitive performance when compared to other machine learning algorithms in classification tasks involving uncertainty, vagueness, partial true, which demand predictors without hard boundaries (Arunpriya and Thanamani, 2015; Riza et al., 2015).

## 2 Methods

### 2.1 Data collection and feature extraction

The data were collected using a digital camera for capturing outer bark images at different heights of the trunk, at a 50 mm distance around the trees. Due to the three-dimensional shape of arboreal trunk, only a central area was used for extracting features, in order to avoid the distortion at the image edge. Then, using a moving mask with 512 x 512 pixels, 2160 samples were obtained, being 108 of each of the 20 tree species from the Brazilian native deciduous forest, shown in Figure 1.

To reduce the influence of the environmental conditions and image acquisition settings, before starting the features extraction the images have been transformed from RGB (red-green-blue) system to HSV (hue-saturation-value) space. Features based on first and second order statistics were extracted using the V channel from the grayscale images.

The first-order statistical parameters included 6 texture features, equivalent to uniformity, entropy, skewness, smoothness, intensity, and standard deviation, described below from Gonzales and Woods (2008).

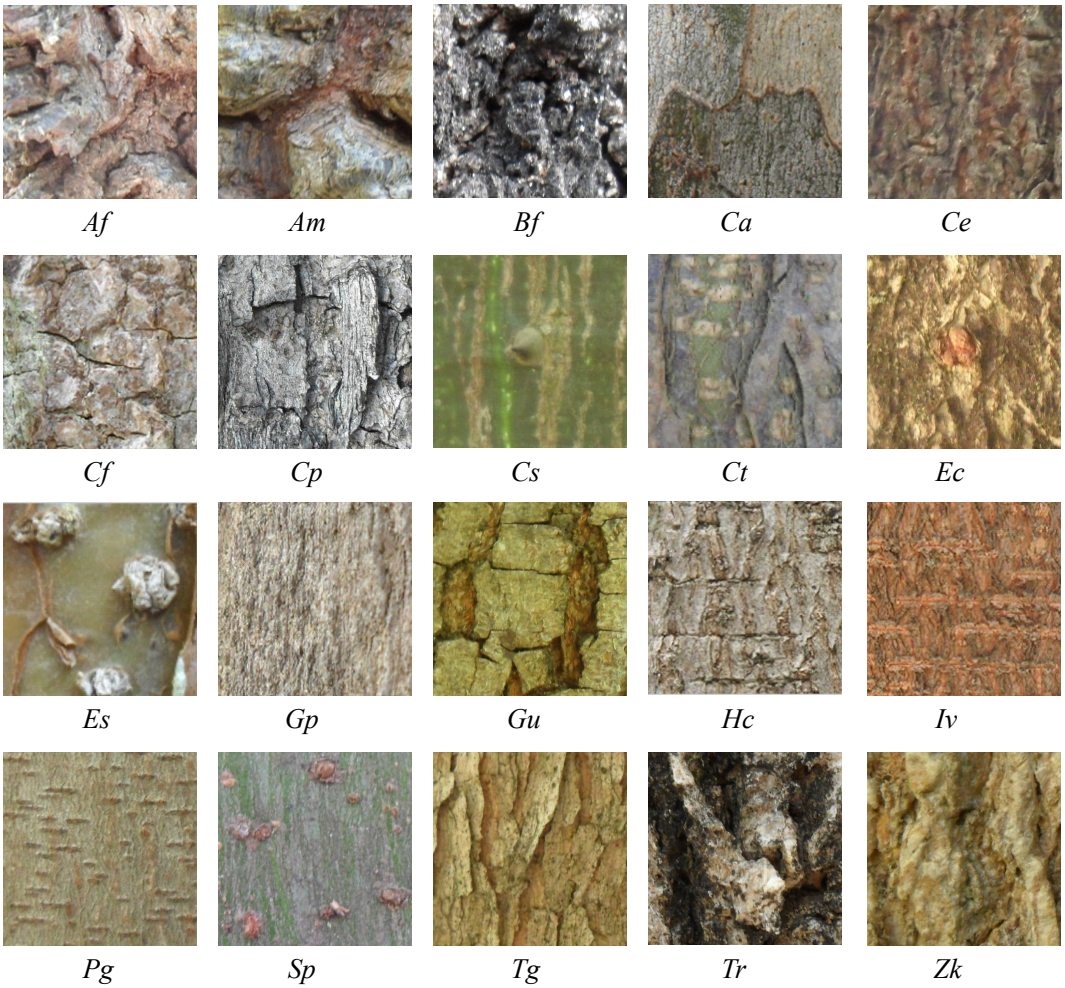



Figure 1. Tree trunk images (512x512 pixels) from: *Anadenanthera falcata* (*Af*), *Anadenanthera macrocarpa* (*Am*), *Bauhinia forficate* (*Bf*), *Caesalpinia peltophoroides* (*Ca*), *Caesalpinia echinata* (*Ce*), *Cedrela fissilis* (*Cf*), *Caesalpinia peltophoroides* (*Cp*), *Ceiba speciosa* (*Cs*), *Centrolobium tomentosum* (*Ct*), *Enterolobium contortisiliquum* (*Ec*), *Erythrina speciosa* (*Es*), *Gochnatia polymorpha* (*Gp*), *Guazuma ulmifolia* (*Gu*), *Hymenaea courbaril* (*Hc*), *Inga vera* (*Iv*), *Piptadenia gonoacantha* (*Pg*), *Schizolobiun parahyba* (*Sp*), *Tibouchina granulosa* (*Tg*), *Tabebuia roseoalba* (*Tr*), and *Zanthoxylum kleinii* (*Zk*).

As a measure of the proximity of the gray levels, the uniformity ($u$) is given by:

$$u = \sum_{i=0}^{L-1} p^2(z_i), \qquad (1)$$

where $L$ correspond to the number of gray levels in the image, $z_i$ is the intensity, and $p(z_i)$ is the image histogram.

The first-order entropy ($e$) measures the randomness in the image, as in:

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \quad (2)$$

The skewness is a measure of the asymmetry ($\mu_3$), and smoothness ($s$) takes in to account the transition of gray shades, respectively obtained by:

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - \mu_1)^2 p(z_i) \quad (3)$$

and

$$s = 1 - \frac{1}{1 + {\mu_2}^2} \quad (4)$$

where $\mu_1$ is the intensity that returns the gray level average, and $\mu_2$ is the standard deviation, calculated by:

$$\mu_1 = \sum_{i=0}^{L-1} z_i p(z_i) \quad (5)$$

and

$$\mu_2 = \frac{\sum_{i=1}^{n} (z_i - \mu_1)^2}{n - 1} \quad (6)$$

where $n$ is the number of image pixels.

In turn, the second order statistics is comprised of the contrast, correlation, energy, and homogeneity, which have been measured at 16 relative positions ($\emptyset$), correspondent to distance between pixels equal to 1, 3, 5 and 7, in the rotation angles 0, 45, 90 and 135 degrees, producing 64 texture features. These descriptors are described below from Harlick et al. (1973) and Gonzales et al. (2009).

Contrast ($c$) compares the intensity of neighboring pixels and it is computed by:

$$c_{\emptyset} = \sum_{i=1}^{k} \sum_{j=1}^{k} (i - j)^2 p_{ij} \quad (7)$$

where $k$ is the co-occurrence matrix dimension, $p_{ij}$ is probability of satisfying $\emptyset$.

The correlation ($r$) measures the probability of occurrence of specified pixel pairs, given by:

$$r_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{(i - m_{row})(j - m_{col})}{\sigma_{row} - \sigma_{col}}p_{ij} \qquad (8)$$

where $m$ is the mean and σ is the standard deviation, both calculated along rows and columns.

Energy ($\varepsilon$) adds the squared elements in the co-occurrence matrix, and homogeneity ($h$) measures the closeness of gray levels in the spatial distribution over image, respectively obtained by:

$$\varepsilon_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k}{p_{ij}}^{2} \qquad (9)$$

and

$$h_{\emptyset} = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{p_{ij}}{1 + |i - j|} \qquad (10)$$

From the foregoing, the total number of measured variables amounted to 70 texture features. So taking into account that some features may be highly correlated, an Exploratory Factor Analysis (EFA) has been performed. As a multivariate analysis technique, the EFA finds a coordinate system that maximizes the variance shared among variables, enabling to reduce the data dimensionality and prevent the use of redundant information (Costello and Osborne, 2005).

In the new $m$-dimensional space found by EFA, the standardized original variables ($z$) correspond to linear combinations of underlying factors ($z'$), given by (Yong and Pearce, 2013):

$$z_j = a_{j1}z'_1 + a_{j2}z'_2 + \ldots + a_{jm}z'_m \qquad (11)$$

where $m$ is the number of underlying factors ($z'_i$), $a_{ji}$ are the factor loadings.

For that, the EFA was carried out using the Spearman's coefficient, a non-parametric alternative regarded as robust for general distributions (non-normal data), the principal factors as extraction method, and the communalities ($h_i$) based on the squared multiple correlations, as in:

$$h_i = \sum_{j=1}^{m} l^2{}_{ij} \qquad (12)$$

where $l_{ij}$ is correlation between the $i^{th}$ principal factor with $j^{th}$ original variable (texture feature), previously standardized by means of:

$$z_i = \frac{x_i - \bar{x}}{\sigma} \qquad (13)$$

where $x_i$ is the measured original variable, $\bar{x}$ and $\sigma$ are respectively its mean and standard deviation.

Thus, the features extracted from tree trunk images have been reduced to fewer latent variables (principal factors), which were used as predictors for generating and learning fuzzy IF–THEN rules in the texture patterns recognition.

## 2.2 Fuzzy modeling for the pattern recognition

From the mid-1990s, the development of the fuzzy modeling for classification tasks is relatively recent in comparison to other applications. Notwithstanding, since then several approaches have already been proposed, including space partitioning (Chi et al., 1996), neural-network-based methods (Nauck and Kruse, 1997), clustering techniques (Abe and Thawonmas, 1997), genetic algorithms (Gonzalez and Perez, 1999), and fuzzy partition using certainty grades (Ishibuchi and Nakashima, 2001).

For the predictive modeling in the present study, we used a fuzzy rule-based classification system, created and described by Riza et al. (2015) as FRBCS.W algorithm, made available in R programming language by means of the '*frbs*' package. The FRBCS.W algorithm has been developed based on the Ishibuchi's method (Ishibuchi and Nakashima, 2001).

As aforementioned, the Ishibuchi's method is a learning method from numerical data that consists of the fuzzy partitioning with certainty grades. In its learning process, the antecedent parts of rules are determined by a grid-type fuzzy partition. That is, the partioning occurs dividing the input space of the predictor variables ($x_i$) into regular fuzzy regions, resulting in uniform and symmetrical intervals correspondent to the antecedent terms ($a_{ij}$), as can be seen in Figure 2.

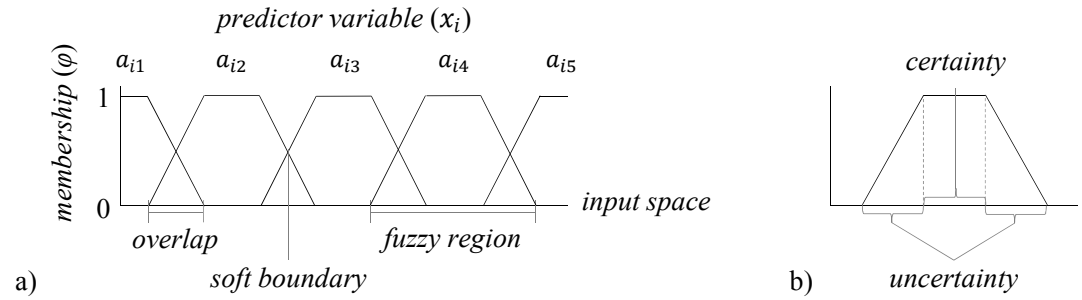



Figure 2. Grid-type fuzzy partition: (a) partitioning of the predictor variable - $x_i$ into regions correspondent to the antecedents terms - $a_{ij}$ using trapezoidal-shaped membership functions; (b) intervals of certainty and uncertainty that comprises the fuzzy region of the antecedent term.

By using the grid-type fuzzy partition, the total number of rules (*N*) is determinate by amount of possible combinations of the antecedent terms. For that, rulebase is generating by pattern matching, calculating membership degrees ($\varphi$) of the training data in the antecedents terms ($a_{ij}$) of each predictor variables ($x_i$). Thus, the consequent part is defined as the dominant categorical variable ($C_j$) in the fuzzy region corresponding to the antecedents of the rule under construction (Ishibuchi and Nakashima, 2001):

Rule $R_\mathrm{j}$: IF $x_1$ is $a_{1\mathrm{j}}$ AND … AND $x_\mathrm{n}$ is $a_{\mathrm{mj}}$ THEN $C_\mathrm{j}$ with $CF_\mathrm{j}$ , $j = 1, 2, \ldots, N$ (14)

where **x** is a *m*-dimensional vector of predictor variables ($x_i$), $a_{\mathrm{ij}}$ is a term of the antecedent part of rule, $CF_\mathrm{j}$ is the certainty grade of the rule $R_\mathrm{j}$, and $C_\mathrm{j}$ is the dominant categorical variable correspondent to output class, determinate taking into account:

$$\sum_{p \in class\ C_j} \varphi_\mathrm{j}(x_p) = \max\{\sum_{p \in \mathrm{class}\ k} \varphi_\mathrm{j}(x_p) : k = 1, 2, \ldots, c\} \quad (15)$$

where $x_p = (x_{p1}, \ldots, x_{pm})$ is a new pattern, and $c$ is the number of output classes.

After generating the predictive model, the classification of new instances is based on a single winner rule, which is determinate by the maximum product of the rule certainty grade ($CF_\mathrm{j}$) by the instance compatibility grade in the rule $R_\mathrm{j}$ ($\varphi_\mathrm{j}$), as in:

$$\varphi_\mathrm{j}(x_p).CF_\mathrm{j} = \max\{\varphi_\mathrm{j}(x_p).CF_\mathrm{j} : j = 1, 2, \ldots, N\} \quad (16)$$

where $\varphi_\mathrm{j}(x_p)$ is the instance compatibility grade given by aggregation of the membership values of its predictor variables vector in the antecedents of the rule. In turn, the rule certainty

grade ($CF_{\mathrm{j}}$) is a real number in the interval [0, 1] that works as the weight of rule, given by:

$$CF_{\mathrm{j}} = \frac{\beta_{class\ c_j}(R_{\mathrm{j}}) - \bar{\beta}}{\sum_{k=1}^{c} \beta_{class\ k}(R_{\mathrm{j}})} \tag{17}$$

where

$$\bar{\beta} = \frac{\sum_{k \neq c_j} \beta_{class\ k}(R_{\mathrm{j}})}{(c-1)} \tag{18}$$

and

$$\beta_{class\ k}(R_{\mathrm{j}}) = \sum_{x_p \in \mathrm{class}\ k} \varphi_{\mathrm{j}}(x_p)\,, k = 1, 2, \ldots, c \tag{19}$$

2.3 Benchmarking experiment

From the database with 2160 samples we used 70% randomly selected for the machine learning process. During this process a 5-fold cross-validation was carried out over learning dataset, in order to find the best control parameters setting. Then, a hold-out validation has been performed using the remaining 30% as testing dataset for assessing the generalization ability of the fuzzy classification system (FRBCS) in the trunk texture pattern recognition (Figure 3).

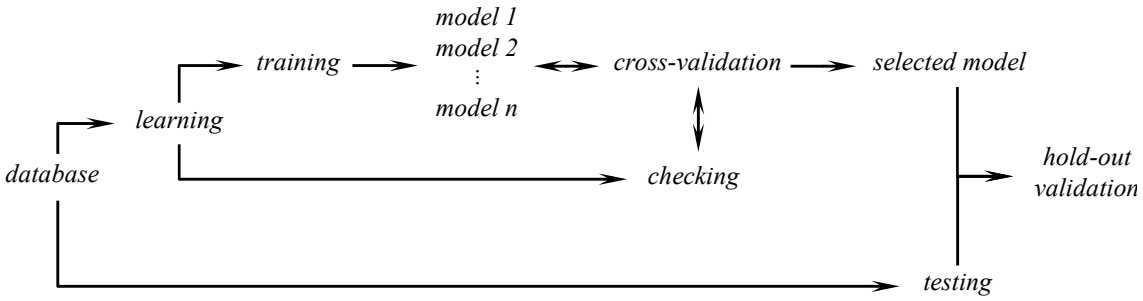



Figure 3. Split of database for the learning process and to assessing the generalization ability based on testing dataset.

Furthermore, as a reference to assess the performance from the fuzzy-based approach, a benchmarking experiment has been carried out using the same database for training, checking and testing other machine learning algorithms as shown in Table 1.

Table 1. Machine learning algorithms considered for performance comparison and control parameters settings adjusted during the learning process, which provide the best results in the cross-validation over the checking dataset.

| Learning algorithm | Main control parameters settings | Available in |
|---|---|---|
| Fuzzy Rule-Based Classification System (FRBCS) | model: frbcs.w, membership function : gaussian, t-norm: product, antecedent terms: 23 | Package *'frbs'* R language |
| Boosted Rule-Based Model (C5) | subset: false, no global pruning: false, CF: 0.25, sample: 0, trials: 100 | Package *'*C5.0*'* R language |
| Cascade-Correlation Neural Network (CNN) | kernel functions: sigmoid and gaussian, neurons: min 0, max $10^3$, candidates $10^2$, epochs $10^3$, overfitting control: prune to optimal size using cross validation over checking dataset | Algorithm *'*CNN*'* C language |
| *k*-Nearest Neighbors (KNN) | model: knn kernel, weighting: gaussian kernel, type: probability | Package *'*CORElearn*'* R language |
| Probabilistic Neural Network (PNN) | sigma: each variable, min $10^{-4}$, max 10, steps 20, kernel function: gaussian, prior probability: frequency distribution in dataset | Algorithm PNN C language |
| Multilayer Perceptron Network (MLP) | layers: 3, overfitting control: minimum holdout validated error over 10% train data, function: logistic, training: scaled conjugate gradient | Algorithm MLP-NN C language |
| Random Decision Tree Forest (Random Forest) | importance: true, proximity: true, number of trees: 300 | Package *'*randomForest*'* R language |
| Single Decision Tree (SDT) | minimum node to split: 3, maximum tree levels: 300, overfitting control: prune to minimum cross-validated error over checking dataset | Algorithm SDT C language |
| Stochastic Gradient Boosting (TreeBoost) | trees number: 300, trees depth: 8, minimum size node to split: 10, prune series to minimum error, minimum trees in series: 10 | Algorithm *'*TreeBoost*'* C language |
| Support Vector Machine (SVM) | type: bound-constraint, kernel function: gaussian, sigma: 0.1, C: 24 | Package *'*kernlab*'* R language |

Based on the testing results, the learning algorithms performance has been assessed according to the overall accuracy ($\theta$), which measures the ratio of samples correctly classified by the total number of samples ($n_T$), as in:

$$\theta = {n_T}^{-1} \sum_{i=1}^{n_{sp}} TP_{sp_i} \tag{20}$$

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

In addition, the Kappa index ($K$) has also been used to assess how well the fuzzy-based model agrees with an already established algorithm, which achieving the best performance during the experiments, by means of (Carletta, 1996):

$$K = \frac{\theta_1 - \theta_2}{1 - \theta_2} \tag{21}$$

and

$$\theta_2 = \frac{1}{n_T{}^2} \sum_{i=1}^{n_{sp}} (V_{sp_i} . I_{sp_i}) \tag{22}$$

where $\theta_2$ is the proportion of times for which an agreement is expected by chance, $I_{sp_i}$ is the total number of samples predicted as belonging to the species $i$ ($sp_i$), and $V_{sp_i}$ is the total number of samples actually belonging to $sp_i$ according to the algorithm adopted as a reference.

## 3 Results and discussion

Regarding the requirements for data pre-processing using multivariate analysis, the Cronbach's alpha equivalent to 0.9 indicated an excellent internal consistency, and the Kaiser-Meyer-Olkin equal to 0.97 confirmed a good sampling adequacy, verifying sufficient conditions to perform the Exploratory Factor Analysis (EFA), whose result is shown in Figure 4.

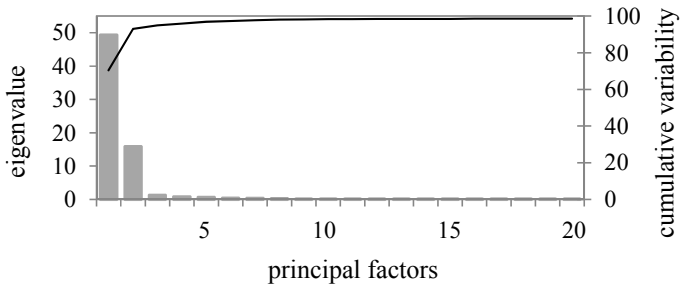



Figure 4. Eigenvalues and cumulative variability explained by the first 20 latent variables (principal factors) produced from the Exploratory Factor Analysis.

From the Figure 4, we find that the first 20 principal factors explain 99.0% of the cumulative variability. Therefore, the EFA was capable of reducing the data dimensionality and at the same time retaining almost all information available in the 70 original variables. Thus, these principal factors were used as predictor variables in the modeling process, affording the results shown in Table 2.

Table 2. Performance of the machine learning algorithms in the benchmarking experiments, based on the settings that reach the best accuracy over checking dataset during the learning process, using the first 20 principal factors as predictor variables.

| Machine learning algorithm | 5-fold Cross-validation (%) | | Hold-out validation (%) |
|---|---|---|---|
| | Training dataset | Checking dataset | Testing dataset |
| Fuzzy Rule-Based Classification System (FRBCS) | 100 | 93.5 | 94.0 |
| Boosted Rule-Based Model (C5) | 100 | 85.3 | 86.5 |
| Cascade-Correlation Neural Network (CNN) | 86.2 | 76.5 | 78.5 |
| *k*-Nearest Neighbors (KNN) | 96.3 | 89.1 | 89.7 |
| Probabilistic Neural Network (PNN) | 100 | 95.3 | 96.1 |
| Multilayer Perceptron Network (MLP) | 94.9 | 88.5 | 90.8 |
| Random Decision Tree Forest (Random Forest) | 100 | 88.9 | 89.5 |
| Single Decision Tree (SDT) | 91.6 | 72.3 | 72.3 |
| Stochastic Gradient Boosting (TreeBoost) | 100 | 85.7 | 87.3 |
| Support Vector Machine (SVM) | 100 | 95.9 | 96.2 |

In general, each machine learning algorithm has properties which can provide better performance than others, depending on the characteristics of the case under analysis. Thus, the performance from the algorithms in the benchmarking experiments has been discussed taking into account such properties. In this sense, by analyzing Table 2 it is noted three performance groups according to the accuracy over testing dataset.

With accuracy less than 80%, in the first group are the Single Decision Tree (SDT) and Cascade-Correlation Neural Network (CNN). The CNN is a self-organizing network that

determines its own size and topologies, by adding neurons to the architecture. The SDT also grows adding nodes to its structure, both for reaching greater preciseness during the learning process. As a consequence, these algorithms can lead to an overfitting to the train data, losing some generalization ability. Then, we use an overfitting control pruning the models to minimum cross-validated error over checking dataset. Despite this, the CNN performance decreases from 86.2% during training to 78.5% in the testing, and SDT from 91.6% to 72.3%, respectively. Therefore, these findings can be considered as an indicator of the complexity of the arboreal trunk texture, making hardier the classification task.

Capable of handling this issue better than the single tree-based model (SDT), the Decision Tree Forest (Random Forest), Stochastic Gradient Boosting (TreeBoost), and Boosted Rule-Based Model (C5.0) are in the second group of algorithms with medium-performance in the testing (from 80 to 90%), along with *k*-Nearest Neighbors (KNN).

The Random Forest and TreeBoost are ensembles based on different strategies of creating a collection of decision trees. The Random Forest uses the bagging (Bootstrap Aggregating) technique for creating trees grown in parallel, which afforded a generalization ability of 89.5%. On the other hand, the TreeBoost uses a sequential training (boosting) that resulted in a series of trees with 87.3% accuracy. Similarly, C5.0 is a voting classification algorithm also based on a boosting technique to create a collection of rules that achieved 85.3% accuracy. The boosting usually provides more accuracy than bagging strategy, except when there is noise in data, such as outliers (Bauer and Kohavi, 1999). Therefore, as the Random Forest outperforms the boosting-based models in the present analysis, we can consider some influence of outliers. Notwithstanding, as the bark texture in the arboreal trunk is a biological feature subject to imperfections, these outliers has not been removed because they can be caused by a natural variability. In turn, the KNN is a non-parametric algorithm of instance-based learning, in which a pattern is recognized by majority voting according to the similarity with the *k* nearest neighbors. By using kernel functions to weight the vote of the neighbors, the KNN provides 89.7% accuracy, slightly higher than ensemble-based models.

The third group with high-performance, more than 90% of accuracy over testing dataset, has been formed by the Support Vector Machine (SVM), Probabilistic Neural Network (PNN), Fuzzy Rule-Based Classification System (FRBCS), and Multilayer Perceptron Neural Network (MLP).

The SVM operates by finding an n-dimensional hyperplane in order to optimize the separation of different data classes. Although similar to artificial neural networks (ANN) in some aspects, the SVM is less prone to overfitting and has good adequacy for dealing with

high dimensional spaces and outliers, because it selects the most suitable features and considers only the most relevant points. Besides that, the SVM has a solution global and unique whilst the ANN can suffer from multiple local minima. Thus, in our analysis the SVM provides a significant improvement in comparison with most of the learning algorithms, reaching 96.2% over testing dataset.

Among the neural networks, the PNN performs classification based on the estimation of probability density functions, capable of dealing with erroneous data and computing nonlinear decision boundaries as complex as necessary, in order to approach the Bayes optimal, i.e., to minimize the error in a probabilistic manner as much as possible. Thus, relatively insensitive to outliers, the PNN achieves virtually the same performance than SVM, with 96.1% accuracy over testing dataset. In turn, the MLP allows nonlinear mappings, using logistic activation functions and back-propagation algorithm for adjusting the neural network weights. To prevent overfitting, we use the MLP architecture with minimum validated error during the learning process, reaching significant generalization ability correspondent to 90.8% accuracy, but still even less than PNN one.

Regarding the FRBCS, to be the focus of the present study, in the following we approach a more detailed description on the machine learning process, before presenting its accuracy over testing dataset. During the training we found that the gaussian curve membership function afforded a performance better than ones achieved with triangular and trapezoidal-shaped functions. Then, using gaussian functions for the fuzzy partitioning, variations of antecedent terms number has been assessed in combination with minimum and product t-norm (Figure 5).

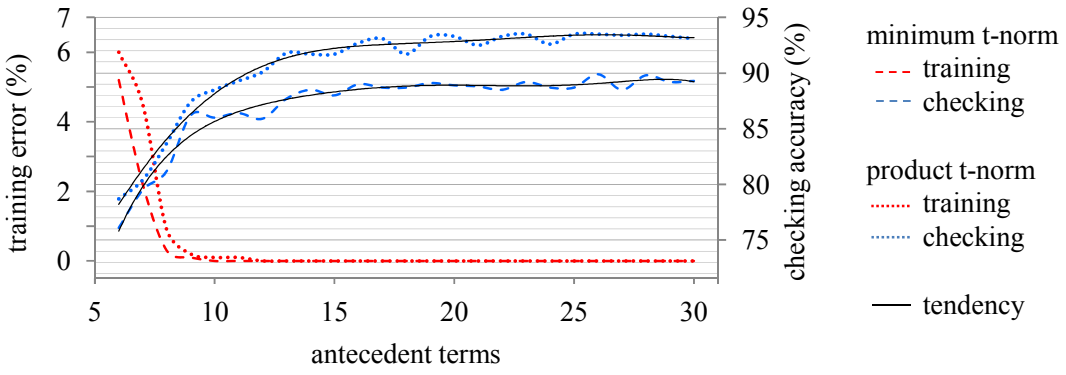



Figure 5. Performance of different setings of the fuzzy rule-based classification model, from the variations of antecedent terms number in combination with minimum and product t-norm.

Analyzing Figure 5 it is noted that, for both t-norms (minimum and product), about 10 antecedent terms were sufficient for the fuzzy rule-based classifier to reduce the error to zero during the training, but a higher accuracy over checking dataset required a greater number of terms. In that regard, one of the main aspects to highlight is the difference of performance provided by minimum and product t-norm.

Both product and minimum t-norm allowed aggregating the predictor variables via fuzzy intersections, modeling the simultaneous occurrence of patterns that characterize the same arboreal species. However, the product t-norm operates multiplying all the membership values and, in contrast, the minimum t-norm takes into account only the lowest membership during the aggregation process (Figure 6).

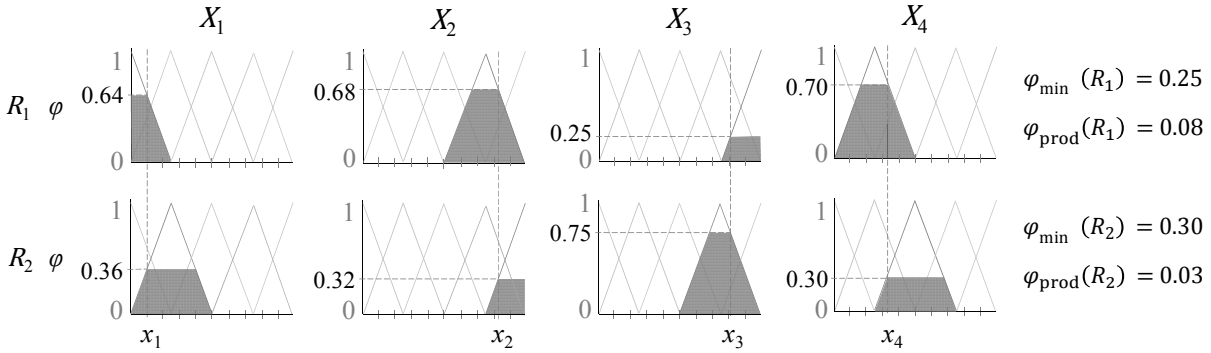



Figure 6. Aggregation process of predictor variables ($x_i$) in the rules 1 ($R_1$) and 2 ($R_2$), using minimum and product t-norm.

In Figure 6 we have a case in which a given sample has features (pattern values) belonging to more than one arboreal species, i.e, a sample with pertinence in both consequent classes of the rules 1 and 2, but with different membership degrees. By using the minimum t-norm the most critical condition given by the lowest membership become decisive, and hence we have a more rigorous classifier, but which can be naive by disregarding the other predictor variables.

As a consequence, for the case in Figure 6 the minimum t-norm would result in the arboreal species identification supported by the rule 2 ($\varphi_{\min}(R_2) > \varphi_{\min}(R_1)$). However, the sample has higher membership in the majority of the fuzzy regions correspondent to the consequent of the rule 1, as computed by the product t-norm ($\varphi_{\text{prod}}(R_1) > \varphi_{\text{prod}}(R_2)$). Thus, by taking account all the predictors, the product t-norm seems to afford a more assertive predictive modeling, so that it provided better performance than minimum t-norm in all

settings assessed in the present study (see Figure 5).

During the learning process we can note a tendency of accuracy improvement over checking dataset with the increase of the fuzzy regions number, which was more significant up to about 15 antecedent terms. This improvement seems to occur due to the increase of the decision areas ($D_j$) formed by each fuzzy if-then rule, as can be seen in Figure 7.

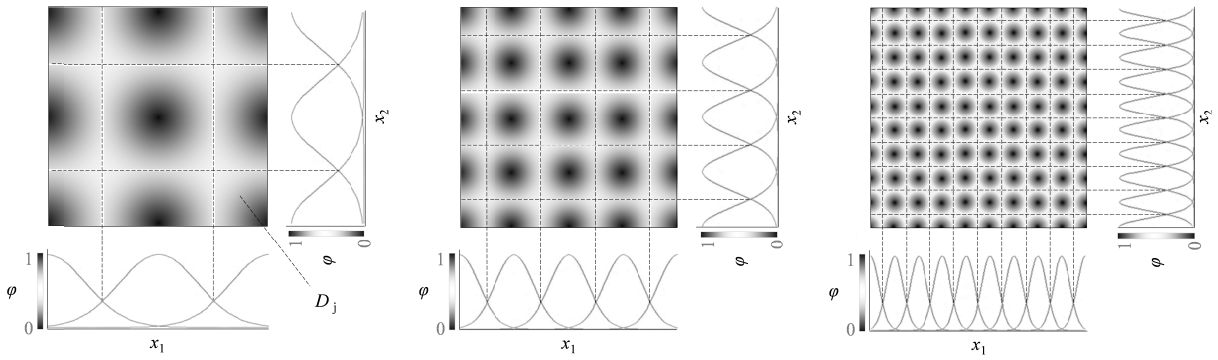



Figure 7. Increase in decision areas formed by the fuzzy if-then rules as consequence of the increment of the antecedent terms numbers.

Nevertheless, after a certain point there was a performance fluctuation that demanded an exhaustive search for the best accuracy over checking dataset (93.5 %), which was found using gaussian curve membership function, product t-norm, and 23 antecedents terms. Then, by using this setting the fuzzy-based model reaches 94.0% accuracy over testing dataset. Furthermore, considering the SVM as a reference, the FRBCS obtained 0.95 Kappa index.

## 4 Conclusions

In the present study we analyzed the enforceability of fuzzy-based pattern recognition for dealing with complexity related to natural variability of texture in the arboreal trunk, which can cause uncertainties due to ambiguity in the pattern matching.

By providing a nonlinear and smooth discriminate function, with the differential of taking into account the graded membership of a given sample in the matching patterns of different classes (arboreal species), the Fuzzy Rule-Based Classification System (FRBCS) afforded a high generalization ability, which outperformed the most of assessed learning algorithms, including ensembles with a lot of classifiers and kernel-based models, such as some artificial neural networks, widely used in pattern recognition tasks.

Furthermore, the Kappa index indicates that the FRBCS had an almost perfect agreement with the classifier with the best accuracy during the benchmarking experiment. Therefore, the fuzzy modeling can be considered an alternative approach, with a competitive and reliable performance for arboreal trunk texture recognition, in order to support the arboreal species identification using computational intelligence.

**References**

Abe, S., Thawonmas, R. 1997. A fuzzy classifier with ellipsoidal regions. *IEEE Trans. Fuzzy Syst.*, 5: 358-368.

Adriaenssens, V., Baets, B., Goethals, P.L.M., Pauw, N., 2004. Fuzzy rule-based models for decision support in ecosystem management. *The Science of the Total Environment*. 319: 1-12.

Arunpriya, C., Thanamani, A.S. 2015. Fuzzy inference system algorithm of plant classification for tea leaf recognition. *Indian Journal of Science and Technology, 8*(S7): 179-184.

Ascough, J.C., Maier, H.R., Ravalico, J.K., Strudley, M.W. 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*. 219: 383-399.

Bauer, E., Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning, 36(*1): 105-139.

Boman, J. 2013. *Tree species classification using terrestrial photogrammetry*. Umeå: Umeå University.

Bressane, A., Roveda, J.A.F., 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.

Bressane, A., Mochizuki, P.S., Caram, R.M., Roveda, J.A.F. 2016. A system for evaluating the impact of noise pollution on the population's health. *Reports in Public Health*. 32(5): 1-

11.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*. 22(2): 249-254.

Chi Z., Yan, H., Pham, T. 1996. *Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition*. World Scientic.

Costello, A.B., Osborne, J.W. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation,* 10(7): 1-9.

Fiel, S., Sablatnig, R. 2011. Automated identification of tree species from images of the bark, leaves and needles. In: *Proceedings of the 2011 Computer vision winter workshop*.

Gonzales, R. C., Woods, R. E. 2008. *Digital image processing*. 3ed. New Jersey: Pearson Prentice Hall.

Gonzales, R. C., Woods, R. E., Eddins, S. L. 2009. *Digital image processing using MATLAB*. 2ed. Gatesmark Publishing.

Gonzalez, A., Perez, R. 1999. SLAVE: A genetic learning system based on an iterative approach. *IEEE Trans. Fuzzy Syst.*, 7:176–191.

Harlick, R. M., Shanmugam, K., & Dinstein, I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6): 610–621.

Huang, Z.K., Huang, D.S., Du, J., Quan, Z.H., Guo, S.B. 2006. Bark classification based on textural features using Artificial Neural Networks. *Lecture notes in computer science*, 3972: 355-360.

Huang, Z.K. 2006. Bark classification using RBPNN based on both color and texture feature. *International journal of computer science and network security*. 6(10):100-103.

Ishibuchi, H., Nakashima, T. 2001. Effect of rule weights in fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, 9(4): 506-515.

Lermontov, A., Yokoyama, L., Lermontov, M., Machado, M.A.S. 2009. River quality analysis using fuzzy water quality index: Ribeira do Iguape river watershed, Brazil. *Ecological Indicators*. 9: 1188–1197.

Liu, L., Zhou, J., An, X., Zhang, Y, Yang, L. 2010. Using fuzzy theory and information entropy for water quality assessment in Three Gorges region, China. *Expert Systems with Applications*. 37: 2517–2521.

Liu, D., Zhou, Z. 2012. Water quality evaluation based on improved fuzzy matter-element method. *Journal of Environmental Sciences*. 24(7): 1210-1216.

Nauck, D., Kruse, R. 1997. A neuro-fuzzy method to learn fuzzy classification rules from data. *Fuzzy Sets Syst.*, 89(3): 277-288.

Pedrycz, W., Gomide, F. 2007. *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Wiley-IEEE Press, New York.

Porebski, A., Vandenbroucke, N., Macaire, L. 2007. Iterative feature selection for color texture classification. In: *Proceedings of the 2007 IEEE International conference on image processing.*

Riza, L.S., Bergmeir, C., Herrera, F., Benítez, J.M. 2015. Frbs: Fuzzy Rule-Based Systems for Classification and Regression in R. *Journal of Statistical Software*, 65(6): 1-30.

Silvert, W. 2000. Fuzzy indices of environmental conditions. *Ecological Modelling*. 130(1-3): 111-119.

Singh, H., Gupta, M.M, Meitzler, T., Hou, Z.G., Garg, K.K., Solo, A.M.G., Zadeh, L. 2013. Real-Life Applications of Fuzzy Logic. *Advances in Fuzzy Systems*, 2013: 1-3. doi.org/10.1155/2013/581879

Wan, Y.Y., Xiang, D.J., Huang, D.S., Chi, Z., Cheung, Y., Wang, X.F., Zhang, G.J. 2004. Bark texture feature extraction based on statistical texture analysis. In: *Proceedings of the 2004 Intelligent multimedia, video and speech processing.*

Yong, A.G., Pearce, S. 2013. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2): 79-94.

Zadeh, L. A., 2008. Is there a need for fuzzy logic? *Information Sciences*. 178(13): 2751-2780.

Zadeh, L. A., 1965. Fuzzy sets. *Information and Control*. 8: 338-353.

## CONSIDERAÇÕES FINAIS

A identificação arbórea é indispensável para muitos fins, tanto econômicos, quanto ambientais. No entanto, o uso de métodos mais frequentes como as chaves de identificação pode ser difícil, impreciso e até mesmo inviável na ausência de caracteres morfológicos, sobretudo de estruturas reprodutivas que são menos sujeitas a variações ambientais.

Assim, o estudo de métodos assistidos por computador tornou-se uma necessidade, não para substituir o conhecimento especializado, mas para o reconhecimento de características que pode ser usado para apoiar a identificação por um especialista, não para substituí-lo.

A partir da revisão de estudos anteriores, verificou-se que a análise de métodos computacionais para apoiar a identificação de espécies arbóreas vem se desenvolvendo nos últimos 20 anos.

Embora com muitos avanços, os estudos têm focado em características extraídas a partir de imagens da folha, que pode não estar disponível, como ocorre com espécies caducifólias em determinadas épocas do ano. Logo, o reconhecimento de padrões baseados na textura em imagens do tronco poderia ser uma alternativa, mas poucos estudos foram encontrados na literatura.

Dessa forma, concluiu-se que a realização de experimentos originais, que proporcionassem novos resultados sobre o reconhecimento da textura em imagens do tronco, poderia contribuir para o avanço na identificação de espécies arbóreas apoiada por inteligência computacional.

Entre os experimentos realizados, o primeiro analisou o desempenho discriminante de parâmetros estatísticos de primeira ordem. Pela análise de histogramas verificou-se que o uso de limites rígidos permitiu uma separação substancial das amostras pertencentes a diferentes classes, mas em alguns casos a complexidade biológica conduziu a uma perda de precisão na identificação das espécies.

Em conclusão, o uso de propriedades estatísticas foi considerado promissor. No entanto, para um desempenho satisfatório associado à inclusão contínua de novas espécies, a avaliação de mais características baseadas na textura em imagens do tronco arbóreo se mostrou necessária.

Considerando esses resultados, o desempenho de descritores de coocorrência foi avaliado em comparação com o uso dos parâmetros estatísticos de primeira ordem analisados no experimento anterior.

Ao considerar o arranjo espacial dos pixels na imagem, os descritores de coocorrência proporcionaram a melhoria esperada no desempenho descriminante dos padrões de textura em imagens do tronco.

Por outro lado, uma análise da importância relativa entre as variáveis preditoras destacou que algumas estatísticas de primeira ordem, como a entropia e a suavidade, superaram alguns descritores de concorrência.

Nesse sentido, concluiu-se que a melhor capacidade de generalização é alcançada combinando os descritores de coocorrência com as estatísticas de primeira ordem, o que proporcionou uma redução significativa nas taxas de erro de comissão e de omissão durante a validação com as amostras de teste.

A diversidade natural da textura em indivíduos de uma mesma espécie e, ao mesmo tempo, a similaridade entre indivíduos de diferentes espécies, sugerem que o acréscimo de variáveis preditoras proporcionaria o alcance de um melhor desempenho na identificação arbórea, o que se confirmou nos experimentos precedentes.

No entanto, os padrões de textura em imagens do tronco estão naturalmente correlacionados, uma vez que são resultantes das mesmas características morfológicas presentes na casca arbórea. Portanto, constatou-se que o estudo de técnicas capazes de tratar informação redundante poderia favorecer o reconhecimento de padrões e, consequentemente, melhorar a acurácia na identificação arbórea.

Para avaliar essa possibilidade, experimentos com técnicas de análise multivariada foram realizados. Como resultado, foi observada uma redução dimensional expressiva no espaço de variáveis preditoras, com efeitos positivos sobre a redução na taxa de erros durante a aprendizagem, e um aumento na capacidade de generalização sobre as amostras de teste.

Entre as técnicas avaliadas, a *Exploratory Factor Analysis* (EFA) foi a que alcançou o melhor desempenho, embora a *Fischer Discriminant Analysis* (FDA) tenha alcançado resultados ligeiramente superiores para condições de avaliação supostamente mais estáveis.

Considerando ainda que o uso da EFA para otimizar o esforço computacional em tarefas de classificação é menos observado na literatura, considerou-se que sua aplicação na última etapa experimental resultaria em maiores contribuições que a FDA que, assim como a *Principal Component Analysis* (PCA), já é amplamente estudada para esta finalidade.

Assim, norteando-se pelo resultado das etapas anteriores, as análises do último experimento foram realizadas a partir da extração de 70 padrões de textura baseados em estatísticas de primeira e segunda ordem que, tratados por meio da *Exploratory Factor*

*Analysis*, resultaram em um conjunto reduzido de variáveis preditoras com maior capacidade descriminante.

Então, o uso da modelagem fuzzy foi finalmente avaliado para lidar com a incerteza relacionada a dispersão na frequência dos valores pertinentes a cada espécie arbórea, que tem como efeito uma sobreposição de subespaços adjacentes no domínio das variáveis preditoras. Essa sobreposição causa uma ambiguidade no reconhecimento das amostras e, consequentemente, torna mais difícil a identificação das espécies por meio de limites rígidos nos padrões de textura em imagens do tronco.

Ao avaliar comparativamente outros algoritmos de aprendizagem amplamente usados para reconhecimentos de padrões, verificou-se o efeito dessa ambiguidade sobre a baixa capacidade de generalização de alguns classificadores, como foi o caso da *Cascade-Correlation Neural Network*.

A superação da técnica *bagging* (*Bootstrap Aggregating*), usada no algoritmo *Random Forest*, sobre a abordagem *boosting*, presente nos classificadores *Boosted Rule-Based Model* e *Stochastic Gradient Boosting*, foi outro indicador da complexidade envolvida, nesse caso, devido a presença de ruídos associados à ocorrência de valores extremos.

Diante desses aspectos, o *Fuzzy Rule-Based Classification System* proporcionou uma alta capacidade de generalização, superando *ensembles* e modelos com uso de descritores *kernel*, além de alcançar uma concordância quase perfeita com o classificador de melhor desempenho durante os testes de validação.

A partir desses resultados, conclui-se que a modelagem fuzzy constitui uma alternativa com desempenho competitivo e confiável para apoiar a identificação das espécies arbóreas caducifólias nativas da flora brasileira, por meio do reconhecimento de padrões de textura em imagens do tronco.

Vale ressaltar que as análises experimentais desenvolvidas ao longo da pesquisa se basearam na coleta de imagens realizada sempre sob as mesmas circunstâncias. Nesse sentido, pesquisas futuras poderiam avaliar como diferentes condições ambientais e configurações de equipamento para registro fotográfico afetam o reconhecimento da textura em imagens do tronco e, dessa forma, orientar procedimentos para um melhor desempenho.

A amostragem considerou apenas árvores adultas, de modo que experimentos contemplando indivíduos jovens representam outra possibilidade para novos avanços. Estudos futuros poderiam ainda avaliar a influência de diferentes condições ou estado de conservação do tronco arbóreo, devido à presença comum de interferentes, como resíduos, insetos e maus tratos.

Por fim, considerando que a abordagem desenvolvida propõe o uso da inteligência computacional para apoiar a identificação arbórea, e não para desempenhá-la de forma independente, o estudo de alternativas para integrar o reconhecimento de padrões ao conhecimento de especialistas constitui outro objetivo a ser alcançado em novos estudos.

## REFERÊNCIAS

BACKES, A. R., CASANOVA, D., BRUNO, O. M. Plant leaf identification based on volumetric fractal dimension. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 23, n. 6, p. 1145-1160, 2009.

BACKES, A. R., BRUNO, O. M. Plant leaf identification using color and Multi-scale Fractal Dimension. *Lecture notes on Computer Science*. v. 6134. p. 463-470, 2010.

BACKES, A. R., CASANOVA, D., BRUNO, O. M. Identificação de plantas por análise de textura foliar. *Anais do VI Workshop de Visão Computacional*, Presidente Prudente, 2011.

BOMAN, J. Tree species classification using terrestrial photogrammetry. 2013. Master Thesis (Computer Science), Umeå University, Umeå, 2013.

BRUNO, O. M., PLOTZE, R., O., FALVO, M., CASTRO, M. Fractal Dimension applied to plant identification. *Information Sciences*, v. 178, p. 2722-2733, 2008.

BUDOWSKI, G. The distinction between old secondary and climax species in tropical central american lowland forests. Tropical Ecology, v. 11, n. 1, p. 44-48, 1970.

CASANOVA, D., SA JUNIOR, J. J. M., BRUNO, O. M. Plant leaf identification using Gabor wavelets. *International Journal of Imaging Systems and Technology*, v. 19, n. 3, p. 236-243, 2009.

CHAKI, J., PAREKH, R. Plant leaf recognition using shape based features and Neural Network classifiers. *International Journal of Advanced Computer Science and Applications*, v. 2, n. 10, p. 41-47, 2011.

CHI, Z., HOUQIANG, L., CHAO, W. Plant species recognition based on bark patterns using novel Gabor filter banks. In: *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, pp. 1035-1038, 2003.

DU, J., HUANG, D., WANG, X., GU, X. Shape recognition based on radial basis probabilistic neural network and application to plant species identification. *Lecture Notes in Computer Science*, v. 3497, p. 281-285, 2005.

DU, J. X., WANG, X. F., GU, X. Computer-aided plant species identification (CAPSI) based on leaf shape matching technique, *Transactions of the Institute of Measurement and Control*, v. 28, n. 3, p. 275-284, 2006.

DU, J. X., WANG, X. F., ZHANG, G. J. Leaf shape based plant species recognition. *Applied Mathematics and Computation*, v. 185, p. 883-893, 2007.

FIEL, S., SABLATNIG, R. Automated identification of tree species from images of the bark, leaves and needles. In: Computer Vision Winter Workshop, 16, p. 67-74, 2011.

FU, H., CHI, Z. A two-stage approach for leaf vein extraction. In: IEEE International Conference on Neural Networks and Signal Processing, *Proceedings*…, Nanjing, 208 - 211, 2003.

GANDOLFI, S., LEITÃO-FILHO, H. F., BEZERRA, C. L. Levantamento florístico e caráter sucessional das espécies arbustivo-arbóreas de uma floresta mesófila semidecídua no município de Guarulhos, SP. Revista Brasileira de Biologia, v. 55, n. 4, p. 753-767, 1995.

GOUVEIA, F., FILIPE, V., REIS, M., COUTO, C., BULAS-CRUZ, J. Biometry: the characterisation of chestnut-tree leaves using computer vision. In: IEEE International Symposium on Industrial Electronics, *Proceedings*…, Guimarães, p. 757-760, 1997.

GU, X., DU, J.-X., WANG, X.-F. Leaf recognition based on the combination of wavelet transform and gaussian interpolation. *Lecture Notes in Computer Science*, v. 3644, p. 253-262, 2005.

HUANG, Z. K. Bark classification using RBPNN based on both color and texture feature. *International Journal of Computer Science and Network Security*, v. 6, n. 10, p. 100-103, 2006.

HUANG, Z. K., HUANG, D. S., DU, J., X. QUAN, Z. H., GUO, S. B. Bark classification based on textural features using Artificial Neural Networks. *Lecture Notes in Computer Science*, v. 3972, p. 355-360, 2006.

IM, C., NISHIDA, H., KUNII, T. L. Recognizing plant species by leaf shapes-a case study of the Acer family. *Pattern Recognition*, v. 2, p. 1171 – 1173, 1998.

_____. Recognizing plant species by normalized leaf shapes. *Vision Interface*, v. 19, n. 21, p. 397-404, 1999.

KADIR, A., NUGROHO, L. E., SUSANTO, A., SANTOSA, P. I. A comparative experiment of several shape methods in recognizing plants. *International Journal of Computer Science & Information Technology*, v. 3, n. 3, p. 256-263, 2011a.

_____. Leaf classification using shape, color, and texture. *International Journal of Computer Trends and Technology*, v. 2, n. 1, p. 225-230, 2011b.

KAUR, G., MONGA, H. Classification of Biological Species Based on Leaf Architecture–A review. *International Journal of Computer Science and Information Technology & Security*, v. 2, n. 2, p. 332-334, 2012.

KIM, S. J., KIM, B. W., KIM, D. P. Tree recognition for landscape using by combination of features of its leaf, flower and bark. In: *Proceedings of the 2011 SICE Annual Conference*, p. 1147–1151, 2011.

LEE, C. L., CHEN, S. Y. Classification of leaf images. *International Journal of Imaging Systems and Technology*, v. 16, n. 1, p. 15-23, 2006.

LI, Y., ZHU, Q., CAO, Y., WANG, C. A leaf vein extraction method based on snakes technique. In: IEEE International Conference on Neural Networks and Brain, *Proceedings*….

Beijing, p. 885-888, 2005.

LORENZI, H. *Arvores brasileiras*: manual de identificação e cultivo de plantas arbóreas nativas do Brasil. Nova Odessa: Plantarum, 1992.

MACHADO, B. B., CASANOVA, D., GONÇALVES, W. N., BRUNO, O. M. Partial differential equations and fractal analysis to plant leaf identification. *J. Phys.: Conf. Ser. 410 012066*, 2013.

MACIEL, M. N. M, WATZLAWICK, L. F., SCHOENINGER, E. R., YAMAJI, F. M. Classificação ecológica das espécies arbóreas. *Revista acadêmica: ciências agrárias e ambientais,* v.1, n.2, p. 69-78, 2003.

NAM, Y., HWANG, E. J., KIM, D. Y. A similarity-based leaf image retrieval scheme: joining shape and venation features. *Comput Vis Image Understand*, v. 110, p. 245-259, 2008.

PLOTZE, R. O., FALVO, M., PÁDUA, J. G., BERNACCI, L. C., VIEIRA, M. L. C., OLIVEIRA, G. C. X., BRUNO, O. M. Leaf shape analysis by the multiscale minkowski fractal dimension, a new morphometric method: a study in passiflora L. (Passifloraceae). *Canadian Journal of Botany*, v. 83, n. 3, p. 287-301, 2005.

POREBSKI, A., VANDENBROUCKE, N., MACAIRE, L. Iterative feature selection for color texture classification. In: IEEE International Conference on Image Processing. *Proceedings…*, p. 509-512, 2007.

PRIYA, A. C., THANAMANI, A. S. A survey on species recognition system for plant classification. *International Journal Computer Technology & Applications*, v. 3, n. 3, p. 1132-1136, 2012.

QI, H., YANG, J.-G. Sawtooth feature extraction of leaf edge based on support vector machine. *Machine Learning and Cybernetics* v. 5, p. 3039-3044, 2003.

ROSSATTO, D. R., CASANOVA, D., KOLB, R. M., BRUNO, O. M. Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with

species from Neotropical Melastomataceae (Miconieae tribe). *Plant Systematics and Evolution,* v. 291, n. 1, p. 103-116, 2011.

SAKAI, A. N., ALLENDORF, F. W., HOLT, J. S., LODGE, D. M., MOLOFSKY, J., WITH, K. A., BAUGHMAN, S., CABIN, R. J., COHEN, J. E., ELLSTRAND, N. C., McCAULEY, D. E., O'NEIL, P., PARKER, I. M., THOMPSON, J. N., WELLER, S. G. The population biology of invasive species. *Annual Review of Ecology and Systematic*, v. 32, p. 305-332. 2001.

SANTOS, N. R. Z., TEIXEIRA, I. F. *Arborização de vias públicas*: ambiente x vegetação. Instituto Souza Cruz. Porto Alegre: Pallotti. 2001.

SILVA, L. C. Plantas ornamentais tóxicas presentes no shopping Riverside Walk em Teresina – PI. *Revista Brasileira de Arborização Urbana,* v. 4, n. 3, p. 69-85, 2009.

SINGH, K., GUPTA, I., GUPTA, S. SVM-BDT PNN and Fourier Moment Technique for classification of leaf shape. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, v. 3, n. 4, p. 67-78, 2010.

SONG, J., CHI, Z., LIU, J., FU, H. Bark classification by combining grayscale and binary texture features. In: Intelligent Multimedia, Video and Speech Processing, *Proceedings*..., p. 450-453, 2004.

SILVA, L. C. Identificação das espécies ornamentais nocivas na arborização urbana de Santiago/RS. *Revista Brasileira de Arborização Urbana,* v. 6, n. 2, p. 44-56, 2011.

WAN, Y. Y., XIANG D, J., HUANG, D. S., CHI, Z., CHEUNG, Y. M., WANG, X. F., ZHANG, G. J. Bark texture feature extraction based on statistical texture analysis. In: *Proceedings of the 2004 Intelligent Multimedia, Video and Speech Processing*, p. 482–485, 2004.

WANG, Z., CHI, Z., FENG, D., WANG, Q. Leaf Image Retrieval with Shape Features. *Lecture Notes in Computer Science*, v. 1929, n. 2000, p. 477-487, 2003.

WANG, X., HUANG, D. S., DU, J. X., XU, H., HEUTTE, L. Classification of plant leaf images with complicated background. *Appl Math Comput*, v. 205, p. 916-926, 2008.

WIT, M. P., CROOKES, D. J., WILGEN, B. W. Conflicts of interest in environmental management: estimating the costs and benefits of a tree invasion. *Biological Invasions*, v. 3, p. 167-178, 2001.

WU, S. G., BAO, F. S., XU, E. Y., WANG, Y-X.,CHANG, Y-F., XIANG, Q-L. A leaf recognition algorithm for plant classification using Probabilistic Neural Network. *The Computing Research Repository*, v. 1, p. 11-16, 2007.

YANIKOGLU, B., APTOULA, E., TIRKAZ, C. Automatic plant identification from photographs. *Machine vision and applications*, v. 25, n. 6, p. 1369-1383, 2014.

YE, Y., CHEN, C., LI, C.-T., FU, H., CHI, Z. A computerized plant species recognition system. In: International Symposium on Intelligent Multimedia, Machine vision and applications, *Proceedings*…, p. 723 - 726, Hong Kong, 2004.