



ISÁIRA LEITE E LOPES

***FEATURE SELECTION* APLICADA À BIOMETRIA
FLORESTAL**

LAVRAS – MG

2022

ISÁIRA LEITE E LOPES

FEATURE SELECTION APLICADA À BIOMETRIA FLORESTAL

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia Florestal, curso de Doutorado, área de concentração em Manejo Florestal, para a obtenção do título de Doutor.

Prof. Dr. Lucas Rezende Gomide
Orientador

Prof^ª. Dra. Angélica Sousa da Mata
Coorientadora

LAVRAS – MG

2022

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Lopes, Isáira Leite e.

Feature selection aplicada à Biometria Florestal / Isáira Leite e
Lopes. - 2022.
106 p. : il.

Orientador(a): Lucas Rezende Gomide.

Coorientador(a): Angélica Sousa da Mata.

Tese (doutorado) - Universidade Federal de Lavras, 2022.

Bibliografia.

1. Inteligência computacional. 2. Crescimento e Produção
Florestal. 3. Colheita Florestal. I. Gomide, Lucas Rezende. II. Mata,
Angélica Sousa da. III. Título.

ISÁIRA LEITE E LOPES

***FEATURE SELECTION* APLICADA À BIOMETRIA FLORESTAL**

FEATURE SELECTION APPLIED TO FOREST BIOMETRY

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia Florestal, curso de Doutorado, área de concentração em Manejo Florestal, para a obtenção do título de Doutor.

APROVADA em 04 de março de 2022.

Prof. Dr. Renato Vinícius Oliveira Castro DEFLO/UFSJ

Prof. Dr. José Roberto Soares Scolforo DCF/UFLA

Dr. Kalill José Viana da Páscoa LEMAF/UFLA

Prof. Dr. Lucas Rezende Gomide

Orientador

Prof^ª. Dra. Angélica Sousa da Mata

Coorientadora

LAVRAS – MG

2022

*A Deus,
À minha mãe,
Maria Geralda Leite
À minha irmã,
Isadora Leite e Lopes
Ao meu pai,
Edison Fernando Oliveira Lopes
Dedico.*

AGRADECIMENTOS

A Deus, entrego, confio e agradeço por todas as oportunidades a mim concedidas: a vida, os aprendizados e as conquistas. Minha imensa gratidão por nunca me desamparar e proporcionar ânimo e força nos momentos mais difíceis.

A Nossa Senhora pelas intercessões constantes, proteção e por passar na frente das situações, guiando meus passos.

A minha amada mãe por nunca medir esforços para proporcionar a mim e a minha irmã, todas as oportunidades de estudo, sempre nos inspirando e incentivando a dar o nosso melhor e também apoiando as nossas decisões.

A minha amada irmã, por ser minha fiel companheira, melhor amiga e por todo apoio.

Ao meu amado pai por todo amor e torcida em todos os momentos da minha vida.

À minha família e amigos pela torcida, apoio e orações nos momentos decisivos.

Aos amigos de moradia, Anny Ataide, Lorena Barbosa e Vítor Abreu por todos os momentos compartilhados e amizade.

Ao professor Lucas Rezende Gomide pela valiosa orientação, paciência, confiança em mim depositada, ensinamentos e toda contribuição e apoio para meu desenvolvimento pessoal e profissional.

A professora Angélica Sousa da Mata pela preciosa coorientação, ensinamentos, contribuições, sempre me animando no desenvolvimento deste trabalho e apoiando a minha trajetória profissional.

A Laís Araújo, Evandro Miranda e Thomaz Bastos pelo inestimável apoio, interação sinérgica, troca intelectual e contribuições no artigo 1.

Ao professor Renato Castro, Dr. Kalill Páscoa e Prof^ª. Carolina Jarochinski pelo apoio na minha trajetória profissional e contribuições essenciais no meu projeto de tese, qualificação do doutorado e na melhoria do artigo 2.

A todos do Laboratório de Planejamento e Otimização Florestal (GOPLAN) por toda convivência, amizade e troca de ideias em um ambiente sinérgico.

Aos amigos do Laboratório de Estudos e Projetos em Manejo Florestal (LEMAF) pela ótima convivência e a todos aqueles que auxiliaram na realização do campo para coleta de dados na Reserva da Matinha.

A Universidade Federal de Lavras (UFLA) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela oportunidade de crescimento profissional.

RESUMO GERAL

Os avanços computacionais oportunizaram uma maior viabilidade da coleta e armazenamento de dados, e processamento de algoritmos com a expansão de big data no setor florestal. Em alinhamento com isso, técnicas de inteligência computacional têm sido adotadas como suporte a tomada de decisão em uma gama de problemas. Dentre suas aplicações, o processo de seleção de atributos (*Feature selection*) contribui com êxito na automatização da tarefa de redução da dimensionalidade dos dados para otimização de um subconjunto de variáveis relevantes na construção de modelos. Diante dessa perspectiva, a tese foca no uso do algoritmo genético juntamente com o *Random Forest* (GA-RF) para seleção de variáveis na modelagem da produtividade de máquina florestal (Artigo 1) e do incremento periódico anual em diâmetro em uma Floresta Estacional Semidecidual Montana (Artigo 2). No artigo 1, o objetivo do trabalho foi testar diferentes abordagens metodológicas na geração de modelos com boa capacidade preditiva além da investigação da importância de variáveis oriundas de condições edafoclimáticas, registros dos operadores e inventário florestal. O GA-RF foi selecionado por apresentar alto poder de generalização com redução do erro das estimativas além da maximização da importância das variáveis relevantes na produtividade da máquina. O artigo 2 objetivou avaliar a incorporação dos efeitos da competição em um modelo de crescimento em nível de árvores individuais, baseando-se na investigação de diferentes categorias de índices clássicos de competição e do uso de métricas de redes complexas, metodologia proposta nesse estudo. A metodologia GA-RF foi eficiente em conciliar aspectos com significado ecológico e melhoria da acurácia por meio da seleção de índices independentes da distância e métricas de redes complexas para a modelagem do crescimento das respectivas espécies, *Xylopia brasiliensis* e *Copaifera langsdorffii*.

Palavras-chave: Inteligência computacional. Algoritmo genético. Crescimento e Produção Florestal. Colheita Florestal.

GENERAL ABSTRACT

Computational advances made possible greater viability of data collection, storage, and algorithms processing with the expansion of big data in the forestry sector. In line with this, computational intelligence techniques have been increasingly applied to support decision-making in several problems. Among their applications, the feature selection process successfully contributes to the task automation of reducing the dimensionality of the data for optimizing a subset of relevant variables in the models building. Given this perspective, the thesis focuses on the genetic algorithms' use in association with the Random Forest (GA-RF) for selecting variables in the modeling of forest machine productivity (Article 1) and the periodic annual diameter increment in a Semideciduous seasonal montane forest in Brazil (Article 2). In article 1, the objective of the work was to test different methodological approaches in the generation of models with good predictive capacity, in addition to investigating the importance of variables arising from soil and climate conditions, operator records, and forest inventory. We selected the GA-RF because it has a high generalization power by reducing the errors' estimates, in addition to maximizing the importance of relevant variables in the machine's productivity. Article 2 aimed to evaluate the incorporation of competition effects in a growth model at individual trees level, based on the investigation of different categories of classical competition indices and an additional methodology proposed in this study, known as metrics of complex networks. The GA-RF methodology was efficient by combining ecological meaning and accuracy improvements. It selected distance-independent indices and complex network metrics for modeling the growth of the species, *Xylopia brasiliensis*, and *Copaifera langsdorffii*, respectively.

Keywords: Computational Intelligence. Genetic Algorithm. Forest Growth and Yield. Forest Harvesting.

LISTA DE FIGURAS

PRIMEIRA PARTE

Figura 1	O centro da cidade de Königsberg (a), o mapa esquemático do problema das pontes de Königsberg (b) e sua representação na forma de grafo (c) ¹	21
Figura 2	Representação de uma pequena rede com oito nós e dez arestas.....	22
Figura 3	Diferenciação entre o grafo não direcionado e direcionado com sua matriz de adjacência.	23
Figura 4	Fluxograma das etapas de funcionamento do Algoritmo genético simples.....	26

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

Figure 1	Flowchart of the Random forest structure.....	44
Figure 2	Flowchart of Artificial neural network structure.	46
Figure 3	Indicators of the association between independent variables and machine productivity.....	49
Figure 4	Analysis of the predictor variable importance.	50
Figure 5	Residuals plots analysis of modeling methods (RF: random forest; GA-RF: Genetic Algorithm and Random Forest; ANN: Artificial Neural Network; GA-ANN: Genetic Algorithm and Artificial Neural Network; and Stepwise), and datasets.....	53
Figure 6	Taylor diagrams of machine productivity modeling methods (RF: Random forest; GA-RF: Genetic Algorithm and Random Forest; ANN: Artificial Neural Network; GA-ANN: Genetic Algorithm and Artificial Neural Network; and Stepwise), and datasets.....	54

ARTIGO 2

Figure 1	(a) Study area map, and (b) native species in spatial arrangements of trees over the Forest Reserve, and the competitor trees selection at Bitterlich method.....	70
Figure 2	Mean increment at diameter classes for each studied species.....	80
Figure 3	Cluster analysis of grouping the set of competition indices/metrics and Pearson's correlation of periodic annual diameter increment (PAI_d) for <i>Copaifera langsdorffii</i> ■ and <i>Xylopia brasiliensis</i> ■	83
Figure 4	Residuals plot with marginal histograms considering the modeling strategies for <i>Copaifera langsdorffii</i> and <i>Xylopia brasiliensis</i>	85

Figure 5 Graphical analysis of the variables selected for <i>Copaifera langsdorffii</i> (■) and <i>Xylopia brasiliensis</i> (■) models.	87
Figure 6 Individual tree diameter increment predicted by Chapman-Richards function (—) and Genetic algorithm with Random forest – (GA-RF—) using complex network metrics (S ₄) in validation dataset (●).	88

LISTA DE TABELAS

SEGUNDA PARTE – ARTIGOS

ARTIGO 1

Table 1. Descriptive values of the studied variables for the analysis (n=297)..... 43

Table 2. Adjustment statistics for the training and validation data, the number of variables, and selected variables using the Stepwise, RF, GA-RF, ANN, and GA-ANN method..... 52

ARTIGO 2

Table 1 - Classical competition indices evaluated in quantifying inter-tree competition. 75

Table 2 - Competition metrics/indices average values within the subject species' diameter classes. 81

Table 3 - Statistics analysis of the periodic annual diameter increment (PAI_d) modeling strategies for *Copaifera langsdorffii* and *Xylopia brasiliensis*..... 84

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO GERAL	14
2 REVISÃO DE LITERATURA.....	17
2.1 Planejamento da colheita.....	17
2.2 Competição entre árvores: Do estado da arte às futuras direções	18
2.3 Redes complexas	21
2.4 Random Forest.....	24
2.5 Algoritmo genético (AG)	25
2.6 Seleção de variáveis.....	27
3 CONSIDERAÇÕES GERAIS	29
REFERÊNCIAS	30
SEGUNDA PARTE – ARTIGOS.....	38
ARTIGO 1 – A comparative approach of methods to estimate machine productivity in wood cutting	38
Introduction	40
Materials and methods.....	41
<i>Experiment description.....</i>	<i>41</i>
<i>Database structure</i>	<i>42</i>
<i>Methods for predicting the machine productivity.....</i>	<i>43</i>
<i>Stepwise</i>	<i>43</i>
<i>Random Forest (RF).....</i>	<i>44</i>
<i>Artificial Neural Networks (ANN)</i>	<i>45</i>
<i>Genetic algorithm for feature selection in Random Forest (GA-RF) and Artificial Neural Network (GA-ANN)</i>	<i>46</i>
<i>Variable selection performance and methods assessment</i>	<i>47</i>
Results.....	48
Discussion	54
Conclusion	59
References.....	60
ARTIGO 2 – Complex network metrics and feature selection for modeling individual tree diameter growth	65
1 INTRODUCTION	67
2 MATERIAL AND METHODS	69
2.1 Site description and tree database.....	69
2.2 Inter-tree competition indices	70

2.2.1	Complex network metrics.....	71
2.2.2	Classical competition indices	74
2.2.3	Indices and metrics analysis	76
2.3	Tree diameter increment modeling.....	76
2.4	Goodness-of-fit metrics for modeling strategies evaluation.....	78
3	RESULTS.....	79
3.1	Forest structure and inter-tree competition.....	79
3.2	Individual tree diameter growth modeling performance	84
4	DISCUSSION.....	88
5	CONCLUSION	92
	REFERENCES	94
	APPENDIX A. SUPPLEMENTARY DATA	101

PRIMEIRA PARTE

1 INTRODUÇÃO GERAL

O aporte de informações fidedignas concernentes aos atributos florestais é criticamente relevante para elaboração de planos de manejo com qualidade e confiabilidade. Esforços na compreensão e quantificação acurada das estimativas configuram seu ponto-chave no suporte a tomadas de decisão para os gestores florestais. Face a crescente geração de *big data* proveniente de diversas fontes de dados, as técnicas de inteligência artificial têm sido potencialmente aplicadas na modelagem de uma gama de problemas florestais (ASHRAF et al., 2015; LIU et al., 2018). A exemplo desses, pode-se citar a projeção de diâmetro e altura (VIEIRA et al., 2018), predição do afilamento (NUNES; GÖRGENS, 2016), estoque de carbono (SAFARI et al., 2017), biomassa acima do solo (BISPO et al., 2020), crescimento e produção florestal (ASHRAF et al., 2013) e produtividade da colheita (ROSSIT et al., 2019). Apesar da consolidação da regressão clássica na modelagem florestal, a alta variabilidade dos dados florestais e as relações complexas não lineares entre as variáveis podem inviabilizar o atendimento as suas pressuposições estatísticas concernentes a independência dos dados, normalidade e homoscedasticidade (ASHRAF et al., 2013; SILVA et al., 2021). Nesse sentido, a inteligência artificial desponta com êxito no fornecimento de resultados satisfatórios devido a suas vantagens competitivas que envolvem aprender à partir de uma amostra de dados limitada (SHAO; LUNETTA, 2012), identificar padrões em dados com relações complexas não-lineares (HAMIDI et al., 2021) e lidar com variáveis de naturezas distintas, alta dimensionalidade de dados (CORTE et al., 2020), presença de ruídos e dados faltantes (BRACKENRIDGE et al., 2022).

Os avanços tecnológicos têm oportunizado a coleta e armazenamento de grande quantidade e variedade de dados no setor florestal, oriundos do monitoramento de atributos florestais desde fontes de sensoriamento remoto até sistemas terrestres com obtenção de dados em tempo real. Apesar da aplicação de *big data* na gestão florestal ainda situar-se em estágio inicial, o seu progresso tem potencial para tornar o manejo florestal inteligente, interconectado e digital no tocante ao atendimento a diferentes serviços e necessidades (ZOU et al., 2019). Frente à complexidade dos dados, é demandado o processamento de diferentes técnicas de

Machine learning, um ramo de inteligência artificial, com destaque para o *Random Forest*, as redes neurais artificiais e a máquina vetor de suporte pelo bom desempenho preditivo evidenciado em inúmeros trabalhos na literatura (HONG et al., 2018; OU; LEI; SHEN, 2019; SILVEIRA et al., 2019; TAVARES JÚNIOR et al., 2020). Entretanto, a alta dimensionalidade dos dados representa um desafio para as análises e tomadas de decisão, sendo a seleção de atributos (*feature selection*) uma tarefa eficiente no solucionamento desse problema. *Feature selection* consiste em uma etapa de pré-processamento na mineração de dados (*Data mining*) e *Machine learning* fundamentada na remoção de atributos irrelevantes e redundantes. A relevância dessa etapa é atribuída a sua comprovada eficiência na construção de modelos mais simples e interpretáveis com ganho no desempenho do algoritmo, melhoria da acurácia dos resultados e redução do tempo computacional (CAI et al., 2018; KHALID; KHALIL; NASREEN, 2014; LI et al., 2018; XUE et al., 2016). Os métodos de *feature selection* são categorizados em abordagens *filter*, *wrapper* e *embedded*. Estas categorias movem de uma abordagem simplificada e computacionalmente rápida com menor acurácia (método *filter*) para a busca do equilíbrio entre tempo computacional e acurácia por meio da combinação dos métodos *filter* e *wrapper*, que consiste no método *embedded* (APOLLONI; LEGUIZAMÓN; ALBA, 2016; GHOSH et al., 2020). Comparado ao método *filter*, a vantagem da abordagem *wrapper* reside em seu critério de seleção de atributos, em que os subconjuntos de atributos testados são avaliados conforme o desempenho do preditor. Desta forma, o preditor é encapsulado em um algoritmo de busca que encontrará um subconjunto otimizado de atributos que fornece uma maior acurácia das estimativas (CHANDRASHEKAR; SAHIN, 2014; HONG et al., 2018). A abordagem *wrapper* com uso do algoritmo genético (AG) tem sido atrativa em termos de operacionalidade e boa capacidade de busca (JIANG et al., 2017). Na literatura, a implementação do AG na seleção de atributos tem proporcionado ganhos em termos de acurácia (HONG et al., 2018; MURTHY; KOOLAGUDI, 2018) e interpretabilidade ecológica e biológica na modelagem do afilamento de árvores (LACERDA et al., 2022) e altura total (MIRANDA et al., 2022).

Face ao exposto, objetivou-se com esta tese implementar o algoritmo genético em associação com algoritmos de *Machine learning* em tarefas de seleção de atributos para modelagem florestal. No artigo 1 realizou-se a investigação de diferentes abordagens metodológicas na modelagem da produtividade da garra traçadora envolvendo regressão linear via *stepwise*, algoritmos de *Machine learning* (*Random Forest* e redes neurais artificiais) e o algoritmo genético na seleção de atributos para os preditores *Random Forest* (GA-RF) e redes

neurais artificiais (GA-RNA). Esta abordagem comparativa teve como objetivo avaliar o desempenho preditivo dos modelos e revelar as variáveis independentes com maior poder explicativo. O artigo 2 despontou da investigação do comportamento da competição na modelagem do crescimento em diâmetro a nível de árvores individuais em uma Floresta Estacional Semidecidual Montana. O escopo desse artigo introduz um comparativo de diferentes categorias de índices de competição clássicos (dependentes, semi-independentes e independentes da distância) frente a metodologia proposta com base em métricas topológicas das redes complexas. Desta forma, o algoritmo genético com o *Random Forest* (GA-RF) foi implementado para seleção dos índices/métricas com maior poder preditivo em termos de acurácia e interpretação ecológica do crescimento em diâmetro.

Com base nesta conjuntura, a estruturação da tese envolveu duas partes. A primeira parte com foco na revisão de literatura para uma introdução e contextualização das temáticas bem como a explanação dos métodos utilizados e a segunda parte configurou um aprofundamento na abordagem da produtividade da colheita florestal e do crescimento em diâmetro em nível de árvores individuais por meio do desenvolvimento dos respectivos artigos, 1 e 2.

2 REVISÃO DE LITERATURA

2.1 Planejamento da colheita

A colheita de madeira demanda alto investimento de capital, sendo de extrema importância a viabilidade econômica de suas operações (SOMAN; KIZHA; ROTH, 2019). Estas operações abrangem desde a derrubada e processamento de árvores até o transporte secundário para as usinas. É atribuído aos gestores florestais o planejamento adequado dessas operações considerando as condições ambientais, sociais e econômicas no detalhamento das atividades, horários, locais e acessos à estrada. As especificações dessas atividades derivam de uma variedade de fatores concernentes ao tipo de floresta, sistema de colheita e produtos finais (FENG; AUDY, 2020).

O planejamento da colheita deve ser delineado de modo a assegurar que a produção florestal atenda satisfatoriamente as especificações de matéria-prima requisitadas para o abastecimento das fábricas, abrangendo diferentes níveis de abordagens categorizados em estratégico, tático e operacional. Em síntese, o planejamento estratégico concerne a determinação das áreas de florestas a serem colhidas afetando as decisões de construção da malha rodoviária florestal. Enquanto o planejamento tático e o operacional envolvem quais as áreas a serem cortadas, quando, quais volumes e sortimentos, quais os requisitos em termos de membros de equipe e equipamentos florestais. O maior detalhamento das informações é inerente ao nível operacional, onde encontram-se muitos desafios técnicos no desenvolvimento e implementação de sistemas de apoio a decisão (FENG; AUDY, 2020)

A complexidade dos fatores que norteiam as operações de colheita tornam difícil a tomada de decisão no tocante ao sistema de colheita mais eficiente (GHAFARIYAN; BROWN, 2013), sendo primordial para suporte às tomadas de decisão, o desenvolvimento de modelos de produtividade das máquinas florestais. Mundialmente, a produtividade de máquinas de colheita tem sido foco de estudos por mais de 25 anos. A literatura existente tem comprovado que a produtividade está sujeita a variação de diferentes fatores como condições do povoamento e do local, configuração do equipamento, objetivos do manejo e experiência do operador (HIESL; BENJAMIN, 2013). Congruente a isso, a indústria florestal tem avançado em direção ao conceito da Indústria 4.0 com a incorporação de tecnologias orientadas a automatização da coleta de dados, configurando novas oportunidades e desafios para a modelagem da

produtividade das operações de colheita (ERIKSSON; LINDROOS, 2014; FENG; AUDY, 2020; LISKI et al., 2020). Nesse sentido, há uma tendência crescente na aplicação de métodos de *Machine learning* para melhorias na acurácia das estimativas e na compreensão da relação entre as variáveis preditoras e a resposta (LISKI et al., 2020). Levers et al. (2014) ressalta a capacidade da abordagem *Machine learning* frente aos modelos de regressão linear em extrair conhecimento a partir da dados de colheita caracterizados por relações não lineares entre as variáveis. Adicionalmente, o uso de métodos de *Machine learning* confere vantagens em termos de inclusão de variáveis categóricas nos modelos. Gonçalves et al. (2022) demonstraram êxito na aplicação desses métodos na modelagem do corte florestal mecanizado via *harvester* resultando em um maior desempenho preditivo da produtividade com uso do algoritmo *Boosted*, seguido pelas redes neurais artificiais e o ANFIS (*Adaptive network-based fuzzy inference system*). Rossit et al. (2019) utilizaram árvores de decisão para predição da produtividade de *harvesters*, alcançando elevado nível de acurácia (em média 90%) e interpretabilidade dos modelos com base na obtenção das variáveis com maior importância preditiva. Portanto, a modelagem da produtividade via métodos de *Machine learning* é uma ferramenta útil no fornecimento de valiosos *insights*, que podem contribuir decisivamente no manejo florestal e no direcionamento de políticas e investimentos florestais (LEVERS et al., 2014). Complementarmente, face ao aumento da quantidade e complexidade dos dados, Liski et al. (2020) destaca que as futuras direções conduzem a operacionalidade de modelos de *Machine learning* em *websites* ou plataformas em nuvem com alimentação dos modelos por meio de dados fornecidos continuamente as previsões.

2.2 Competição entre árvores: Do estado da arte às futuras direções

O estudo da dinâmica florestal busca entender as mudanças na estrutura e composição da floresta ao longo do tempo, incluindo seu comportamento em resposta a perturbações antrópicas e naturais (PRETZSCH, 2009). A simulação da dinâmica florestal em nível de árvores individuais é fundamentada em um sistema de equações que agregam o crescimento, o ingresso e a mortalidade das árvores (BURKHART; TOMÉ, 2012). Portanto, torna-se fundamental a identificação de fatores que impulsionam a dinâmica florestal (DING et al., 2019). Dentre estes fatores, tem-se o clima, micro-ambiente, características genéticas, tamanho da árvore, idade e competição. A competição entre as árvores, em nível de vizinhança local, destaca-se como um dos principais fatores que fornece uma melhor compreensão sobre o

desenvolvimento das árvores em uma floresta (BURKHART; TOMÉ, 2012; JIANG et al., 2018; OHEIMB et al., 2011), tornando-se parte importante do manejo florestal, bem como para a ecologia (HUI et al., 2018).

As árvores em uma floresta encontram-se em contínuo estado de competição por recursos limitados, acima do solo (luz), abaixo do solo (água e minerais), ou ambos. Assim, seu desenvolvimento é resultante das seguintes condições: o quanto cada árvore cresce depende do seu tamanho, do tamanho de seus vizinhos e das distâncias desses vizinhos. À medida que as árvores mudam de tamanho, seu crescimento contínuo é influenciado pelos incrementos de tamanho que eles e seus vizinhos já fizeram (SCHNEIDER et al., 2006; VATRAZ et al., 2016)

Nesse sentido, a competição atua como um fator decisivo quanto às taxas de crescimento e conseqüentemente, na produtividade florestal (SABATIA; BURKHART, 2012; SCHNEIDER et al., 2006). Em nível de árvores individuais, usualmente a competição ocasiona a redução do crescimento em termos de diâmetro, diminuição ou estagnação do comprimento da copa e aumenta a probabilidade de mortalidade (WEISKITTEL et al., 2011). Logo, a mesma deve ser quantificada com maior confiabilidade, uma vez que são requeridas para a elaboração de modelos de crescimento e produção (CONTRERAS; AFFLECK; CHUNG, 2011). Esta quantificação consiste em um desafio contínuo na área florestal. Para isto, são utilizados os índices de competição que consistem em formulações matemáticas com o intuito de expressar o quanto cada árvore é afetada por seus vizinhos. Os índices variam desde formulações simples, expressando a posição hierárquica da árvore dentro do povoamento ou parcela, até índices mais complexos que consideram o tamanho, a distância e o número de vizinhos locais (BURKHART; TOMÉ, 2012).

Na tentativa de retratar a competição de forma apropriada são exploradas uma variedade de índices de competição, cuja maioria está concentrada em processos acima do solo (KUEHNE; WEISKITTEL; WASKIEWICZ, 2019). Os mais difundidos na literatura são classificados como: índices dependentes da distância que requerem as coordenadas espaciais de cada árvore e a definição de vizinhança para estabelecer as árvores da vizinhança que competem com a árvore-objeto e os índices independentes da distância que não exigem as coordenadas espaciais das árvores. Além desses, tem-se os semi-independentes da distância que surgiram a posteriori, sendo similares aos índices independentes, mas que são computados, considerando-se parcelas circulares ao redor da árvore-objeto (BURKHART; TOMÉ, 2012; POMMERENING; MEADOR, 2018; STAGE; LEDERMANN, 2008).

Numerosos estudos abordam de maneira comparativa a eficiência entre os diferentes tipos de índices. À exemplo destes estudos, podem ser citados: Ledermann, (2010), Castro et

al. (2014), Lambrecht et al. (2019) e Kuehne, Weiskittel e Waskiewicz (2019). Porém, o desempenho dos índices varia conforme o tipo e as condições da floresta, não havendo unanimidade de um índice superior aos demais (CONTRERAS; AFFLECK; CHUNG, 2011), sendo assim, estes devem ser analisados sob condições específicas de modo a designar sua aplicabilidade (HUI et al., 2018). Diante disso, alguns estudos também têm sido voltados para a melhoria ou a criação de novos índices, como Pedersen et al. (2013) que propôs índices à partir de métricas de varredura a laser, Hui et al. (2018) que formulou o índice baseado em estrutura espacial (*Structure-based Competition Index, SCI*) e Boeck et al. (2014) que utilizou índices fundamentados na competição por luz. Outra abordagem da competição com caráter inovador foi introduzida por meio dos estudos de Nakagawa, Yokozawa e Hara (2016) e Mongus et al. (2018), que utilizaram métricas de redes complexas para investigar a interação entre as árvores. Por se tratar de uma técnica inovadora no campo das ciências florestais, dedicou-se a próxima seção (seção 2.3 Redes complexas) para explicação de conceitos básicos sobre o assunto, que são fundamentais para melhor compreensão do artigo 2.

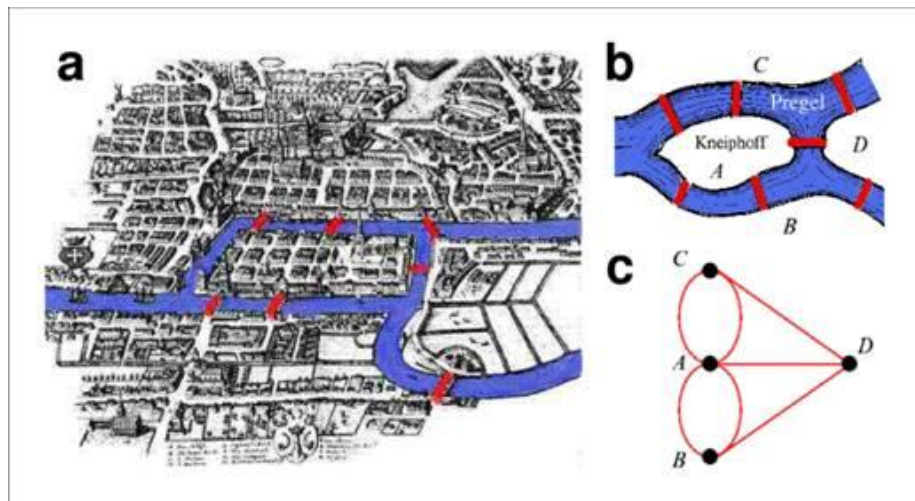
No entanto, ainda restam muitas oportunidades de melhoria, que podem solucionar, inclusive, as limitações dos índices/métricas na incorporação de conceitos de competição como: a competição entre árvores da mesma espécie (intraespecífica), entre árvores de diferentes espécies (interespecífica) (MALEKI; KIVISTE; KORJUS, 2015), quando todas as árvores recebem a mesma quantidade de recursos, independentemente de seus tamanhos (simétrica) ou quando as árvores de maior tamanho tem uma parcela desproporcional de grande parte dos recursos em detrimento das menores (assimétrica) (PRETZSCH; BIBER, 2010). A análise de ambos os conceitos de competição, foram abordados no estudo de (RÍO; CONDÉS; PRETZSCH, 2014) que destacou como sendo uma ferramenta útil para explorar as interações entre espécies.

Desta forma, um único índice/métrica de competição não consegue sintetizar todos os componentes da competição, uma vez que a mesma é compreendida como um processo contínuo, complexo e altamente dinâmico, tanto espacial quanto temporalmente (WEISKITTEL et al., 2011). Isto posto, esforços têm sido direcionados com base no uso do sensoriamento remoto para avaliar o comportamento da competição, como o estudo de Téó et al. (2015) que realizou a espacialização do estresse competitivo em uma Floresta Ombrófila Mista.

2.3 Redes complexas

Historicamente, o estudo das redes iniciou com o desenvolvimento do ramo da matemática discreta, conhecido como teoria dos grafos. O nascimento dessa teoria é atribuído ao matemático suíço Leonhard Euler por ter solucionado, em 1736, o problema das sete pontes de Königsberg que cruzavam o rio Pregel, conectando duas ilhas. O problema consistia em encontrar uma viagem de ida e volta que atravessasse apenas uma única vez cada uma das sete pontes da cidade. Euler estudou o problema utilizando uma abstração matemática denominada de grafo, em que cada parte da cidade era representada por meio de pontos (vértices/nós) e as pontes da cidade eram as ligações (arestas) que conectavam esses pontos. Ele demonstrou que não existia solução para o problema. A falta de solução foi relacionada ao número ímpar de conexões (ligações) dos vértices, assim com estes vértices só seria possível iniciar ou terminar o caminho, sendo necessário passar mais de uma vez por uma mesma ligação (FIGURA 1) (BOCCALETTI et al., 2006; SILVA; ZHAO, 2016; NEWMAN, 2003).

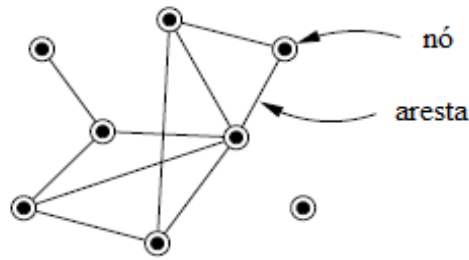
Figura 1- O centro da cidade de Königsberg (a), o mapa esquemático do problema das pontes de Königsberg (b) e sua representação na forma de grafo (c)¹.



¹ Figura retirada de <<http://macsmundi.blogspot.com/2010/09/grafosredes.html>>.

Nesse sentido, uma rede é definida como qualquer sistema passível de representação matemática abstrata na forma de um grafo (FIGURA 2), em que os nós (vértices) identificam os elementos do sistema e a presença de uma relação ou interação entre os elementos são representadas por um conjunto de ligações (arestas) (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

Figura 2 - Representação de uma pequena rede com oito nós e dez arestas.



Fonte: Newman (2003).

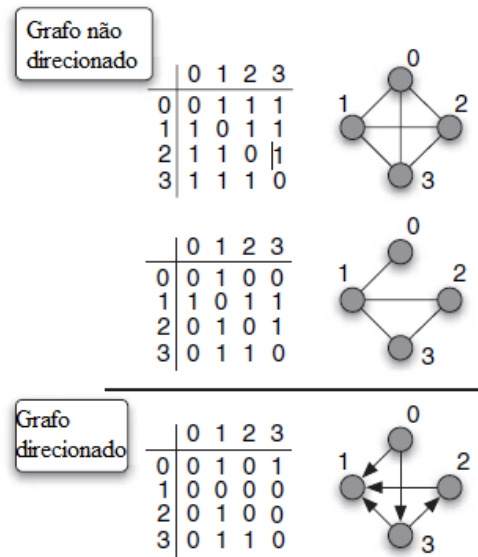
Um grafo G é considerado não direcionado quando estabelecido por um par de conjuntos $G = (V, E)$, em que V é um conjunto não vazio de elementos, denominados de vértices ou nós, e E é um conjunto de pares não ordenados de diferentes nós, denominados de arestas ou links. Diferente deste, em um grafo direcionado (dígrafo), o conjunto E é constituído de pares ordenados de diferentes nós que são chamadas arestas direcionadas, cuja a natureza dirigida das arestas é representada por meio de uma seta, que indica a direção de cada aresta. Desta forma, em um grafo não direcionado, a presença de uma aresta entre os nós i e j conecta os mesmos em ambas as direções. Entretanto, a presença de uma aresta i e j em um grafo direcionado não implica necessariamente a presença da aresta inversa entre j e i (BARRAT; BARTHELEMY; VESPIGNANI, 2008). A aresta (i, j) une os nós i e j , que são referidos como adjacentes ou vizinhos (LATORA; NICOSIA; RUSSO, 2017).

A representação matemática de um grafo é determinada com base na matriz de adjacência $A = \{a_{ij}\}$ por meio da Equação 1. Esta matriz possui dimensão $N \times N$, em que N consiste no número total de nós do grafo.

$$a_{ij} = \begin{cases} 1, & \text{se } (i, j) \text{ são vizinhos } \in A \\ \text{caso contrário, } 0 & \text{se } (i, j) \notin A \end{cases} \quad (1)$$

Sendo assim, os grafos não direcionados são expressos por uma matriz de adjacência simétrica ($a_{ij} = a_{ji}$). Porém, para os grafos direcionados, esta matriz não é simétrica (FIGURA 3) (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

Figura 3 - Diferenciação entre o grafo não direcionado e direcionado com sua matriz de adjacência.



Fonte: Barrat, Barthelemy e Vespignani (2008).

Os grafos também são definidos como não ponderados, assumindo uma natureza binária, em que é estipulada a presença ou ausência de arestas entre os nós. No entanto, diante da ampla heterogeneidade na capacidade e na intensidade das conexões em várias redes reais, os sistemas podem ser melhor descritos em termos de grafos ponderados. Nestes grafos, cada aresta está associada a um peso (valor) numérico real positivo, que representa a intensidade da conexão (BOCCALETTI et al., 2006; LATORA; NICOSIA; RUSSO, 2017). Sua matriz de adjacência consiste em uma matriz quadrada com dimensão $N \times N$, a qual é constituída por um conjunto de pesos $W = \{w_1, w_2, w_3 \dots w_k\}$. Assim, w_{ij} corresponde ao peso (valor) da aresta que conecta o nó i ao nó j , em contrapartida $w_{ij} = 0$ quando os nós i e j não estiverem conectados. Também vale ressaltar que $w_{ii} = 0 \forall i$.

Em geral, a representação da estrutura de interesse como uma rede é parte do processo de investigação em redes complexas, que também envolve uma análise das características topológicas da representação obtida, produzida em termos de um conjunto de medidas informativas. As medidas de rede são compreendidas como um recurso direto ou subsidiário em muitas investigações de rede, potencialmente utilizadas para representar, caracterizar, classificar e modelar sistemas complexos compostos por elementos de interação. Diante disto, nos últimos anos, as pesquisas em redes complexas têm ganhado relevância, sendo aplicada em diversas áreas como biologia, economia, linguística, medicina, ciências sociais, tecnologia e

transporte, o que lhe conferiu um caráter multidisciplinar (COSTA ET AL., 2007; COSTA; RODRIGUES; CRISTINO, 2008).

Existem uma variedade de métricas que podem ser utilizadas para a caracterização topológica das redes. A exemplo de algumas medidas básicas podem ser citadas: o grau do nó que é determinado pelo número de arestas conectadas ao nó, o coeficiente de agrupamento (*Clustering coefficient*) que mede a presença de ciclos de ordem 3 (triângulos), o tamanho do caminho que consiste no número de arestas em um caminho que conecta os vértices i e j , com destaque para o menor caminho entre os vértices i e j , que é definido como o menor número de ligações existentes entre i e j (COSTA et al., 2007), centralidade de intermediação (*Betweenness centrality*), centralidade de autovetor (*Eigenvector centrality*) e PageRank (LIAO et al., 2017). Tais medidas estão relacionadas com a análise disposta no artigo 2, seção 2.2.1, onde serão detalhadas. Estas concentram-se em medidas de análise individual do nó, sendo sua importância definida com base na característica a ser avaliada.

2.4 Random Forest

O *Random Forest* (RF), foi introduzido por Breiman (2001), como uma melhoria em relação a árvore de decisão, por meio do uso de um comitê ou "*ensemble*" de árvores de decisão, portanto "floresta", que foram descorrelacionadas de modo a proporcionar a redução da variância e conseqüentemente, um aumento do desempenho preditivo (CHENG et al., 2019; JAMES et al., 2013).

O funcionamento do RF consiste na combinação de árvores de decisão, em que cada árvore é construída com base em um vetor aleatório amostrado de forma independente em relação aos outros vetores, porém com a mesma distribuição para todas as árvores da floresta. Este vetor refere-se às amostras dentro da bolsa (do inglês, *in bag*) que constituem 2/3 do conjunto de dados original utilizadas para o treinamento de cada árvore de decisão. O 1/3 dos dados remanescentes consistem nas amostras fora da bolsa (do inglês, *out-of-bag* – OOB) que compõe o conjunto teste, utilizado como técnica de validação-cruzada interna do RF (AURET; ALDRICH, 2012; BELGIU; DRĂGU, 2016; BREIMAN, 2001).

As árvores de decisão são cultivadas por meio de divisões no conjunto de dados, em que cada divisão é realizada a partir da seleção aleatória de um subconjunto de variáveis preditoras (m) menor que o número total de variáveis disponíveis (p). Assim, são geradas árvores distintas e a agregação das predições das mesmas é realizada para obter a predição final, que é

resultante da média (para regressão) ou do voto majoritário (para classificação) (BREIMAN, 2001; AURET; ALDRICH, 2012; BELGIU; DRĂGU, 2016).

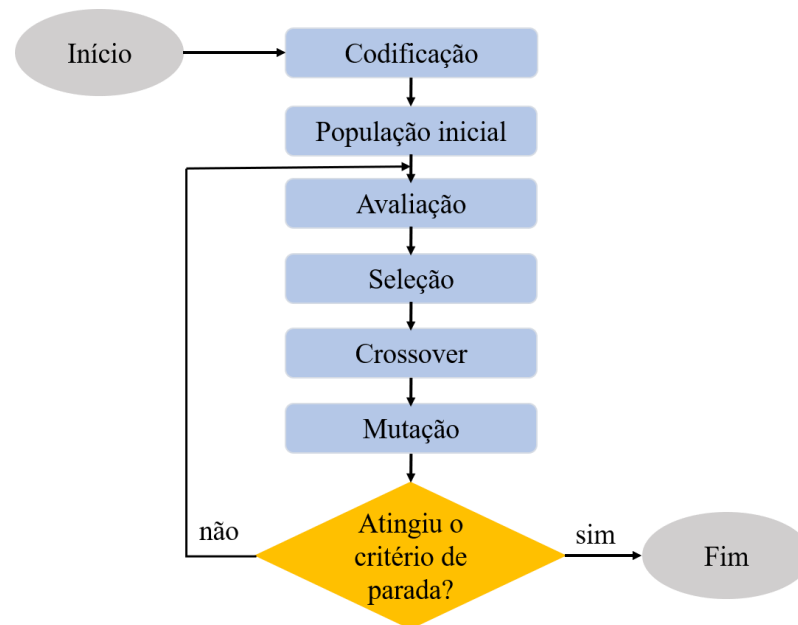
O ajuste do RF demanda a definição de três parâmetros, sendo estes: o número de árvores a serem cultivadas (*ntree*) com base na quantidade de amostras *bootstrap* oriundas do conjunto de dados original, o número de preditores *m* a ser selecionado aleatoriamente para divisão de cada nó da árvore (*mtry*) e o número mínimo de observações nos nós terminais de cada árvore (*nodesize*) (LAHSSINI et al., 2015).

A popularidade do RF como uma ferramenta excepcional na resolução de problemas de regressão e classificação nas mais diversas áreas científicas, pode ser atribuída a facilidade de ajuste de seus parâmetros, o alto desempenho preditivo, o ranqueamento das variáveis conforme suas medidas de importância e sua capacidade de lidar com valores faltantes, ruídos e alta dimensão de dados (AURET; ALDRICH, 2012; HAPFELMEIER; ULM, 2014; JANITZA et al., 2016; LI et al., 2018).

2.5 Algoritmo genético (AG)

O algoritmo genético (AG) tem sido utilizado extensivamente e com êxito na solução de problemas de otimização. O mesmo foi introduzido na década de 70 por John Holland e colaboradores da Universidade de Michigan como uma heurística inspirada na teoria da evolução darwiniana que tem como princípio “a sobrevivência do mais apto” (PATTANAIK; BASU; DASH, 2018) por meio dos mecanismos de seleção natural, herança e variabilidade (HONDA, 2018). Portanto, seu método de busca pela solução ideal é realizado com base na simulação dos mecanismos supracitados (ZHI; LIU, 2019). Logo, o funcionamento do AG pode ser descrito conforme o fluxograma disposto na Figura 4.

Figura 4 - Fluxograma das etapas de funcionamento do Algoritmo genético simples.



Fonte: Adaptado de Abdulhamed, Tawfeek e Keshk (2018).

Inicialmente,   necess rio codificar uma solu o candidata adequada ao problema, por meio de uma representa o cromoss mica. O cromossomo (indiv duo) consiste em um vetor de comprimento definido, cuja codifica o mais usual   de natureza bin ria, ou seja, cada gene assume o valor 0 ou 1 que determina as informa es de cada indiv duo (ABDULHAMED; TAWFEEK; KESHK, 2018).

A partir da codifica o, s o gerados indiv duos de forma aleat ria que constituem na popula o inicial. Cada indiv duo   submetido a avalia o, em que seu desempenho   mensurado por meio de uma fun o objetivo, denominada de *fitness* (aptid o) (PATTANAIAK; BASU; DASH, 2018). Desta forma, os indiv duos s o comparados buscando, dentro da popula o corrente, aqueles que s o mais aptos para participar da forma o de novos indiv duos. Esta opera o   definida como sele o, podendo ser realizada por meio de m todos distintos como: torneio ou roleta. A sele o por torneio   comumente utilizada devido a sua efici ncia e simplicidade na implementa o de problemas de maximiza o ou minimiza o (MIASAKI; ROMERO, 2007).

Posterior a sele o dos indiv duos,   aplicado o operador gen tico crossover que visa a gera o de melhores indiv duos. Para isto, s o escolhidos um par de indiv duos (pais) que tem seus genes alternados de modo a produzir dois indiv duos diferentes. Em seguida,   aplicado outro operador, denominado de muta o que permite escapar de  timos locais, uma vez que a

geração de um novo indivíduo é conduzida a partir da alteração, de modo aleatório, de valores de um ou mais genes em um indivíduo. Em geral, adota-se uma probabilidade de mutação pequena, diferente da probabilidade de crossover (cruzamento) que é próxima de 1 (ABDULHAMED; TAWFEEK; KESHK, 2018; METAWA; HASSAN; ELHOSENY, 2017). Estas são consideradas estratégias com a finalidade de diversificar a população, bem como melhorar o *fitness* dos indivíduos da nova população (CERRADA et al., 2015). O procedimento é repetido desde a etapa de avaliação até a mutação, finalizando somente quando o critério de parada é satisfeito. Este pode ser estabelecido como um número máximo de iterações (MIASAKI; ROMERO, 2007).

2.6 Seleção de variáveis

O aumento exponencial na quantidade de dados disponíveis tem configurado um desafio para extração de conhecimento, sendo substancial uma etapa de pré-processamento de dados em problemas de regressão e classificação. A seleção de atributos (*feature selection*) consiste no pré-processamento com base na redução da dimensionalidade dos dados, sendo definido um subconjunto de atributos relevantes a partir de um conjunto original. Esta etapa visa a remoção de atributos irrelevantes, redundantes e ruídos que podem conduzir a uma baixa interpretabilidade e desempenho do modelo (MASOUDI-SOBHANZADEH, MOTIEGHADER; MASOUDI-NEJAD, 2019; MIAO; NIU, 2016).

Na literatura são difundidas diferentes abordagens de seleção de atributos que variam desde métodos clássicos a algoritmos de *Machine learning*. No tocante aos métodos clássicos tem-se o *stepwise* que seleciona as variáveis independentes a serem incorporadas em um modelo de regressão. Esta seleção consiste em um processo iterativo por meio da adição (passo *forward*) ou remoção (passo *backward*) de variáveis com base no critério de seleção adotado. Os critérios mais usuais são o teste F, coeficiente de correlação linear múltipla, erro quadrático total e critério de informação de Akaike (ALVES; LOTUFO; LOPES, 2013). À exemplo de estudos florestais, Li et al. (2014) evidenciaram que a aplicação do *stepwise* na seleção de atributos para a construção de modelos de regressão OLS (*Ordinary least squares*) e GAM (*Generalized additive model*) favoreceram a acurácia das estimativas de biomassa e carbono em uma floresta mista. A regressão PLS (*Partial least squares*) é uma análise multivariada fundamentada na transformação das variáveis de processo x e de produto y em um número reduzido de combinações lineares, sendo uma opção plausível em face de sua capacidade de

lidar com grande quantidade de variáveis correlacionadas e ruído (ZIMMER; ANZANELLO, 2014). A sua aplicação tem-se estendido no meio florestal com êxito na predição de atributos florestais, como a altura média de Lorey a partir de dados LiDAR (LIU et al., 2018) e a biomassa acima do solo a partir de dados LiDAR e hiperespectrais (LAURIN et al., 2014). Uma abordagem popularmente difundida é o método de regressão denominado de Lasso (*Least absolute shrinkage and selection operator*) que minimiza a soma dos quadrados dos resíduos por meio da penalização L_1 aos coeficientes do modelo (KUKREJA; LÖFBERG; BRENNER, 2006). Este procedimento induz a redução dos coeficientes, em que alguns coeficientes assumem o valor zero podendo ser removidos do modelo juntamente com os coeficientes negativos. Desta forma, são realizadas simultaneamente a regularização dos coeficientes e a seleção de atributos (GHOSH et al., 2021; YAN; YAO, 2015). Kankare et al. (2013) obtiveram maior acurácia das estimativas de biomassa em nível de árvore individual a partir de dados ALS (*Airborne laser scanning*) com seleção de atributos via Lasso comparado aos modelos clássicos baseados em diâmetro e altura derivados de ALS.

A seleção de atributos é compreendida como um problema de otimização combinatória face a inviabilidade de testar todas as soluções possíveis. Neste sentido, maiores avanços nas técnicas de seleção envolvem o uso de meta-heurísticas na otimização do processo de busca por um subconjunto de atributos. Os algoritmos evolucionários baseados em população são capazes de encontrar boas soluções sem a necessidade de explorar todo o espaço de busca. À exemplo desses tem-se o algoritmo genético (*Genetic algorithm* – GA), o algoritmo de otimização por colônia de formigas (*Ant Colony Optimization* – ACO) e por enxame de partículas (*Particle Swarm Optimization* – PSO) (HONG et al., 2018; VIEIRA et al., 2012; YANG; OLAFSSON, 2006). Estes algoritmos são utilizados como métodos *wrapper*, cujas soluções são avaliadas em termos de acurácia com base no desempenho de um preditor como o *Random Forest*, máquina de vetor de suporte e redes neurais artificiais (HONG et al., 2018; PATEL; GIRI, 2016). Estudos têm comprovado ganhos em termos de acurácia e interpretabilidade na modelagem da susceptibilidade a incêndios florestais (HONG et al., 2018), altura de árvores (MIRANDA et al., 2022) e estoque de carbono orgânico no solo (WANG et al., 2018).

3 CONSIDERAÇÕES GERAIS

A inteligência artificial tem sido aplicada em diversos problemas florestais, constituindo uma ferramenta importante no fornecimento de informações assertivas para um adequado manejo e planejamento florestal. Dentre suas diversas aplicações, a tarefa de seleção de atributos permeia um campo multidisciplinar com êxito na extração de conhecimento dos dados favorecendo a obtenção de modelos simplificados com boa capacidade preditiva e explicativa. Portanto, a seleção de atributos surge como uma abordagem potencial a ser explorada, face a sua incipiência na modelagem florestal. Esse contexto impulsiona o deslocamento de um panorama puramente estatístico fundamentado em acurácia para a integração da propriedade interpretativa aos modelos em termos biológicos e ecológicos. Tal interpretabilidade advém da tentativa de explicar os efeitos subjacentes ao subconjunto de variáveis otimizado. A relevância da interpretabilidade dos modelos concernentes a produtividade da colheita florestal (Artigo 1) e ao crescimento em nível de árvores individuais (Artigo 2) é atribuída a sua atuação decisiva na confiabilidade, agilidade e suporte as tomadas de decisão pelos gestores florestais.

REFERÊNCIAS

- ABDULHAMED, A. A.; TAWFEEK, M. A.; KESHK, A. E. A genetic algorithm for service flow management with budget constraint in heterogeneous computing. **Future Computing and Informatics Journal**, New Cairo, v. 3, n. 2, p. 341–347, 2018.
- ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Applied and Computational Mathematics**, São Carlos, v. 1, n. 1, p. 1–6, 2013.
- ASHRAF, I. M. et al. Integrating biophysical controls in forest growth and yield predictions with artificial intelligence technology. **Canadian Journal of Forest Research**, Ottawa, v. 43, p. 1162–1171, 2013.
- ASHRAF, M. I. et al. A novel modelling approach for predicting forest growth and yield under climate change. **PLoS ONE**, San Francisco, v. 10, n. 7, p. 1–18, 2015.
- AURET, L.; ALDRICH, C. Interpretation of nonlinear relationships between process variables by use of random forests. **Minerals Engineering**, Falmouth, v. 35, p. 27–42, 2012.
- BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A. **Dynamical Processes on Complex Networks**. Cambridge: Cambridge University Press, 2008.
- BELGIU, M.; DRĂGU, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 114, p. 24–31, 2016.
- BISPO, P. da C. et al. Woody aboveground biomass mapping of the brazilian savanna with a multi-sensor and machine learning approach. **Remote Sensing**, Basel, v. 12, n. 2685, p. 1–22, 2020.
- BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics Reports**, Amsterdam, v. 424, n. 4–5, p. 175–308, 2006.
- BOECK, A. et al. Predicting tree mortality for European beech in southern Germany using spatially explicit competition indices. **Forest Science**, Bethesda, v. 60, n. 4, p. 613–622, 2014.
- BRACKENRIDGE, R. E. et al. Improving Subsurface Characterisation with ‘Big Data’ Mining and Machine Learning. **Energies**, Basel, v. 15, n. 1070, p. 1–23, 2022.
- BREIMAN, L. Random Forests. **Machine learning**, Dordrecht, v. 45, p. 5–32, 2001.
- BURKHART, H. E.; TOMÉ, M. **Modeling forest trees and stands**. Dordrecht: Springer

Science & Business Media, 2012.

CAI, J. et al. Feature selection in machine learning: A new perspective. **Neurocomputing**, Amsterdam, v. 300, p. 70–79, 2018.

CASTRO, R. V. O. et al. Competition indices in individual tree level in a Semideciduous Montana forest. **Silva Lusitana**, Oeiras, v. 22, n. 1, p. 43–66, 2014.

CERRADA, M. et al. Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. **Sensors**, Basel, v. 15, n. 9, p. 23903–23926, 2015.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, United Kingdom, v. 40, p. 16–28, 2014.

CHENG, L. et al. Applying a random forest method approach to model travel mode choice behavior. **Travel Behaviour and Society**, Netherlands, v. 14, p. 1–10, 2019.

CONTRERAS, M. A.; AFFLECK, D.; CHUNG, W. Evaluating tree competition indices as predictors of basal area increment in western Montana forests. **Forest Ecology and Management**, Amsterdam, v. 262, n. 11, p. 1939–1949, 2011.

CORTE, A. P. D. et al. Forest inventory with high-density UAV-Lidar: Machine learning approaches for predicting individual tree attributes. **Computers and Electronics in Agriculture**, Oxford, v. 179, p. 105815, 2020.

COSTA, L. D. F. et al. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Abingdon, v. 56, n. 1, p. 167–242, 2007.

COSTA, L. da F.; RODRIGUES, F. A.; CRISTINO, A. S. Complex networks: The key to systems biology. **Genetics and Molecular Biology**, Ribeirão Preto, v. 31, n. 3, p. 591–601, 2008.

DING, Y. et al. Intraspecific trait variation and neighborhood competition drive community dynamics in an old-growth spruce forest in northwest China. **Science of the Total Environment**, Amsterdam, v. 678, p. 525–532, 2019.

ERIKSSON, M.; LINDROOS, O. Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. **International Journal of Forest Engineering**, Philadelphia, v. 25, n. 3, p. 179–200, 2 set. 2014.

FENG, Y.; AUDY, J. F. Forestry 4.0: A framework for the forest supply chain toward Industry 4.0. **Gestão e Produção**, São Carlos, v. 27, n. 4, p. 1–21, 2020.

GHAFFARIYAN, M. R.; BROWN, M. Selecting the efficient harvesting method using multiple-criteria analysis: A case study in south-west Western Australia. **Journal of Forest Science**, Prague, v. 59, n. 12, p. 479–486, 2013.

- GHOSH, P. et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. **IEEE Access**, Piscataway, v. 9, p. 19304–19326, 2021.
- GONÇALVES, S. B. et al. Machine learning techniques to estimate mechanised forest cutting productivity. **Southern Forests - A Journal of Forest Science**, Grahamstown, Latest Articles, p. 1–8, 2022. doi: 10.2989/20702620.2021.1994342
- HAMIDI, S. K. et al. Analysis of plot-level volume increment models developed from machine learning methods applied to an uneven-aged mixed forest. **Annals of Forest Science**, Les Ulis, v. 78, n. 4, p. 1–16, 2021.
- HAPFELMEIER, A.; ULM, K. Variable selection by Random Forests using data with missing values. **Computational Statistics & Data Analysis**, Amsterdam, v. 80, p. 129–139, 2014.
- HIESL, P.; BENJAMIN, J. G. Applicability of international harvesting equipment productivity studies in Maine, USA: A literature review. **Forests**, Basel, v. 4, n. 4, p. 898–921, 2013.
- HONDA, M. Application of genetic algorithms to modelings of fusion plasma physics. **Computer Physics Communications**, Amsterdam, v. 231, p. 94–106, 2018.
- HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, Amsterdam, v. 630, p. 1044–1056, 2018.
- HUI, G. et al. A novel approach for assessing the neighborhood competition in two different aged forests. **Forest Ecology and Management**, Amsterdam, v. 422, p. 49–58, 2018.
- JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013.
- JANITZA, S.; TUTZ, G.; BOULESTEIX, A. L. Random forest for ordinal responses: Prediction and variable selection. **Computational Statistics & Data Analysis**, Amsterdam, v. 96, p. 57–73, 2016.
- JIANG, S. et al. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. **Expert Systems with Applications**, Oxford, v. 82, p. 216–230, 2017.
- JIANG, X. et al. Interspecific variation in growth responses to tree size, competition and climate of western Canadian boreal mixed forests. **Science of the Total Environment**, Amsterdam, v. 631–632, p. 1070–1078, 2018.
- KANKARE, V. et al. Single tree biomass modelling using airborne laser scanning. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 85, p. 66–73, 2013.

- KHALID, S.; KHALIL, T.; NASREEN, S. A survey of feature selection and feature extraction techniques in machine learning. **Science and Information Conference**, London, v. 2014, p.27-29, 2014
- KUEHNE, C.; WEISKITTEL, A. R.; WASKIEWICZ, J. Comparing performance of contrasting distance-independent and distance-dependent competition metrics in predicting individual tree diameter increment and survival within structurally-heterogeneous, mixed-species forests of Northeastern United States. **Forest Ecology and Management**, Amsterdam, v. 433, p. 205–216, 2019.
- KUKREJA, S. L.; LÖFBERG, J.; BRENNER, M. J. A Least Absolute Shrinkage and Selection Operator (Lasso) for Nonlinear System Identification. **IFAC Proceedings Volumes**, Newcastle, v. 39, n. 1, p. 814–819, 2006.
- LACERDA, T. H. S. et al. Feature selection by genetic algorithm in nonlinear taper model. **Canadian Journal of Forest Research**, Ottawa, Just-IN version, 2022.
- LAHSSINI, S. et al. Predicting Cork Oak Suitability in Maâmora Forest Using Random Forest Algorithm. **Journal of Geographic Information System**, [s.l.], v. 07, p. 202–210, 2015.
- LAMBRECHT, F. R. et al. Competição em floresta natural de araucária na região noroeste do Rio Grande do Sul-Brasil. **Scientia Forestalis**, Piracicaba, v. 47, n. 121, p. 131–138, 2019.
- LATORA, V.; NICOSIA, V.; RUSSO, G. **Complex networks: Principles, Methods and Applications**. Cambridge: Cambridge University Press, 2017.
- LAURIN, G. V. et al. Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 89, p. 49–58, 2014.
- LEDERMANN, T. Evaluating the performance of semi-distance-independent competition indices in predicting the basal area growth of individual trees. **Canadian Journal of Forest Research**, Ottawa, v. 40, p. 796–805, 2010.
- LEVERS, C. et al. Drivers of forest harvesting intensity patterns in Europe. **Forest Ecology and Management**, Amsterdam, v. 315, p. 160–172, 2014.
- LI, J. et al. Feature Selection: A Data Perspective. **ACM Computing Surveys**, New York, v. 50, n. 6, p. 1–45, 30 nov. 2018.
- LI, M. et al. Forest biomass and carbon stock quantification using airborne LiDAR data: A case study over Huntington wildlife forest in the Adirondack Park. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Piscataway, v. 7, n. 7, p. 3143–3156, 2014.
- LI, Y. et al. Random forest regression for online capacity estimation of lithium-ion batteries. **Applied Energy**, Oxford, v. 232, p. 197–210, 2018.

LIAO, H. et al. Ranking in evolving complex networks. **Physics Reports**, Amsterdam, v. 689, p. 1–54, 2017.

LISKI, E. et al. Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods. **International Journal of Forest Engineering**, Philadelphia, v. 31, n. 3, p. 253–262, 2020.

LIU, K. et al. Estimating forest structural attributes using UAV-LiDAR data in Ginkgo plantations. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 146, p. 465–482, 2018.

LIU, Z. et al. Application of machine-learning methods in forest ecology: Recent progress and future challenges. **Environmental Reviews**, Ottawa, v. 26, n. 4, p. 339–350, 2018.

MALEKI, K.; KIVISTE, A.; KORJUS, H. Analysis of individual tree competition effect on diameter growth of silver birch in Estonia. **Forest Systems**, Madrid, v. 24, n. 2, p. 1–13, 2015.

MASOUDI-SOBHANZADEH, Y.; MOTIEGHADER, H.; MASOUDI-NEJAD, A. FeatureSelect: A software for feature selection based on machine learning approaches. **BMC Bioinformatics**, London, v. 20, n. 170, p. 1–17, 2019.

METAWA, N.; HASSAN, M. K.; ELHOSENY, M. Genetic algorithm based model for optimizing bank lending decisions. **Expert Systems with Applications**, Oxford, v. 80, p. 75–82, 2017.

MIAO, J.; NIU, L. A Survey on Feature Selection. **Procedia Computer Science**, Amsterdam, v. 91, p. 919–926, 2016.

MIASAKI, C. T.; ROMERO, R. Um algoritmo genético especializado aplicado ao planejamento da expansão do sistema de transmissão com alocação de dispositivos de compensação série. **Revista SBA - Controle & Automação**, Campinas, v. 18, n. 2, p. 210–222, 2007.

MIRANDA, E. N. et al. Variable selection for estimating individual tree height using genetic algorithm and random forest. **Forest Ecology and Management**, Amsterdam, v. 504, p. 119828, 2022.

MONGUS, D. et al. Predictive analytics of tree growth based on complex networks of tree competition. **Forest Ecology and Management**, Amsterdam, v. 425, p. 164–176, 2018.

MURTHY, Y. V. S.; KOOLAGUDI, S. G. Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS). **Expert Systems with Applications**, Oxford, v. 106, p. 77–91, 2018.

NAKAGAWA, Y.; YOKOZAWA, M.; HARA, T. Complex network analysis reveals novel essential properties of competition among individuals in an even-aged plant population. **Ecological Complexity**, Amsterdam, v. 26, p. 95–116, 2016.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, Philadelphia, v. 45, n. 2, p. 167–256, 2003.

NUNES, M. H.; GÖRGENS, E. B. Artificial Intelligence Procedures for Tree Taper Estimation within a Complex Vegetation Mosaic in Brazil. **PLoS ONE**, San Francisco, v. 11, n. 5, p. 1–16, 2016.

OHEIMB, G. VON et al. Individual-tree radial growth in a subtropical broad-leaved forest: The role of local neighbourhood competition. **Forest Ecology and Management**, Amsterdam, v. 261, p. 499–507, 2011.

OU, Q.; LEI, X.; SHEN, C. Individual tree diameter growth models of Larch-Spruce-Fir mixed forests based on machine learning algorithms. **Forests**, Basel, v. 10, n. 187, p. 1–20, 2019.

PATEL, R. K.; GIRI, V. K. Feature selection and classification of mechanical fault of an induction motor using random forest classifier. **Perspectives in Science**, [s.l.], v. 8, p. 334–337, 2016.

PATTANAİK, J. K.; BASU, M.; DASH, D. P. Improved real coded genetic algorithm for dynamic economic dispatch. **Journal of Electrical Systems and Information Technology**, United Kingdom, v. 5, p. 349–362, 2018.

PEDERSEN, R. Ø. et al. On the evaluation of competition indices - The problem of overlapping samples. **Forest Ecology and Management**, Amsterdam, v. 310, p. 120–133, 2013.

POMMERENING, A.; SÁNCHEZ MEADOR, A. J. Tamm review: Tree interactions between myth and reality. **Forest Ecology and Management**, Amsterdam, v. 424, p. 164–176, set. 2018.

PRETZSCH, H. **Forest Dynamics, Growth and Yield**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

PRETZSCH, H.; BIBER, P. Size-symmetric versus size-asymmetric competition and growth partitioning among trees in forest stands along an ecological gradient in central Europe. **Canadian Journal of Forest Research**, Ottawa, v. 40, p. 370–384, 2010.

RÍO, M. del; CONDÉS, S.; PRETZSCH, H. Analyzing size-symmetric vs. size-asymmetric and intra- vs. inter-specific competition in beech (*Fagus sylvatica* L.) mixed stands. **Forest Ecology and Management**, Amsterdam, v. 325, p. 90–98, 2014.

ROSSIT, D. A. et al. A Big Data approach to forestry harvesting productivity. **Computers and Electronics in Agriculture**, Oxford, v. 161, p. 29–52, 2019.

SABATIA, C. O.; BURKHART, H. E. Competition among loblolly pine trees: Does genetic variability of the trees in a stand matter? **Forest Ecology and Management**, Amsterdam, v. 263, p. 122–130, 2012.

SAFARI, A. et al. A comparative assessment of multi-temporal Landsat 8 and machine learning algorithms for estimating aboveground carbon stock in coppice oak forests. **International Journal of Remote Sensing**, Abingdon, v. 38, n. 22, p. 6407–6432, 2017.

SCHNEIDER, M. K.; LAW, R.; ILLIAN, J. B. Quantification of neighbourhood-dependent plant growth by Bayesian hierarchical modelling. **Journal of Ecology**, Malden, v. 94, n. 2, p. 310–321, 2006.

SHAO, Y.; LUNETTA, R. S. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 70, p. 78–87, 2012.

SILVA, J. P. M. et al. Prognosis of forest production using machine learning techniques. **Information Processing in Agriculture**, Beijing, in press, 2021. doi: <https://doi.org/10.1016/j.inpa.2021.09.004>

SILVA, T. C.; ZHAO, L. **Machine Learning in Complex Networks**. Cham: Springer International Publishing, 2016.

SILVEIRA, E. M. O. et al. Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment. **International Journal of Applied Earth Observation and Geoinformation**, Amsterdam, v. 78, p. 175–188, 2019.

SOMAN, H.; KIZHA, A. R.; ROTH, B. E. Impacts of silvicultural prescriptions and implementation of best management practices on timber harvesting costs. **International Journal of Forest Engineering**, Philadelphia, v. 30, n. 1, p. 14–25, 2019.

STAGE, A. R.; LEDERMANN, T. Effects of competitor spacing in a new class of individual-tree indices of competition: semi-distance-independent indices computed for Bitterlich versus fixed-area plots. **Canadian Journal of Forest Research**, Ottawa, v. 38, n. 4, p. 890–898, 2008.

TAVARES JÚNIOR, I. da S. et al. Machine learning: Modeling increment in diameter of individual trees on Atlantic Forest fragments. **Ecological Indicators**, Amsterdam, v. 117, p. 106685, out. 2020.

TÉO, S. J.; FILHO, A. F.; LINGNAU, C. Análise espacial do estresse competitivo, incremento diamétrico e estrutura de uma floresta ombrófila mista, Irati, PR. **Floresta**, Curitiba, v. 45, n. 4, p. 681–694, 2015.

VATRAZ, S.; ALDER, D.; SILVA, J. N. M. Índices de competição dependentes da distância do estrato arbóreo na Amazônia brasileira. **Revista Espacios**, Caracas, v. 37, n. 27, p. 1–12, 2016.

VIEIRA, G. C. et al. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment**, Amsterdam, v. 619–620, p. 1473–1481, 2018.

VIEIRA, S. M. et al. Metaheuristics for feature selection: application to sepsis outcome prediction. **IEEE World Congress on Computational Intelligence**, Brisbane, p. 1–8, 2012.

WANG, B. et al. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. **Ecological Indicators**, Amsterdam, v. 88, p. 425–438, 2018.

WEISKITTEL, A. R. et al. **Forest Growth and Yield Modeling**. Chichester, UK: John Wiley & Sons, Ltd, 2011.

XUE, B. et al. A Survey on Evolutionary Computation Approaches to Feature Selection. **IEEE Transactions on Evolutionary Computation**, Piscataway, v. 20, n. 4, p. 606–626, 2016.

YAN, Z.; YAO, Y. Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v. 146, p. 136–146, 2015.

YANG, J.; OLAFSSON, S. Optimization-based feature selection with adaptive instance sampling. **Computers & Operations Research**, Oxford, v. 33, n. 11, p. 3088–3106, 2006.

ZHI, H.; LIU, S. Face recognition based on genetic algorithm. **Journal of Visual Communication and Image Representation**, San Diego, v. 58, p. 495–502, 2019.

ZIMMER, J.; ANZANELLO, M. J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Production**, Porto Alegre, v. 24, n. 1, p. 84–93, 2014.

ZOU, W. et al. A Survey of Big Data Analytics for Smart Forestry. **IEEE Access**, Piscataway, v. 7, p. 46621–46636, 2019.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 – A comparative approach of methods to estimate machine productivity in wood cutting

Isáira Leite e Lopes^{a*}, Laís Almeida Araújo^a, Evandro Nunes Miranda^a, Thomaz Aurelio Bastos^a, Lucas Rezende Gomide^a, Gustavo Pereira Castro^b.

^aFederal University of Lavras, Department of Forest Sciences, Lavras, Minas Gerais, Brazil,

^bFederal University of Paraná, Department of Forest Sciences, Curitiba, Paraná, Brazil.

*Corresponding author: isairaleite2010@gmail.com

Abstract

Forest harvesting planning requires careful analysis of the variables that influence machine productivity. This information is crucial for better decision-making. Thus, we aimed to compare models for predicting the excavator-based grapple saw productivity in wood cutting with variables from environmental data, forest inventory, and operator records. We applied Stepwise linear regression, Random Forest (RF), and Artificial Neural Networks (ANN) to estimate machine productivity (mp). Hybrid methods were also designed to perform the feature selection procedure. A Genetic algorithm (GA) was combined with RF (GA-RF), and ANN (GA-ANN). These methods were assessed according to error metrics and accuracy. Although the order of the variables' importance changed based on these methods, the operator's experience was the main factor in the mp behavior, regardless of the model. The work shift impacted the machine productivity, but not as significantly as the operator's experience. The mean individual tree volume and precipitation also made a considerable contribution to the mp estimates of the GA-RF and GA-ANN models, respectively. Our findings indicate that the RF and GA-RF methods perform best and with high accuracy to estimate mp. Furthermore, we highlight that GA-RF performed a robust selection of the variables that really influenced the mp behavior.

Keywords: Machine learning; forest harvesting; feature selection; forest management; genetic algorithm.

Introduction

Global changes in forest industry have forced forest managers to find better models to reduce investment risks. Improvements in current trends have affected our perception, and new technologies are being designed to support the decision-making process. The forest industry has been investing in big data technologies to model and integrate their entire process into a better economy and sustainable principles. Parallel to the development of machines, the accuracy of field information or its prediction has a high aggregate value and is attracting investment. However, greater statistical and computational efforts are desirable to guarantee an accurate plan of the wood supply chain. This may be a useful outcome for the demand of the wood supply market (She, Chung and Kim, 2018; Tolan and Visser, 2015), and for the control of forest harvesting, which has critical operations with high investment costs (Soman, Kizha and Roth, 2019). Nevertheless, harvest organization and control process face technical and socio-economic challenges without reliable field information or predictions. Machine productivity modeling plays a key role in this process, being a complex task since it has several variables (species, tree size, terrain surface, site conditions, management objectives, machines performance, and operator experience) that affect its behavior (Hiesl and Benjamin, 2013; Silayo and Migunga, 2014).

Today, advances in the acquisition of data at a reasonable cost and high precision make a significant contribution to any area of scientific knowledge (She, Chung, and Kim 2018), and, as a result, there is a great deal of information available to correlate with our researches. Hence, the modeling process requires an optimal method to deal with large or multidimensional sets of variables (Strandgard, Mitchell and Acuna, 2016). The advantage of using powerful models is to enable greater accuracy in guessing unknown or extreme values (Rossit et al., 2019) and highlighting new patterns of variables that can be used to explain events (Hong et al., 2018). Currently, the modeling of several forest management problems using machine learning methods has been increasingly applied. It means that these methods have been achieving high performance. Most studies have highlighted them as promising models to work with a huge range of variables (Ou, Lei and Shen, 2019), such as in tree diameter growth (Vieira et al., 2018), volumetric prediction (Dantas et al., 2020), aboveground biomass (Silva et al., 2019), carbon stock (Safari et al., 2017), and tropical selective log mapping (Hethcoat et al., 2019).

Artificial neural networks (ANN) and Random Forest (RF) have been widely applied machine learning methods over the last decades. Usually, they denote a robust and unbiased method with high accuracy (Rossit et al., 2019; Xing et al., 2019), working with any set of

variable types (Freitas et al., 2020). Nevertheless, the variables heterogeneity is a problem when modeling any complex system, as it causes the risk of inaccurate prediction (Rossit et al., 2019). The choice of non-parametric learning methods may overcome this problem when a high performance is required (Das, Das and Ghosh, 2017). First, the accuracy should be improved using any feature selection technique to reduce the complexity of a task. This step decreases the number of irrelevant and redundant variables (Hong et al., 2018). Apolloni, Leguizamón, and Alba (2016) define strategies to reduce the database size and dimensions. They present three methods (filter, wrapper, embedded) with high potential use, but finally suggest the wrapper method since, it presents excellent results. However, it requires high computational effort, as it must perform interactions between variables several times with different subsets (Mafarja and Mirjalili 2018; Ghosh et al. 2020). On the other hand, the meta-heuristics use for searching deeply for an optimal set of variables may accelerate the final convergence of this response, giving a high performance (Hong et al. 2018; Rossit et al. 2019). Hence, combining these algorithms with machine learning methods may reduce the prediction errors. This may be particularly important to model complex systems as observed in machine productivity performing forest harvesting tasks. Conversely, the robustness of the variable selection method is still an open problem. Thus, efforts must be made to fill this gap in forest science. The study aimed to compare modeling strategies of an excavator-based grapple saw's productivity in wood cutting with several variable types. The study investigates the most suitable procedure, taking into account the linear regression model, machine learning methods, and hybrid algorithms using the genetic algorithm to perform feature selection. Furthermore, it is also timely to reflect on the influence of variables on predicting machine productivity through the performance of the proposed methods. Therefore, this study was addressed to answer the following research questions: a) Which modeling strategies have the best predictive performance? and b) What are the variables that most influence machine productivity?

Materials and methods

Experiment description

The experiment data was acquired over the instruction period (February to October/2018) considering 11 machine operators and the full tree harvest system. The training process has a learning curve of forest harvest operators with a sigmoidal model shape (Purfürst 2010). This process has a substantial performance of training by taking into account three

learning levels (beginner, medium, and professional). The learning rate decreases at the last stage, which validates the professional skills (Malinen, Taskinen and Tolppa, 2018). Operators with high scores may reach the production target of the company (Lopes and Pagnussat 2017). They are trained in operating Doosan DX300LL Hydraulic Excavators with 267 hp and 1800 rpm. These machines have a Rotobec grapple saw bar (157 cm or 62 inches), grapple area (1 m²), and are previously set for the company's length pattern (7.2 m). Two metrics were calculated to describe the machine performance. Machine productivity (*mp*) is the amount of volume (m³) harvested per hour (h). These values are associated with the total volume (forest inventory) and the effective hours worked by an operator in the field. Moreover, the machine utilization rate percentage (*mur*) reflects the efficiency of the operator training, which is defined by the scheduled machine hours (*smh*) and the productive machine hours (*pmh*).

Database structure

The step before processing consists of structuring the database. This procedure aimed to explore the variables since their influence on the machine productivity behavior is a key issue to analyze. The company has 183,515 ha of *Eucalyptus spp* trees managed for cellulose pulp production. The study covers 65 forest stands, available for harvest. They are spatially located around 24°13'19" S and 50°32'33" W in Paraná state, Brazil. According to the Köppen climate classification, the local is Cfa/Cfb, which consists of a humid subtropical transition to an oceanic climate. The annual mean temperature in the coldest month is below 18°C, and the hottest month above 22°C (Alvares et al., 2013). The quantitative variables were obtained from the forest inventory (stand age and mean individual tree volume), weather stations in the field (precipitation, maximum and minimum temperatures), and machine variables defined previously (Table 1). The qualitative variables were Eucalyptus species/hybrid (*E. dunnii*, *E. grandis*, *E. saligna*, *E. paniculata*, and *E. grandis* x *E. urophylla*), 2 soil classes (inceptisol and red latosol), 3 classes of soil texture (clayey, very clayey, and sandy-loam) and 2 work shifts (6:00 a.m. to 4:00 p.m. and 4:00 p.m. to 1:37 a.m.). The final database has 297 observations and we randomly divided them into two independent sets (80% - adjust/train, 20% - validation) to assess the tested methods.

Table 1. Descriptive values of the studied variables for the analysis (n=297).

Variables	Mean (\pm Sd)	Variables limits		Units	CV (%)
		Min.	Max.		
mp	97.96 \pm 30.0	38.20	183.60	m ³ hour ⁻¹	30.62
exp	113.65 \pm 63.5	0.00	242.00	days	55.90
a	8.36 \pm 2.6	6.00	15.00	years	30.82
eh	5.85 \pm 2.0	1.00	9.00	hours	34.83
v	0.39 \pm 0.1	0.23	0.80	m ³	28.48
tmax	22.84 \pm 4.0	13.26	31.15	°C	17.61
tmin	14.29 \pm 3.9	3.93	22.44	°C	27.11
pp	2.57 \pm 6.6	0.00	43.60	mm	257.60
mur	63.1 \pm 22.2	10.4	100	%	35.3

Where: Sd: standard deviation values, mp: machine productivity, exp: operator experience, a: stand age, eh: effective hours, v: mean individual tree volume, tmin: minimum temperature, tmax: maximum temperature, pp: precipitation, mur: machine utilization rate, Min.: the minimum value of the variable, Mean: the mean value of the variable, Max: the maximum value of the variable, CV: coefficient of variation as a percentage (%), and n=sample size.

Methods for predicting the machine productivity

Once in possession of the database, Pearson's correlation was used as an indicator of the ability of the independent variables to explain the response variable. This descriptive analysis considers only the linear association between two variables. Therefore, regression methods were applied to examine the power of the interaction of independent variables in predicting the response variable. The procedure used was based on the evaluation of five methodological approaches for predicting machine productivity, and is described below.

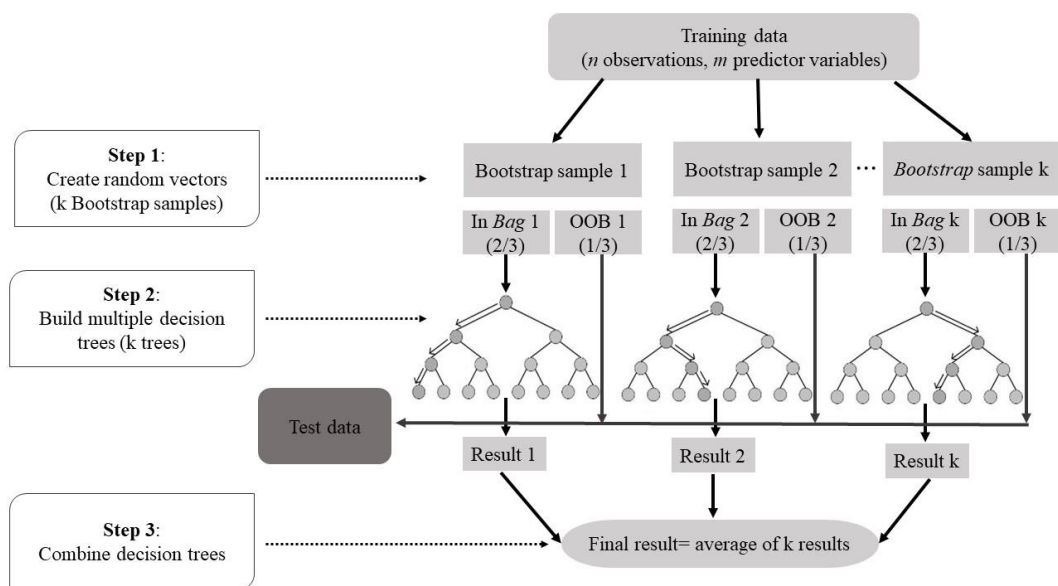
Stepwise

Multiple linear regression (MLR) is recognized as a relevant statistical method to explain the relationship between the predictive and response variables (Ciulla and D'Amico, 2019). However, there are several variables available that can be empirically suggested as inputs to the linear model in our study. Hence, we applied the stepwise building model to formulate a new optimized version. This technique is an iterative procedure adding (forward step) or removing (backward step) variables according to the selection criteria (Alves, Lotufo and Lopes, 2013). We used both steps under the Akaike Information Criterion (AIC) metric for model selection and final convergence. Meanwhile, the F statistic evaluates the contribution of each independent variable selected in the model. Further, we checked the multicollinearity applying the Variance Inflation Factor (VIF) and eliminating variables with values superior to 10. All analysis was performed under the “car” package (Fox and Weisberg, 2020), lm, and step functions available in the R software (R CORE TEAM, 2018) were used.

Random Forest (RF)

The Random Forest algorithm is a non-parametric estimation technique based on a decision trees ensemble (Breiman 2001). Generally, this ensemble reduces the variance and increases the predictive performance (Figure 1). The algorithm sets that each tree structure relies on a random vector of variables. This vector refers to the samples in the bag, in which the trees are built using a bootstrap sample different from the original data set. This sample consists of $2/3$ of the original data set used for training each decision tree. The remaining data consists of out-of-bag samples (OOB) that compose the test set. Further, it is also used for internal cross-validation (Auret and Aldrich, 2012; Breiman, 2001). In training, the decision trees are grown through divisions in the data set. Each division is performed based on the random selection of a subset of predictor variables (m) less than the total number of available variables. This procedure leads to the building of several trees with different results. Then, these trees are combined to predict the response variable, which is calculated by the average of their results (Ahmed et al., 2015; Auret and Aldrich, 2012; Breiman, 2001). The learning method has initial parameters to define the algorithm rules and they may affect the performance. Usually, these parameters are set under preliminary tests to fit the model, and we used 500 trees (ntrees), 2 chosen attributes (mtry), and the observation number of nodes equal to 4 (nodesize). The performance of the algorithm was established based on 50 repetitions to obtain the best model with the lowest mean square error (mse). We used the randomForest package (Liaw and Wiener, 2018) in R software (R CORE TEAM, 2018).

Figure 1. Flowchart of the Random forest structure.



Source: Adapted from Rodriguez-Galiano et al. (2016), and Cheng et al. (2019).

Artificial Neural Networks (ANN)

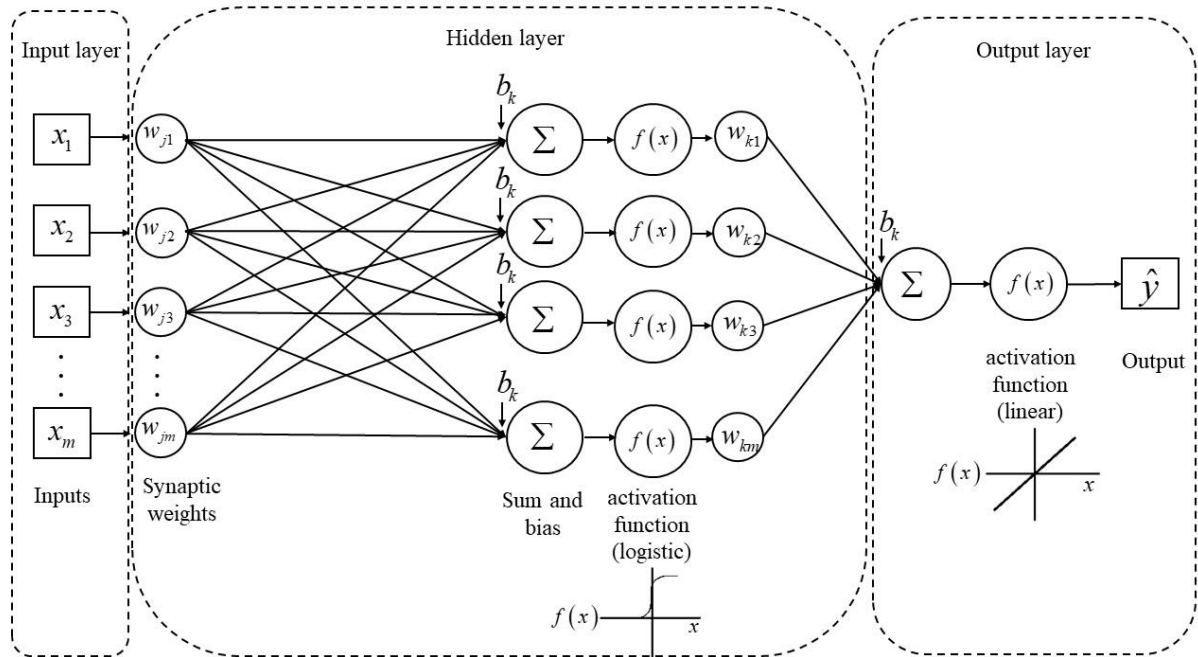
The ANN is inspired by the human brain learning process by linking the input/output data and correlating them (Zhao et al., 2019). The artificial neuron is the basic unit of information processing (Eq. 1), which represents the dendrites of the natural neuron (Tavares Júnior et al., 2020). In the hidden layer, each input x_j is weighted by a respective synaptic weight w_{kj} , configuring m synapses. The bias (b_k) is understood as a synaptic weight with fixed input. Subsequently, the sum of each input multiplied by its respective weight occurs, resulting in a linear combination of the inputs together with the bias. The activation function ($f(x)$) transforms these values to define the range of the artificial neurons in the output layer (Y) (Haykin, 2009). The modeling process starts with the pre-treatment of the data to deal with continuous and categorical variables. The categorical variables are transformed into a binary coding $\{0,1\}$. As for the continuous variables, these are normalized $\{0-1\}$ taking into account the Max-Min procedure, in which v_{new} is the normalized value of the data set and v is the original value of the data set. (Eq. 2).

$$Y = f \left(\sum_{j=1}^m w_{kj} \cdot x_j + b_k \right) \quad (1)$$

$$v_{new} = \frac{v - \min(v)}{\max(v) - \min(v)} \quad (2)$$

After the initial procedures, the modeling of machine productivity was performed using networks with Multilayer Perceptron (MLP) architecture (Figure 2). This structure consisted of one input layer or more hidden layers that receive and process the data. Previously, we tested the ANNs configurations for high performance with the lowest mse. The parameters set are: a) the number of neurons in the hidden layer (4), b) the activation function in the hidden (logistic) and output layer (linear) and c) learning algorithm (Resilient Propagation – Rprop +). Resilient Propagation is the most widely used algorithm for function approximation problems in forest science (Freitas et al., 2020). Finally, the 50 ANNs were trained under the previous configuration until the stopping criterion (100,000 iterations) was reached. The neural network was implemented using the *neuralnet* package (Fritsch et al., 2019) in R software (R CORE TEAM, 2018).

Figure 2. Flowchart of Artificial neural network structure.



Genetic algorithm for feature selection in Random Forest (GA-RF) and Artificial Neural Network (GA-ANN)

Variable selection is a challenging task that seeks to optimize models with a minimal number of variables. This procedure is trivial when building models under statistical rules. However, we are suggesting a new approach to figure out only those variables that best explain the machine productivity. This strategy aims to remove redundant and weak variables that are not correlated with the dependent variable. A deep search of the combinatorial problem is necessary to solve this task. Instead of testing all possible combinations, we applied the genetic algorithm (GA) to find a fast and approximate solution in the machine learning methods. This algorithm selects a set of variables for the Random Forest (GA-RF) and artificial neural network (GA-ANN) over the generations. The GA is inspired by Darwinian evolution theory with the principles of the survival of individuals (Pattanaik, Basu and Dash, 2018), and also natural selection mechanisms (Honda, 2018). The major part of the meta-heuristics has to define a set of parameters or rules to solve a constrained or unconstrained problem. Thus, we first carried out tuning tests to reach the best parameterization: a) population size (100); b) generations (10); c) crossover (50%); d) mutation rate (10%); and e) selection operator (tournament).

The individual has a randomized vector with a fixed size (number of variables) and variable position (gene). The algorithm operates the search procedure over this string structure. The gene has a binary code $\{0, 1\}$ that selects $\{1\}$ or not $\{0\}$ the variable for the next stage.

Hence, a set of enabling features is the input of RF or ANN algorithms to model the machine productivity. It is worth mentioning that we applied the same RF and ANN configuration as described earlier (see the Random Forest and Artificial Neural Networks sections for details). The fitness function evaluates the solution performance to guide the search procedure. In contrast to other studies, we denote this function as a multi-objective form that includes minimization of the number of variables used and the error. Instead of only focusing on the error, the number of variables may inflate the application of the final model. A key resource for balancing between these two factors is a normalized scale with 0-1 values (Eq. 3). The first component of the equation relates to the error, which has a maximum utopic value found for all the variables selected. This hypothetical function behavior works properly as a maximum value in the worst case of training. The second part is the ratio between the number of selected variables (n) and the total number available (N).

$$fitness = \left(\frac{mse}{\max(mse)} + \frac{n}{N} \right) \quad (3)$$

Variable selection performance and methods assessment

Due to a wide range of investigated models under the genetic algorithm, we have only highlighted the best structure found to evaluate the performance (RMSE - root mean squared error and B – bias). Currently, these two metrics (Equations 4 and 5) validate most modeling studies in the literature. Residual plots and histograms were also applied as a complementary statistical analysis. These plots were based on percentage error (%) (Eq. 6), in which Y_i is the measured value of mp in the i observation, \hat{Y}_i is the predicted value of the i observation and n is the total number of observations. Moreover, we summarized the correlation coefficient, standard deviation, and the root of the mean square error in Taylor's diagram (Taylor, 2001). This graph displays the measure distance of the method's performance relying on the prediction and real values (Yaseen et al., 2018).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (4)$$

$$B = \frac{\sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i}{n} \quad (5)$$

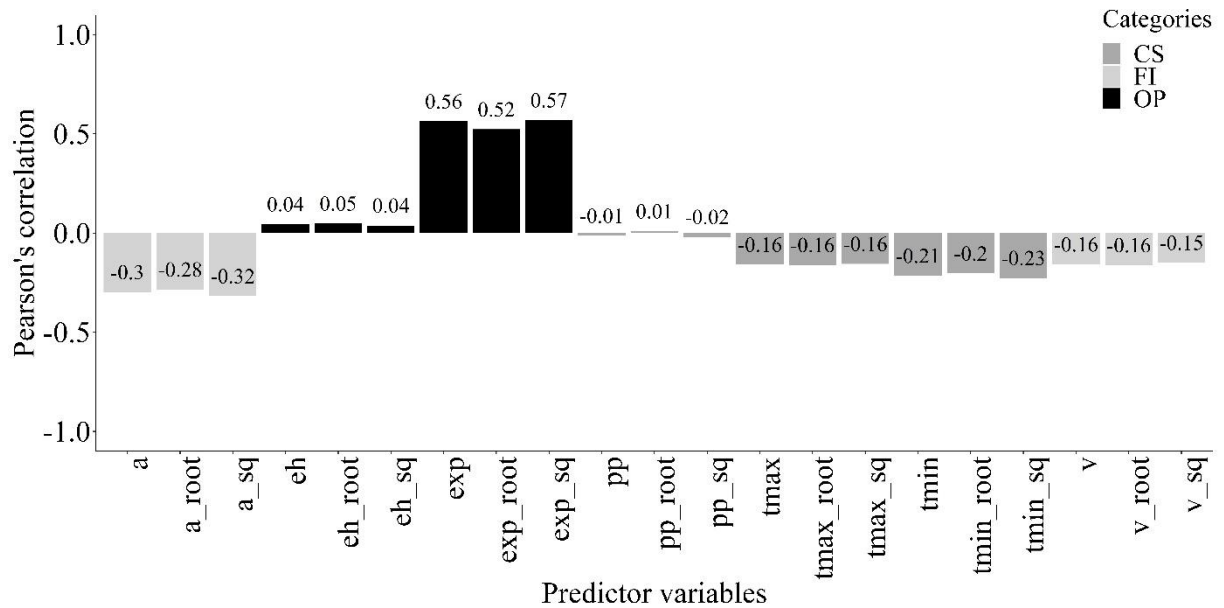
$$Error(\%) = \frac{(Y_i - \hat{Y}_i)}{Y_i} 100 \quad (6)$$

Two independent procedures were taken into account to evaluate the importance of each selected variable. The Increment in Mean Square Error (%IncMSE) was used only for the Random Forest algorithm, in which high values suggest the importance gradient tendency (Miao et al., 2018). Thus, each predictor variable belonging to the out-of-bag (OOB) sample set has its values exchanged. Meanwhile, the values of the other variables remain fixed. The increase in mse is computed when the disturbance in a significant variable reduces the predictive capacity of the model. Finally, we applied the Garson algorithm to measure the relative importance of each variable based on the extracted weights of the ANN (Olden and Jackson, 2002). For this, we used the NeuralNetTools package (Beck, 2018) in R software (R CORE TEAM, 2018).

Results

The real data used in the current study is crucial to support the training process of machine operators to achieve high productivity, as we noticed an average rate increase of 37.5% in productivity from the intermediate period of operator training. As expected, task repetition led to moderate linear operator improvement, demonstrated by a positive correlation with an average of 0.55. There are several variables with a positive correlation with machine productivity (Figure 3). Nevertheless, the advantages of operator skills over environmental and forest inventory variables are highlighted initially due to the higher correlation values. The impact of operator experience on the machine productivity is clear when analyzed specifically by work shift. Although productivity was higher in the first shift, a strong linear association between these variables was 36% higher in the second shift than in the first. This fact denotes a greater dependence on well-trained operators in the second shift to achieve good productivity. The second level of correlation importance is a mix of these two last classes of variables. In addition, the forest stand age is inversely proportional to our investigated variable. Our data points to greater machine productivity in younger stands, with an average difference of 23% compared to older stands. Null correlations were found for variables related to effective hours of work and precipitation. Thus, we can assume that these two variables have no significant linear impact on machine productivity performance under our study condition.

Figure 3. Indicators of the association between independent variables and machine productivity.



Where: CS: climate and soil variables, FI: variables from Forest Inventory, OP: variables related to the operator, a: stand age, a_root: square root of stand age, a_sq: second power of stand age, pp: precipitation, pp_root: square root of precipitation, pp_sq: second power of precipitation, eh: effective hours, eh_root: square root of effective hours, eh_sq: second power of effective hours, v: mean individual tree volume, v_root: square root of mean individual tree volume, v_sq: second power of mean individual tree volume, exp: operator experience, exp_root: square root of operator experience, exp_sq: second power of operator experience, tmin: minimum temperature, tmin_root: square root of minimum temperature, tmin_sq: second power of minimum temperature, tmax: maximum temperature, tmax_root: square root of maximum temperature and, tmax_sq: second power of maximum temperature.

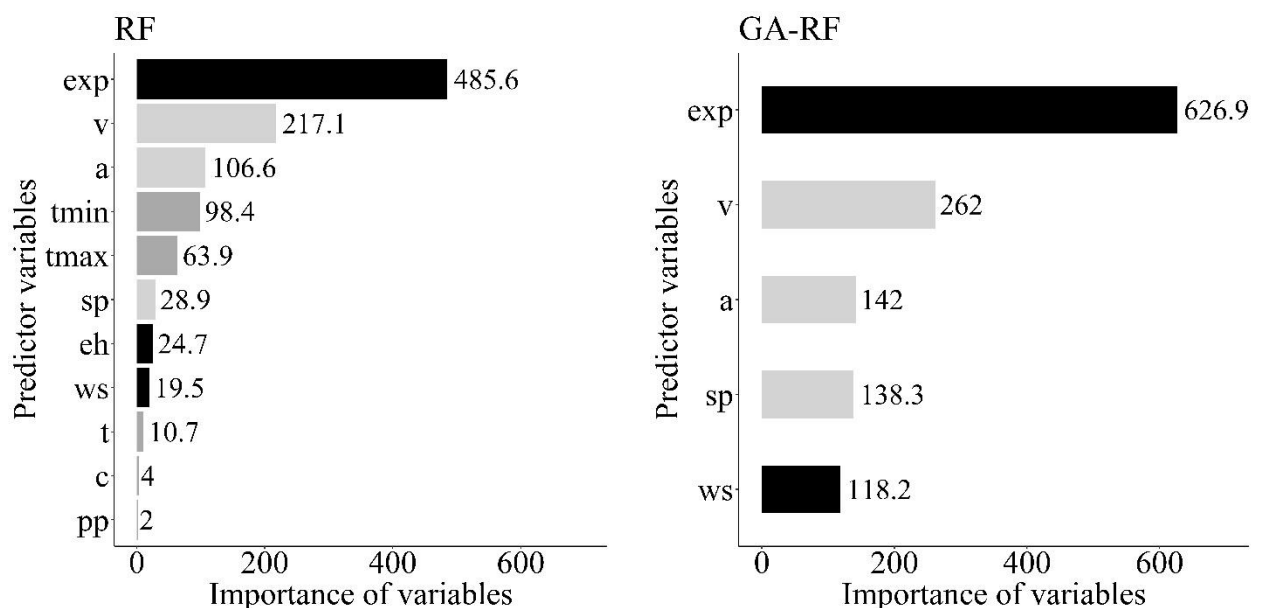
The predictor variables' importance and their order change according to the method. For instance, the stepwise model has only two significant variables (Eq. 7). They explain 93% of the machine productivity variation in the training data. All the predictors were significant at the 0.1% level. The feature selection processes cope with multi-dimensional problems to boost the model accuracy. This procedure reduced the range of variables for RF (80%) and ANN (72%). The selected variables are not similar in each of the other tested methods, which denotes the algorithms' divergences in the modeling procedure. Moreover, the genetic algorithm defined an optimal set of variables for RF that were associated with the forest inventory (species, stand age, and volume), work shift, and operator experience. On the other hand, the environmental variables (soil texture, temperature (max and min), and precipitation), work shift, operator experience, and effective hours were selected for the ANN. As observed, both GA-ANN and GA-RF took advantage of the effect of operator-related variables. Nevertheless, species, stand age, and volume only had an influence in GA-RF, since GA-ANN opted for soil texture, temperature, and precipitation. The inferior precision of GA-ANN may be associated with the lower correlation of these selected input variables to explain the machine productivity. For

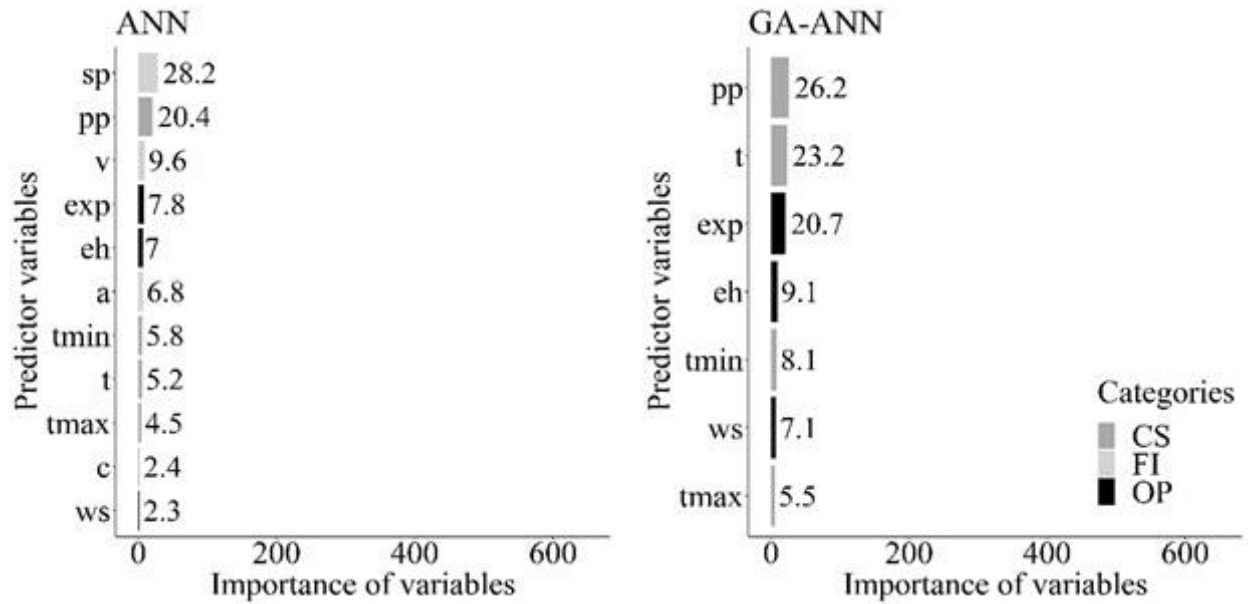
instance, the experiment analysis was realized under dry weather conditions in the study region which affected the final results.

$$mp = 0.0013 * exp^2 + 30.80 * \sqrt{eh} \quad (7)$$

In addition, the machine learning methods under the genetic algorithm amplify and increase the importance of the variables in the RF (v - 20.68%, exp - 29.09%, sp - 378.54%, a - 25.35%, and ws - 506.15%) and ANN model (pp - 28.43%, t - 348.07%, exp - 165.38%, eh - 30%, tmin - 39.65%, ws - 208.69, and tmax - 22,22%). It is noteworthy that GA-ANN did not show a clear trend in the selection of variables. This approach omitted the variables that contributed most to ANN (FI variables). On the other hand, GA-RF optimized the set of variables, enhancing the importance of those that contributed most to the RF (FI and OP variables) and consequently removing the irrelevant ones in its prediction procedure. Most of the variables showed similar behavior in the RF and GA-RF methods, despite the different contribution levels of the variables to the machine productivity. Remarkably, both the RF and GA-RF models indicated that operator experience (exp) was by far the most important variable affecting machine productivity. Although this variable does not occupy the most prominent position for the other models, it still has a decisive role in predicting machine productivity. This fact emphasizes the relevance of the operator's training level in the efficiency of the operation. The work shift also influenced the machine productivity, but less than the operator's experience. As expected, the remaining variables (Figure 4) tended to contribute part of the model variation at lower levels of importance.

Figure 4. Analysis of the predictor variable importance.





Where: CS: climate and soil variables, FI: variables from Forest Inventory, OP: variables related to the operator, *sp*: specie, *ws*: work shift, *t*: soil texture, *c*: soil class, *a*: stand age, *pp*: precipitation, *eh*: effective hours, *v*: mean individual tree volume, *exp*: operator experience, *tmin*: minimum temperature and *tmax*: maximum temperature. The measure of the variables' importance is intrinsic to each algorithm, being in percentage unit (%) for models based on artificial neural networks and in the error increment (IncMSE%) for models based on random forest.

Modeling machine productivity is not a simple task, as we found from testing various methods that achieved different results. Although the Stepwise model resulted in similar prediction error trends in the training/validation datasets, we do not recommend it due to the high RMSE values (Table 2). In contrast, the ANN has the best predictive performance on the same dataset, nevertheless proved deficient in terms of validation. This fact reflects the overfitting problem that has been intrinsic to this procedure. Both RF and GA-RF are indicated as suitable to model our dependent variable as they provided the most precise and stable of errors tendency. There was also a reduction in the error values under the feature selection procedure of the variables. Thus, the accuracy of the mp estimates for the validation data was improved over RMSE by 21% for ANN and 0.4% for RF after using the hybrid methods (GA-ANN, and GA-RF). In terms of computational time demands, the genetic algorithm inflated the time consumption (RF - 39.47 s and ANN - 446.47 s) by significant values compared with their standard forms.

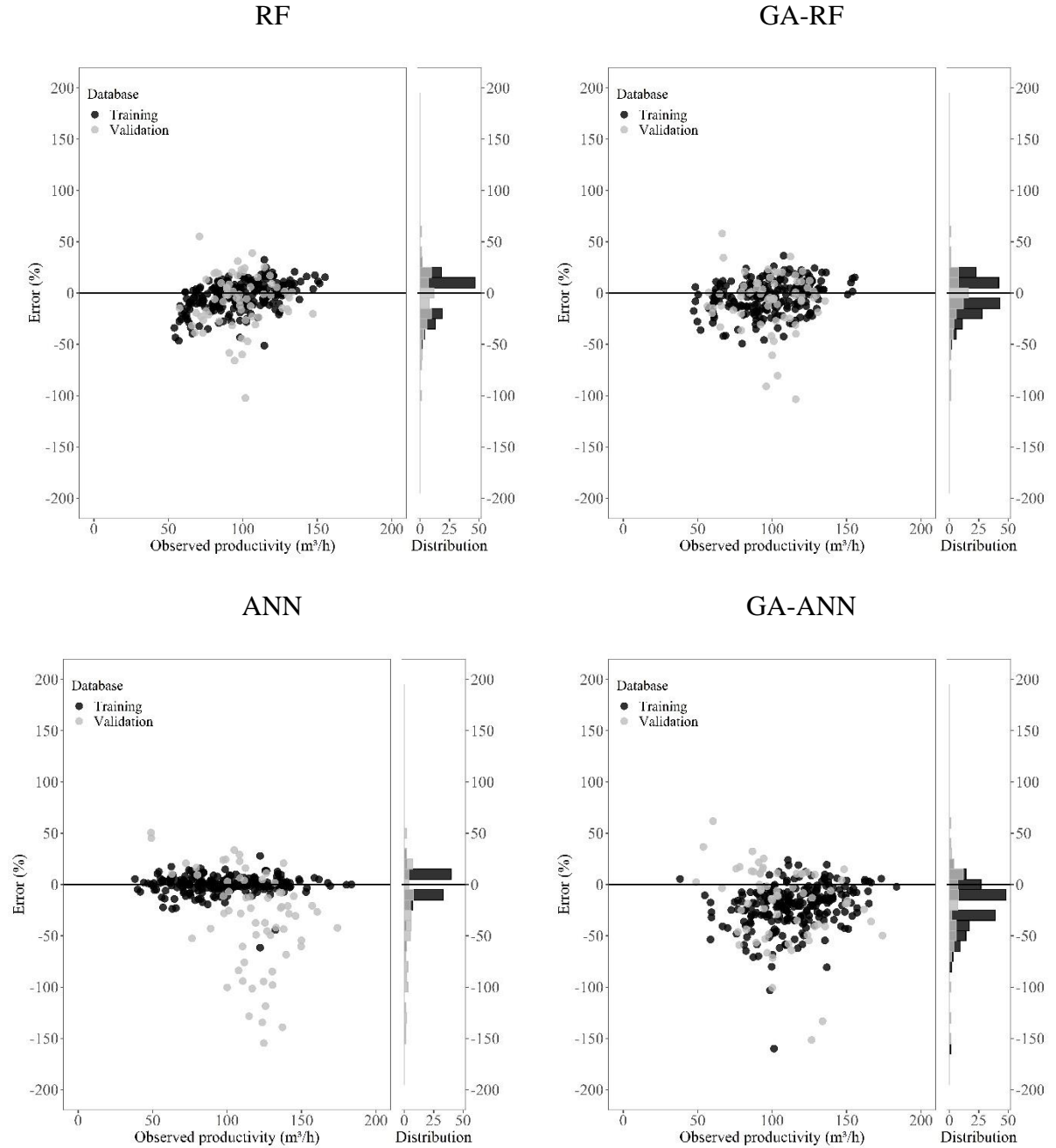
Table 2. Adjustment statistics for the training and validation data, the number of variables, and selected variables using the Stepwise, RF, GA-RF, ANN, and GA-ANN method.

Methods	Dataset	RMSE	RMSE (%)	Bias	N	Selected variables	Time (sec)
Stepwise	Training	27,28	27,75	2,28	2	exp_sq* + eh_root*	0,02
	Validation	27,02	27,94	2,93			
RF	Training	12,8	13,02	-0,06	25	All variables	0,53
	Validation	23,99	24,72	-1,51			
GA-RF	Training	14,68	14,93	-0,05	5	ws + sp + a_sq + v_sq + exp_root	40,0
	Validation	23,89	24,72	-1,65			
ANN	Training	8,08	8,21	0,36	25	All variables	13,97
	Validation	39,26	40,61	-20,53			
GA-ANN	Training	22,83	23,22	-16,83	7	t + ws + eh + eh_root + tmax + pp + tmin + exp_sq	460,44
	Validation	31,04	32,11	-10,53			

Where: RF: random forest; GA-RF: Genetic Algorithm and Random Forest; ANN: Artificial Neural Network; GA-ANN: Genetic Algorithm and Artificial Neural Network; N: number of selected variables; *sp*: specie, *ws*: work shift, *t*: soil texture, *a_sq*: stand age, *pp*: precipitation, *eh*: effective hours, *eh_root*: square root of effective hours, *v_sq*: second power of mean individual tree volume, *exp_root*: square root of operator experience, *exp_sq*: second power of operator experience, *tmin*: minimum temperature, and *tmax*: maximum temperature; * Significant at 99.9% of probability and VIF less than 5.

The residuals plot suggests a complementary analysis to check the tendencies of the prediction (Figure 5). We observed a slightly biased prediction of the ANN and GA-ANN methods mainly for the validation data set. The stepwise model also had problems in predicting extreme values. However, RF and GA-RF present good error tendencies concentrating them within $\pm 50\%$ of the y-axis for both the tested data sets. Taylor's diagram also corroborates the previous analysis by showing the striking differences in the methods' performance in a single plot (Figure 6). Regardless of the data set type, RF and GA-RF perform better than other models.

Figure 5. Residuals plots analysis of modeling methods (RF: random forest; GA-RF: Genetic Algorithm and Random Forest; ANN: Artificial Neural Network; GA-ANN: Genetic Algorithm and Artificial Neural Network; and Stepwise), and datasets.



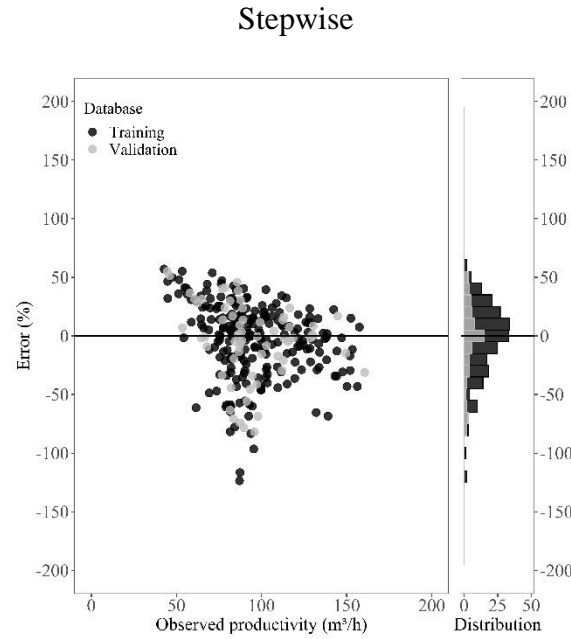
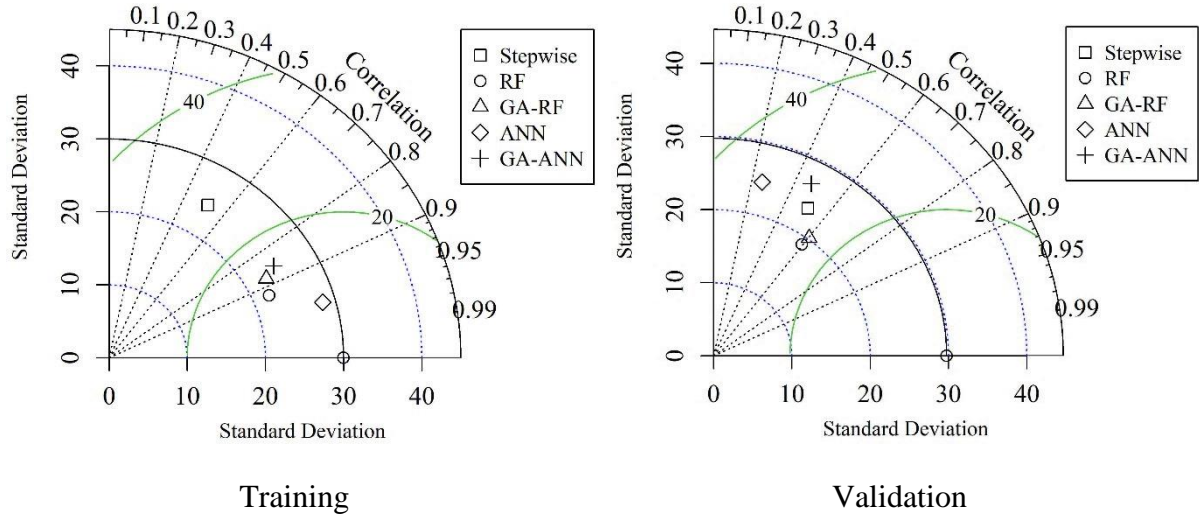


Figure 6. Taylor diagrams of machine productivity modeling methods (RF: Random forest; GA-RF: Genetic Algorithm and Random Forest; ANN: Artificial Neural Network; GA-ANN: Genetic Algorithm and Artificial Neural Network; and Stepwise), and datasets.



Discussion

The modern forestry industry is highly dependent on machines to meet wood supply demand (Purfürst, 2010). Therefore, information about a machine’s productivity to guide its proper allocation at the stands can ensure high performance and low risk to operations. However, the reliability of the machine productivity is subject to a series of variables that, if well understood, can support good decision-making. An advantage of our study relies on the

heterogeneity of the database, with several variable types (weather conditions, soil, forest inventory, and machine operator) that affect machine productivity performance. Knowledge of the effect of these variables may help explain a truly representative part of the variation in machine productivity. In our study, the mean productivity of the excavator-based grapple saw was $97.96 \text{ m}^3 \text{ h}^{-1}$ (with 30.62% of variation). This value was 22% higher than the maximum performance ($76.57 \text{ m}^3 \text{ h}^{-1}$) reported by Lopes et al. (2008) for a stand with a volume of $300 \text{ m}^3 \text{ ha}^{-1}$. Even so, both results have the same tendency as those found in Rocha et al. (2009), since their limits were $58 - 118 \text{ m}^3 \text{ h}^{-1}$. Several factors inflate this variation range, most of them associated with environmental conditions (Liski et al., 2020), log length (Spinelli et al., 2021), forest yield (Lopes et al., 2008), and the different levels of operator training, which was one of the conditioning factors of our study.

One key challenge for decision-makers is to comprehend which variables most influence the machine's productivity. The importance of each variable differed between the modeling strategies tested in this study. Nevertheless, our findings highlighted a set of variables that contributed strongly to predicting machine productivity based on the models with the greatest accuracy (RF and GA-RF). Previous studies by Liski et al. (2020) and Rossit et al. (2019) corroborate our findings, as they proved the efficiency of tree-based algorithms in predicting the productivity of cut-to-length (CTL) forest harvesting systems. They also identified that the operator had a significant influence on machine productivity. A similar tendency was obtained in the current study, since operator experience was among the highest contributing variables in the models, even though the results varied for each model. Liski et al. (2020) also confirm that the influence of the operator depends on the modeling approach. Recent research also suggests that machine operators play an essential role in the efficiency of tasks (Dvořák et al. 2019) and their modeling (Liski et al. 2020). Purfürst and Eler (2011) evaluated the effect of machine operator training, and found a high level of productivity for those with a higher training rate. There is a consensus on the importance of operator skills to increase efficiency, and how motivated they are daily. Besides the incentives, a recurrent training program may also avoid risks and delays (Cho et al., 2019), mainly when the operator performance results in a lower machine utilization rate, as presented in our study. Other performance-related factors include living conditions (quantity and quality of sleep), working environment conditions (available light, air quality, vibration, and noise inside the cabin), and shift arrangements (Malinen, Taskinen and Tolppa, 2018). Concerning shifts, productivity is strongly associated with how long time the work takes at the operation. In fact, operator exhaustion is always associated with long shifts and unplanned timetables. According to Passicot and Murphy (2013), the efficiency

of harvesters (41%) and processors (61%) reduces after 18 hours of continuous working time. As the shift lengths were similar in the current study for the first and second shifts, the main difference in productivity can be assigned to the time of day (Murphy, Marshall and Dick, 2014). We observed a drop in both machine productivity (on average 6%) and the machine utilization rate (on average 10%) on the night shift. In fact, operators working at night face challenges in performing tasks due to reduced alertness and availability of light, thereby decreasing their productivity. Factors such as dim lighting, shadowing, and glare lead to reduced visibility, and affect log-making and machine-positioning accuracy (Nicholls, Bren and Humphreys, 2004).

The forest stand structure usually influences the harvest, skid/drag, forward (Hiesl and Benjamin, 2013), and processors operations (Passicot and Murphy, 2013). As our results showed, the mean tree volume was also an important factor affecting the productivity of the excavator-based grapple saw. This result is parallel to the finding by Strandgard, Mitchell, and Acuna (2016). They clarified that the mean tree volume was the most significant independent variable (79%) when modeling the harvester productivity for eucalyptus plantations in Australia. The relationship between these variables is characterized as directly proportional since the harvester productivity increases as the tree volume increases (Norihiro et al., 2018). Another effect is that heavier (big) trees result in more mechanical problems and a greater decline in productivity than small trees. The last point is especially the case when the operator is working under conditions of physical and mental fatigue (Passicot and Murphy, 2013).

In the present study, the benefit of the interaction between tree size and stand age in machine productivity modeling is evident. In the models that produced the best estimates, these variables were among the most important. The stand age and management regimes are also factors that limited the performance due to the tree size and barriers to the machine transit in mechanical thinning operations (Mederski et al., 2016). Generally, lower machine productivity is achieved for younger stands on poor sites. On the other hand, species characteristics such as shape, tree health, percentage of bark, and size of branches may also impact the machine productivity (Olivera et al. 2016; Rossit et al. 2019). Passicot and Murphy (2013) found greater productivity impacts from harvesters operating with stands of *E. globulus* than stands of *E. nitens*. We noted a substantial trend of the impact of these variables and their interactions working in *Eucalyptus spp.* plantations. This statement is supported by the good estimates of productivity in most of the models tested with the data available for our study.

Currently, computational advances have been applied in many companies with significant returns. They are also investing in acquiring a variety of field data and storing it for further

analysis. This technology is a reality that makes achieving high efficiency accessible to the modern industry. However, the challenges of modeling procedures by hand with high-dimensional data or non-linear patterns reduce the final accuracy. Eriksson and Lindroos (2014) point out that the search for acceptable accuracy is fundamental to the control and planning of harvesting operations, as outlined in our study objectives. Insights from machine learning techniques in forest harvesting science have been applied to solve many of the issues debated by researchers and practitioners. These issues include mapping the forest stand susceptibility to damage during harvest (Shabani, Pourghasemi and Blaschke, 2020), the establishment of the optimal operation mode for handling chainsaws to reduce the emission of pollutants into the atmosphere (Dimou et al., 2018), and predicting the productivity of harvesting systems (Liski et al., 2020). In alignment with the latter issue, machine productivity modeling has been a hard task, regardless of the method. In general, machine learning methods have several advantages over MLR regression, due to the use of qualitative variables, the data type, non-linearity (Were et al., 2015), noise problems, and outliers (Auret and Aldrich, 2012). In the current study, we find the best Stepwise model regression by selecting only two obvious variables (operator experience and effective hours). Although this technique has achieved good accuracy (Fujiwara et al., 2009) with a limited set of variables, it has proved to be inferior to machine learning methods when training models from a reduced data set. The machine learning techniques and hybrid methods (HM) under the genetic algorithm differ slightly from the regression analysis. Thus, the accuracy analysis was critical to define the best strategy, in which the ANN obtained the lowest errors in the training data. Nevertheless, the overfitting was pretty evident when checking the difference in RMSE (%) between the training and validation set (32.4%). This model behavior leads to a low generalization capacity, and it consequently cannot be applied reliably to other data sets (Mohammadi et al., 2019). On the other hand, the RF has an internal validation mechanism with less susceptibility to overfitting (Breiman, 2001), as observed in our results. It can be inferred that RF provided a high quality estimate to predict the machine's productivity without the previous feature selection. The feature selection procedure using the genetic algorithm simplified the model but provided little additional gain to the predictive accuracy of the RF. This may occur because the ideal set of variables is not decided by machine learning methods (Jadhav, He and Jenkins, 2018). The larger the number of predictor variables (more than 25 attributes), the genetic algorithm should improve the efficiency of machine learning methods when dealing with large sets of variables. On the other hand, it requires more computational effort. We could have done some variable transformations to reach at least 50 variables and tested the simplification power of the model with greater gain in precision.

However, we may extend this assumption for other further works. Concerning ANN, we do not suggest its use within a genetic algorithm, taking into account studies with similar data sets. In contrast, Murthy and Koolagudi (2018) applied GA-ANN and GA-RF with a high success rate for accuracy (>90%), which differs from our study and application instances.

Our findings highlight the relevance of including the genetic algorithm, especially for a large number of variables, as reported by Fassnacht et al. (2014) and Li et al. (2016). In this sense, we do not detect that a database reduction makes a significant contribution. This fact is attributed to our database already containing a small number of available variables. Regardless of the final accuracy, in forest harvesting, studies focusing on the comparison/implementation of several modeling strategies have been increasing. In contrast, a direct comparison of our results with those taken from other researchers, such as Rossit et al. (2019) and Liski et al. (2020), is hard to conduct. The reason is the differences in the models utilized, the data sources, the set of predictor variables, and the machines evaluated. Furthermore, they did not implement an automatic feature selection approach, which still reflects a gap in recent studies in this field of science. Thus, this novelty is one of the main contributions proposed in this study using the GA implementation. Remarkably, the GA-RF method provides a strong model simplification with 5 selected variables, the collection of which demands less operational effort in the field. Therefore, the selection of predictor variables is a credible strategy for estimating machine productivity. At this point, our results indicate a promising choice to extract relevant variables from a dataset. Furthermore, this methodology may be replicated for other machine studies, adding also a wide range of variables (wood log details, equipment consumption, spatial and surface information, extreme environmental conditions, and forest variables).

As stated by Liski et al. (2020), as the harvesting database continues to increase, new machine learning methods will be needed to make accurate predictions. A common strategy is to build competing machine learning models based on the same (continuously updated) streaming data, all of them being shared and saved on a website or a cloud platform. In terms of applicability, the modeling approach employed in our research can also be useful to feed systems with more accurate machine productivity. Therefore, machine productivity associated with more information would assist in scheduling activities in accordance with the stand age, tree size, human factors and weather conditions. This information can be in high demand in future applications as it serves as an input to improve the operational planning and costs of harvest activities.

Conclusion

Forest machinery operates under a range of variables that must be considered by decision-makers when planning harvesting operations. This study proposed an approach to analyze the impact of weather conditions, forest inventory data, and operator records in excavator-based grapple saw productivity. In this context, we highlight operator experience as the main component for estimating the excavator-based grapple saw productivity. The work shift had an impact on the machine productivity, but not as significantly as the operator's experience. A key role in machine productivity was also played by the stand characteristics (mean individual tree volume and stand age). Also, we found that moving from a background of using only statistical linear regression to applying RF and GA-RF as modeling techniques enabled considerable predictive improvements. Here, we highlighted the ability of the genetic algorithm in conjunction with machine learning techniques (hybrid methods), and especially GA-RF, to perform a robust selection of the variables that truly affect the machine productivity. Thus, the outcomes of the current research, integrating forest harvesting and machine learning techniques, provide valuable findings for science and companies, as they can be the inputs to an operational planning model in future applications.

Acknowledgments

The authors are especially grateful to the forest company for providing the database and supporting the initial premises of this work; the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brazil) for financial support under Finance Code 001; the Editor and Reviewers for the contributions to the study improvement.

References

- AHMED, O. S. et al. Characterizing stand-level forest canopy cover and height using Landsat time series, samples of airborne LiDAR, and the Random Forest algorithm. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 101, p. 89–101, 2015.
- ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.
- ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Applied and Computational Mathematics**, v. 1, n. 1, p. 1–6, 2013.
- APOLLONI, J.; LEGUIZAMÓN, G.; ALBA, E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. **Applied Soft Computing Journal**, v. 38, p. 922–932, 2016.
- AURET, L.; ALDRICH, C. Interpretation of nonlinear relationships between process variables by use of random forests. **Minerals Engineering**, v. 35, p. 27–42, 2012.
- BECK, M. W. Package ‘**NeuralNetTools**’, 2018.
- BREIMAN, L. Random Forests. **Machine learning**, v. 45, p. 5–32, 2001.
- CHENG, L. et al. Applying a random forest method approach to model travel mode choice behavior. **Travel Behaviour and Society**, v. 14, p. 1–10, 2019.
- CHO, M. J. et al. Comparison of productivity and cost between two integrated harvesting systems in South Korea. **Forests**, v. 10, n. 763, p. 1–13, 2019.
- CIULLA, G.; D'AMICO, A. Building energy performance forecasting: A multiple linear regression approach. **Applied Energy**, v. 253, p. 113500, 2019.
- DANTAS, D. et al. Multilevel nonlinear mixed-effects model and machine learning for predicting the volume of Eucalyptus spp. trees. **Cerne**, v. 26, n. 1, p. 48–57, 2020.
- DAS, A. K.; DAS, S.; GHOSH, A. Ensemble feature selection using bi-objective genetic algorithm. **Knowledge-Based Systems**, v. 123, p. 116–127, 2017.
- DIMOU, V. et al. Comparative analysis of exhaust emissions caused by chainsaws with soft computing and statistical approaches. **International Journal of Environmental Science and Technology**, v. 15, n. 7, p. 1597–1608, 2018.
- DVOŘÁK, J. et al. Long-term Cost Analysis of Mid-performance Harvesters in Czech Conditions. **Austrian Journal of Forest Science**, v. 136, n. 4, p. 351–372, 2019.
- ERIKSSON, M.; LINDROOS, O. Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. **International Journal of Forest Engineering**, v. 25, n. 3, p. 179–200, 2014.
- FASSNACHT, F. E. et al. Importance of sample size, data type and prediction method for

remote sensing-based estimations of aboveground forest biomass. **Remote Sensing of Environment**, v. 154, p. 102–114, 2014.

FOX, J.; WEISBERG, S. **Package ‘car’**, 2020.

FREITAS, E. C. S. DE et al. Modeling of eucalyptus productivity with artificial neural networks. **Industrial Crops and Products**, v. 146, p. 112149, 2020.

FRITSCH, S. et al. **Package “neuralnet” Training of Neural Networks**, 2019. Disponível em: <<https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf>>

FUJIWARA, K. et al. Soft-sensor development using correlation-based just-in-time modeling. **AIChE Journal**, v. 55, n. 7, p. 1754–1765, 2009.

GHOSH, M. et al. A wrapper-filter feature selection technique based on ant colony optimization. **Neural Computing and Applications**, v. 32, n. 12, p. 7839–7857, 2020.

HAYKIN, S. **Neural Networks and Learning Machines**. 3rd. ed. New Jersey: Pearson Education, 2009.

HETHCOAT, M. G. et al. A machine learning approach to map tropical selective logging. **Remote Sensing of Environment**, v. 221, n. November 2018, p. 569–582, 2019.

HIESL, P.; BENJAMIN, J. G. Applicability of international harvesting equipment productivity studies in Maine, USA: A literature review. **Forests**, v. 4, n. 4, p. 898–921, 2013.

HONDA, M. Application of genetic algorithms to modelings of fusion plasma physics. **Computer Physics Communications**, v. 231, p. 94–106, 2018.

HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, v. 630, p. 1044–1056, 2018.

JADHAV, S.; HE, H.; JENKINS, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. **Applied Soft Computing Journal**, v. 69, p. 541–553, 2018.

LI, M. et al. A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments. **International Journal of Applied Earth Observation and Geoinformation**, v. 49, p. 87–98, 2016.

LIAW, A.; WIENER, M. **randomForest**, 2018.

LISKI, E. et al. Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods. **International Journal of Forest Engineering**, v. 31, n. 3, p. 253–262, 2020.

LOPES, E. DA S.; PAGNUSSAT, M. B. Effect of the behavioral profile on operator performance in timber harvesting. **International Journal of Forest Engineering**, v. 28, n. 3, p. 134–139, 2017.

LOPES, S. E. et al. Technical and economical evaluation of a slacher , operating under

different productivities. **Scientia Forestalis**, v. 36, n. 79, p. 215–222, 2008.

MAFARJA, M.; MIRJALILI, S. Whale optimization approaches for wrapper feature selection. **Applied Soft Computing**, v. 62, p. 441–453, 2018.

MALINEN, J.; TASKINEN, J.; TOLPPA, T. Productivity of cut-to-length harvesting by operators' age and experience. **Croatian Journal of Forest Engineering**, v. 39, n. 1, p. 15–22, 2018.

MEDERSKI, P. S. et al. Estimating and modelling harvester productivity in pine stands of different ages, densities and thinning intensities. **Croatian Journal of Forest Engineering**, v. 37, n. 1, p. 27–36, 2016.

MIAO, S. et al. Random Forest Algorithm for the Relationship between Negative Air Ions and Environmental Factors in an Urban Park. **Atmosphere**, v. 9, n. 463, p. 1–13, 2018.

MOHAMMADI, F. et al. Modelling and Optimizing Pyrene Removal from the Soil by Phytoremediation using Response Surface Methodology, Artificial Neural Networks, and Genetic Algorithm. **Chemosphere**, v. 237, p. 124486, 2019.

MURPHY, G.; MARSHALL, H.; DICK, A. Time of day impacts on machine productivity and value recovery in an off-forest central processing yard. **New Zealand Journal of Forestry Science**, v. 44, n. 1, p. 1–9, 2014.

MURTHY, Y. V. S.; KOOLAGUDI, S. G. Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS). **Expert Systems with Applications**, v. 106, p. 77–91, 2018.

NICHOLLS, A.; BREN, L.; HUMPHREYS, N. Harvester Productivity and Operator Fatigue: Working Extended Hours. **International Journal of Forest Engineering**, v. 15, n. 2, p. 57–65, 2004.

NORIIHIRO, J. et al. Productivity model for cut-to-length harvester operation in South African eucalyptus pulpwood plantations. **Croatian Journal of Forest Engineering**, v. 39, n. 1, p. 1–13, 2018.

OLDEN, J. D.; JACKSON, D. A. Illuminating the “black box”: Understanding variable contributions in artificial neural networks. **Ecological Modelling**, v. 154, p. 135–150, 2002.

OLIVERA, A. et al. Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a *Eucalyptus spp.* harvesting operation in Uruguay. **International Journal of Forest Engineering**, v. 27, n. 1, p. 15–28, 2016.

OU, Q.; LEI, X.; SHEN, C. Individual tree diameter growth models of Larch-Spruce-Fir mixed forests based on machine learning algorithms. **Forests**, v. 10, n. 187, p. 1–20, 2019.

PASSICOT, P.; MURPHY, G. E. Effect of work schedule design on productivity of mechanised harvesting operations in Chile. **New Zealand Journal of Forestry Science**, v. 43, n. 2, p. 1–10, 2013.

PATTANAİK, J. K.; BASU, M.; DASH, D. P. Improved real coded genetic algorithm for dynamic economic dispatch. **Journal of Electrical Systems and Information Technology**,

v. 5, p. 349–362, 2018.

PURFÜRST, F. T. Learning Curves of Harvester Operators. **Croatian Journal of Forest Engineering**, v. 31, n. 2, p. 89–97, 2010.

PURFÜRST, F. T.; ERLER, J. The Human Influence on Productivity in Harvester Operations. **International Journal of Forest Engineering**, v. 22, n. 2, p. 15–22, 2011.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2018.

ROCHA, E. B. DA et al. Productivity and costs of a full tree harvesting system. **Cerne**, v. 15, n. 3, p. 372–381, 2009.

RODRIGUEZ-GALIANO, V. F. et al. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. **Biogeosciences**, v. 13, p. 3305–3317, 2016.

ROSSIT, D. A. et al. A Big Data approach to forestry harvesting productivity. **Computers and Electronics in Agriculture**, v. 161, p. 29–52, 2019.

SAFARI, A. et al. A comparative assessment of multi-temporal Landsat 8 and machine learning algorithms for estimating aboveground carbon stock in coppice oak forests. **International Journal of Remote Sensing**, v. 38, n. 22, p. 6407–6432, 2017.

SHABANI, S.; POURGHASEMI, H. R.; BLASCHKE, T. Forest stand susceptibility mapping during harvesting using logistic regression and boosted regression tree machine learning models. **Global Ecology and Conservation**, v. 22, p. e00974, 2020.

SHE, J.; CHUNG, W.; KIM, D. Discrete-event simulation of ground-based timber harvesting operations. **Forests**, v. 9, n. 11, p. 1–20, 2018.

SILAYO, D. S.; MIGUNGA, G. Productivity and costs modeling for tree harvesting operations using chainsaws in plantation forests, Tanzania. **International Journal of Engineering & Technology**, v. 3, n. 4, p. 464, 2014.

SILVA, J. P. M. et al. Computational techniques applied to volume and biomass estimation of trees in Brazilian savanna. **Journal of Environmental Management**, v. 249, p. 109368, 2019.

SOMAN, H.; KIZHA, A. R.; ROTH, B. E. Impacts of silvicultural prescriptions and implementation of best management practices on timber harvesting costs. **International Journal of Forest Engineering**, v. 30, n. 1, p. 14–25, 2019.

SPINELLI, R. et al. A Low-Investment Option for the Integrated Semi-mechanized Harvesting of Small-Scale, Short-Rotation Poplar Plantations. **Small-scale Forestry**, v. 20, p. 59–72, 2021.

STRANDGARD, M.; MITCHELL, R.; ACUNA, M. General productivity model for single grip harvesters in Australian eucalypt plantations. **Australian Forestry**, v. 79, n. 2, p. 108–113, 2016.

TAVARES JÚNIOR, I. DA S. et al. Machine learning: Modeling increment in diameter of individual trees on Atlantic Forest fragments. **Ecological Indicators**, v. 117, p. 106685, 2020.

TAYLOR, K. E. Summarizing multiple aspects of model performance in a single diagram. **J. Geophys. Res.**, v. 106, n. D7, p. 7183–7192, 2001.

TOLAN, A.; VISSER, R. The effect of the number of log sorts on mechanized log processing productivity and value recovery. **International Journal of Forest Engineering**, v. 26, n. 1, p. 36–47, 2015.

VIEIRA, G. C. et al. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment**, v. 619–620, p. 1473–1481, 2018.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, v. 52, p. 394–403, 2015.

XING, J. et al. Predictive single-step kinetic model of biomass devolatilization for CFD applications: A comparison study of empirical correlations (EC), artificial neural networks (ANN) and random forest (RF). **Renewable Energy**, v. 136, p. 104–114, 2019.

YASEEN, Z. M. et al. Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. **Advances in Engineering Software**, v. 115, p. 112–125, 2018.

ZHAO, Q. et al. Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. **Forest Ecology and Management**, v. 434, p. 224–234, 2019.

ARTIGO 2 – Complex network metrics and feature selection for modeling individual tree diameter growth

ABSTRACT

Several methods have been applied to measure the inter-tree competition over decades in forest science. Mathematical approaches are often proposed to explain the growth increment taking into account the neighborhood of trees. The interaction between them denotes a network structure with a tradeoff for each individual in the ecological system, described by topological metrics. This information may support the tree's growth modeling for better accuracy at the individual level in consistency with the ecological process. In this sense, our main objective was to compare the performance of classical indices with complex network metrics. They were compared in terms of similarity by Cluster analysis and Pearson's correlation with periodic annual diameter increment (PAI_d). The experimental area is the semideciduous seasonal montane forest in Brazil. Due to the high diversity of the tropical forest, we selected two species (*Xylopia brasiliensis* and *Copaifera langsdorffii*) to be the subject trees of our research during the period of 2010-2017, because they are naturally dominant with the highest importance value indices (%IVI) in our area. Before modeling, the Bitterlich procedure was simulated (BAF=4) to figure out tree competitors using their geographic coordinates (x, y). Further, the PAI_d was modeled under four strategies assisted by the Genetic algorithm and Random Forest method. The strategies encompassed trees variables (diameter at breast height – dbh, basal area, and geographic coordinates), competition indices (distance-dependent, distance-independent and semi-independent), and topological metrics related to complex networks. We assess modeling performance based on error metrics analysis, ranked according to a scale unit. Regardless of the species, both methods have similar importance to explain the PAI_d . However, our findings suggest the use of distance-independent indices and topological metrics for *X. brasiliensis* and *C. langsdorffii*, respectively. Our results revealed the applicability of complex networks to measure effectively the inter-tree competition and their effects on the individual-tree diameter growth. For this reason, we hope that our findings encourage the progress of this interdisciplinary tool in generating insights into the field of Forest Sciences.

Keywords: Competition indices. Tropical forest. Random Forest. Genetic Algorithm.

1 INTRODUCTION

Tropical forests have been extensively studied for conservation and sustainable management uses. They have a complex structure which denotes a high composition and interactions between a range of species (AAKALA et al., 2013). These forests have a major diversity of trees and fauna with high levels of endemism subject to extinction in the actual condition (REZENDE et al., 2018). This global hotspot of biodiversity (MYERS et al., 2000) is at serious risk of deforestation due to land-use changes (REZENDE et al., 2018). These ongoing changes in microenvironmental conditions shape the forest structure into ecological groups. In this sense, the responses of trees growth are constrained by available resources and species demands (ABDO et al., 2016). Recent studies have investigated a set of dominant factors that drives their relations with the trees' development pattern (ALBUQUERQUE et al., 2019; GONZAGA et al., 2017; SILVA; SOUZA; VITÓRIA, 2021). In this sense, the tree growth function may support decision-making for forest restoration programs or economic plantation (SCOLFORO et al., 2017). Nevertheless, tropical forest growth modeling is not a trivial task due to its structural heterogeneity, temporal and spatial dynamics (FIEN et al., 2019). The tree growth rate is driven by several factors such as light assimilation (TANG; DUBAYAH, 2019), water availability (CAMPOE et al., 2016), silvicultural interventions (AVILA et al., 2017), age (OUYANG et al., 2019), topography, soil nutrients (SCHOLTEN et al., 2017), tree size, and neighborhood competition (ZHANG et al., 2016). According to Cruz et al. (2020), the annual growth increments are strongly regulated by the weather condition, soil water, photoperiod, and temperature at higher latitudes and altitudes. Nevertheless, the inter-tree competition has a significant effect on their accessibility to available resources (SOARES et al., 2017). Hence, the forest dynamics are influenced by the local neighborhood competition (OHEIMB et al., 2011). It is well-known that the high intensity of competition leads to a decrease in growth and recruitment rates and an increase in the mortality rate (AVILA et al., 2017; CAILLERET et al., 2016). The inter-tree competition acts decisively in the mortality, recruitment, and growth models (FERNÁNDEZ-TSCHIEDER; BINKLEY, 2018; SABATIA; BURKHART, 2012; SCHNEIDER; LAW; ILLIAN, 2006; VANCLAY et al., 2013; ZHANG; HUANG; HE, 2015), with the diameter growth being one of the most sensitive variables to its intensity (OHEIMB et al., 2011).

The inter-tree competition is usually measured by indices with mathematical assumptions to describe the spatial interactions between trees (AAKALA et al., 2013). They numerically express the effect of the neighborhood (competitor trees) on a target tree (subject

tree) (SUN et al., 2018). Most studies involve the classical approach of competition indices (KUEHNE et al., 2019; OU et al., 2019; TAVARES JÚNIOR et al., 2020), which consists of distance-dependent (IDD), semi-distance-independent (ISI), and distance-independent indices (IDI) (LEDERMANN, 2010). These indices only diverge into the criteria and formulas used to express relations between trees (CASTRO et al., 2014). Although they work appropriately, these categories are limited by not capturing the effects of local variation or by the need for tree data not commonly collected in forest inventory (LEDERMANN, 2010). There is no consensus about the best strategy since their performance depends on the type and conditions of the forest (CONTRERAS; AFFLECK; CHUNG, 2011; KUEHNE; WEISKITTEL; WASKIEWICZ, 2019). Therefore, the current state of knowledge has advanced by proposing new methods, such as competition spatialization using Geographic information system (GIS) (TÉO; FILHO; LINGNAU, 2015), indices based on airborne laser scanning (ALS) (PEDERSEN et al., 2013), spatial structure (HUI et al., 2018), crown area (KUEHNE; WEISKITTEL; WASKIEWICZ, 2019), light interception (BOECK et al., 2014), and complex networks (MONGUS et al., 2018).

Complex network (CN) is an interdisciplinary science applied to a range of fields, such as statistical physics, computer science, biology and sociology (BOCCALETTI et al., 2006; MATA, 2020). Its versatility has been allowed investigating livestock (TRIGUERO-OCAÑA et al., 2020) and vector-borne diseases (ZHANG, 2020), synchronization in the power grid (MOTTER et al., 2013), vegetation-atmosphere feedbacks (ZEMP et al., 2017), extreme-rainfall teleconnections (BOERS et al., 2019), links between fungal community and phosphorus cycling in mixed forests plantations (PEREIRA et al., 2021). The first researches of applying CN in Forest Sciences are Nakagawa et al. (2016) and Mongus et al. (2018). These pioneers' studies had revealed spatial patterns of complex network metrics concerning trees growth and survival rates. Complex networks are represented by graphs which are composed of many components - that are called nodes or vertices in the complex network context and the interactions between them are represented by links or edges (ALBERT; BARABÁSI, 2002). Forests are natural systems of tree sets arranged into an ecological network with heterogeneous interactions. Therefore, they can be described mathematically as a graph linking trees spatially. Such graphs can support researches in quantifying inter-tree competition with insights into the ecology and forest management processes in modeling.

The novelty of the current study aggregates CN into a feature selection procedure using a Genetic algorithm combined with Random Forest (GA-RF). The GA-RF has proven efficiency in modeling forest attributes in terms of accuracy and variables' importance

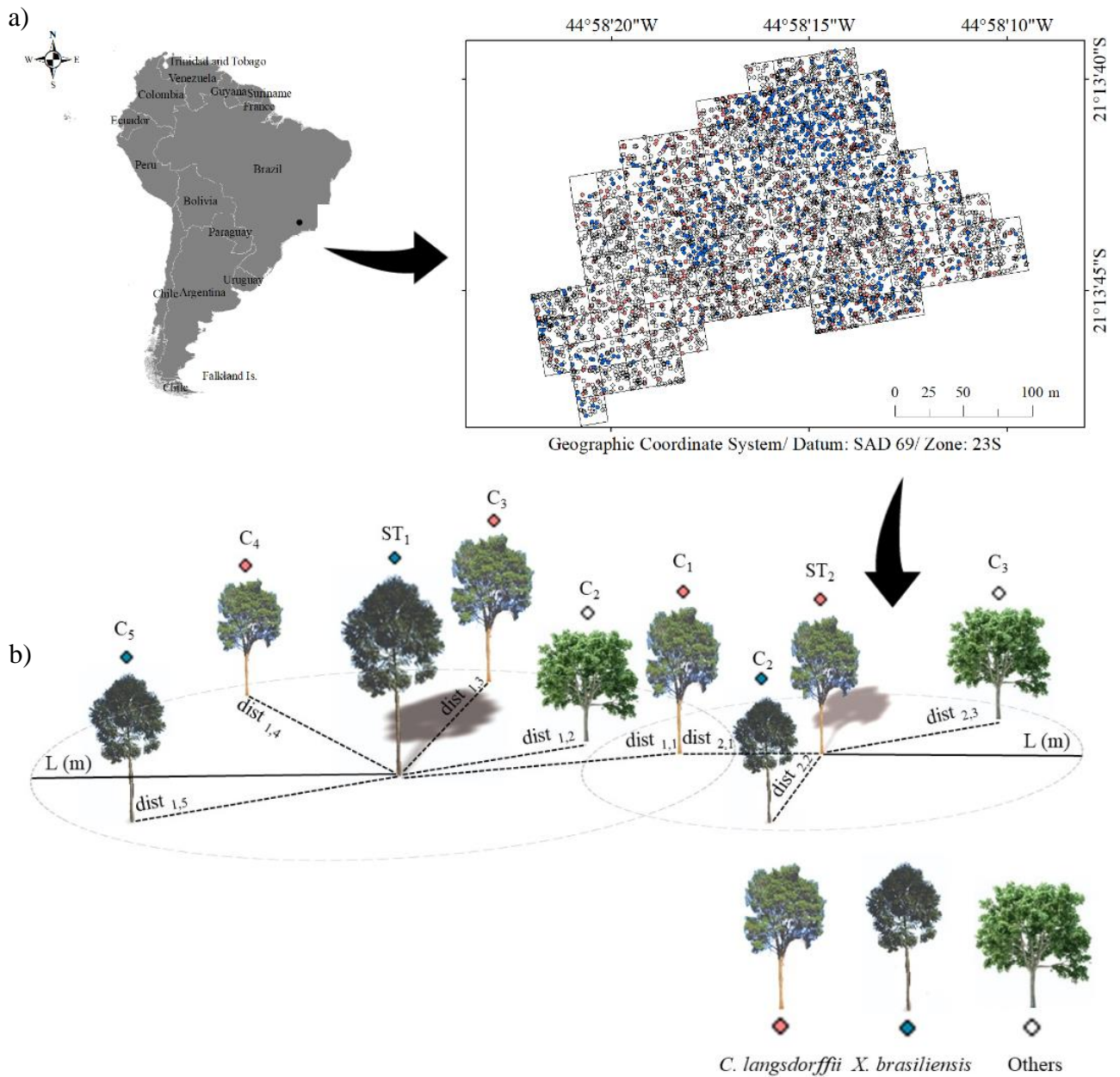
explaining (HONG et al., 2018; LOPES et al., 2021). Our proposal moved from a background developing models based only on statistical assumptions to append a valuable contribution to tree diameter modeling with ecological meaning. Therefore, our study aimed to: 1) investigate if there is a similarity between all the indices by using a dendrogram; 2) assess if the predictive performance of CN metrics is at a similar level to classical indices in quantifying inter-tree competition, and 3) verify if CN metrics incorporation into a predictive tool (GA-RF algorithm) improves the ecological interpretation of variables' influence on tree diameter increment modeling. As exposed, the methodology of this study sought to understand the predictor variables that play a decisive role in the submodel of periodic annual diameter increment (PAId).

2 MATERIAL AND METHODS

2.1 Site description and tree database

The experimental area covers 5,8 ha of Seasonal Semideciduous Montane Forest located in 21°13'40"S / 44°57'50"W coordinates at about 930 m above sea level (FIGURE 1 - a). The soil types are mainly dystrophic Red-Yellow Argisol (PVAd) and eutroferric Red Nitosol (NVef). The Köppen climate classification is a humid subtropical zone with dry winters and temperate summers (Cwb). Most of the precipitation (80%) occurs during October - March, while the dry season extends during April - September (ALVARES et al., 2013). This forest has a long-term census program measuring the DBH of trees with $DBH \geq 5\text{cm}$. These trees were identified botanically by specialists and using aluminum tags. We have chosen only the period of 2010-2017 for our analysis. Over this interval, the stand varied in density (872-953 trees ha^{-1}) and basal area (20.7-23.3 $\text{cm} \text{ha}^{-1}$). The two selected species (*Copaifera langsdorffii* and *Xylopia brasiliensis*) dominate the experimental area, achieving in 2017 the highest importance value index – IVI% (22%) and representing 23.4% of all trees. Then, the other 180 species were considered competitor trees. We have applied the Bitterlich method (Basal area factor – BAF=4) to define the competition buffer of each subject tree (FIGURE 1 - b). Later, only in 2019, we collected the geographic coordinates of all trees. We matched the spatial location of trees to their growth data into the validation process of living trees within this period.

Figure - 1 (a) Study area map, and (b) native species in spatial arrangements of trees over the Forest Reserve, and the competitor trees selection at Bitterlich method.



Where: ST_i= subject tree *i*, C_j= competitor tree *j* and dist_{ij}= the distance between a subject tree *i* and its competitor *j*, L= the maximum distance ($L = 0.5 \text{ DBH}_i / \sqrt{\text{BAF}}$) allowed for a selection of the set of competitors from each subject tree *i*.

2.2 Inter-tree competition indices

The growth conditions play a decisive role in the allometry of the trees. These conditions involve a complex set of variables and factors, which denotes an interactive system. Generally,

a substantial variation in tree allometry is shaped by their plasticity in response to competitive interactions among them. Thus, it is essential to consider the influence of competitive status as a modifier of tree allometry, which is quantified in different ways by competition indices (HUI et al., 2018; RÍO et al., 2019). The index predictive performance is subject to the particularities of the forest structure and composition. Hence, we need to test them under local conditions to determine their applicability (HUI et al., 2018). In this sense, several indices belonging to the categories found in the literature were calculated for all living trees corresponding to the two subject species. However, data were collected for all living trees, regardless of species, as information on possible competitor trees. To describe tree development, we tested four strategies. These strategies included the different categories of classic competition indices and complex network metrics.

2.2.1 Complex network metrics

We propose an inter-tree competition network derived from complex network tools to analyze the relationship between trees. Under these circumstances, the complex networks characterize the spatial arrangement of trees. Complex networks (CN) are a promising technique for describing and modeling ecological structures (COSTA et al., 2008). The CN is represented mathematically by a graph $G = (N, E)$ defined by a set of nodes $N = \{n_i\}$, and edges $E = \{e_{ij}\}$ connecting them. Thereby, a node n_i denotes a subject tree, and its interaction with the competitor (n_j) is established into a directed network if nodes i and j are connected. We also use a diameter-based function $(\sqrt{\max(dbh_i, dbh_j)})^{-1}$ to weigh the edges between the nodes i and j (w_{ij}). Under this structure, the subject tree is influenced by the size of the largest individual.

The inter-tree competition was quantified using topological metrics from the graph G structure. We used the *igraph* package (CSARDI, 2015) of R software (R CORE TEAM, 2018) to extract all numerical information. Among several metrics available in the literature, we selected the most promising ones with biological interpretations. First, we extracted the simplest topological metric, named node degree, which defines the number of nodes connected by edges to a given node. N is the total number of nodes, and a_{ij} represents the existence of a connection between the nodes i and j based on the adjacency matrix A . This means, $a_{ij} = 1$ if

there is a directed connection leaving i and arriving at j , and $a_{ij} = 0$, otherwise. The directed network allows calculating two components: the number of edges that leave the node i , named k_i^{out} (out-degree) (EQUATION 1), and the number of nodes that arrives at the node i , named k_i^{in} (in-degree) (EQUATION 2) (LATORA et al., 2017). The sum of these components results in the total degree of node i (EQUATION 3). The higher the value of this measure, the greater number of connections of that node (MO; DENG, 2019).

$$k_i^{out} = \sum_{j \neq i}^n a_{ji} \quad (1)$$

$$k_i^{in} = \sum_{j \neq i}^n a_{ij} \quad (2)$$

$$k_i = \sum k_i^{in} + k_i^{out} \quad (3)$$

Furthermore, the weighted average nearest neighbors' degree ($K_{nn,i}^w$) measures the probability of a given node to connect with nodes that have a degree similar or not with its own degree. In other words, this means that when nodes with a high degree have a larger probability to be connected with nodes that also have high degree, the network has a positive correlation (assortative networks). Otherwise, a negative correlation is presented by the network if most of the neighbors connected to a high degree nodes have a lower degree (disassortative networks) (BARRAT et al., 2004; WANG et al., 2017). In weighted networks, the $K_{nn,i}^w$ (EQUATION 4) is calculated based on the normalized weight of the connecting edges, w_{ij}/s_i , in which s_i (EQUATION 5) is the node strength.

$$k_{nn,i}^w = \frac{1}{s_i} \sum_{j=1}^N a_{ij} w_{ij} k_j \quad (4)$$

$$s_i = \sum_{j=1}^N a_{ij} w_{ij} \quad (5)$$

We also calculate the Eigenvector centrality (EC) for each node that relies on its neighbors' degrees (EQUATION 6). It considers an eigenvector of the adjacency matrix A and a constant (λ). For instance, a node i with fewer neighbors can have more influence in the network than a node j with more links, if the neighbors of node i , on average, have more connections than the neighbors of node j (MOGHADAM et al., 2019).

$$C_E(i) = \frac{1}{\lambda} \sum_{j=1}^N A_{ji} C_E(j) \quad (6)$$

Closeness centrality (CL) is another metric to assess the accessibility of a given node, where d_{ij} is the distance between i and j (EQUATION 7). This measure indicates a node centrality based on its distance from all other nodes (LIU et al., 2016). In this context, we measured distance between any two nodes according to the number of edges between them (SUN; WANG; GAO, 2016). For weighted networks, this is measured based on the weights (values) applied to the edges (AHMED; THOMO, 2017; TSIOTAS; CHARAKOPOULOS, 2018). This allows identifying the reach of information from a node to other nodes in the network (LÜ et al., 2016).

$$Cl_i = \left[\sum_{j=1, j \neq i}^N (d_{ij}) \right]^{-1} \quad (7)$$

We also take into account the coreness (CR), which measures the influence of a node based on its location in the network. Nodes with high coreness are interpreted as the most central in a network (LÜ et al., 2016). The value is measured by an iterative process of k -core decomposition (GAO et al., 2019). This process categorizes the network into hierarchical shells from the core to the periphery (LIU et al., 2016). Thus, a node i belongs to a shell layer $c(i, G)$ that consists of the coreness of node i (EQUATION 8).

$$c(i, G) = \max \{k \mid i \in C_k(G)\} \quad (8)$$

Another useful metric is the local clustering coefficient (CC) that measures the density of triangles in a network (NEWMAN, 2003). Triangles as subgraphs of the network provide a detailed view of the neighborhood interconnection. It means how many neighbors of a node are connected between them (EQUATION 9). The k_i is the number of neighbors of node i (degree of node i), $|\mathcal{E}(\Gamma_i)|$ is the number of real edges between the neighbors of node i and, $k_i(k_i - 1)$ is the maximum number of possible edges between them (total number of triangles formed by node i) (GHANBARI; JALILI; YU, 2018; TSIOTAS; CHARAKOPOULOS, 2018).

$$C_i = \frac{|\mathcal{E}(\Gamma_i)|}{k_i(k_i - 1)} \quad (9)$$

The Betweenness centrality (BC) is also an important centrality measure that denotes the node's ability to control the flow of the network (EQUATION 10). It works as a bridge connecting any two nodes through the shortest path that connects them. The higher values characterize the most central nodes in the network that often participate in the shortest paths between any pair of nodes (MAGLARAS et al., 2016). The b_{jh} corresponds to the total number of possible minimum paths between nodes j and h , and $b_{jh}(i)$ represents the number of minimum paths between them that pass through node i .

$$BC_i = \sum_{j \neq h \neq i} \frac{b_{jh}(i)}{b_{jh}} \quad (10)$$

Finally, the PageRank (PR) is a Google search engine to classify web pages relevance. It simulates the users' behavior when browsing the Web to rank pages (network nodes) and hyperlinks (edges). The PageRank value of each node is related to the probability that it will be accessed more often in a random search (EQUATION 11) (HENNI; MEZGHANI; GOUIN-VALLERAND, 2018). Where c is a constant between 0 and 1, $PR(p)$ is the PageRank value for node p and, $b_{out}(p)$ is the number of edges coming out of node p .

$$PR_i = (1 - c) + c \sum_{p \in b(i)} \frac{PR(p)}{|b_{out}(p)|} \quad (11)$$

2.2.2 Classical competition indices

In general, classical competition indices incorporate the tree size and geographic location when required to measure the inter-tree competition. They have a range of strategies and formulas to express the competition level. Here, we have tested all strategies and indices often described in literature. We applied six distance-dependent indices to synthesize the competitive influence of the neighbors' size and their distance on subject tree. The distance-independent indices reflect the effect of the entire stand on a subject tree, so we used 11 due to their ease of obtaining being location-independents. We also computed two semi-distance-independent indices that capture spatially explicit competition by only considering trees within the same plot of the subject tree (TABLE 1).

Table 1 - Classical competition indices evaluated in quantifying inter-tree competition.

Distance-dependent indices (IDD)		
Code	Author	Formula
IDD1	Hegyí (1974)	$\sum_{j=1}^n (d_j / d_i l_{ij})$
IDD2	Rouvinen and Kuuluvainen (symmetric, 1997)	$\sum_{j=1}^n d_j / l_{ij}$
IDD3	Rouvinen and Kuuluvainen (asymmetric, 1997)	$\sum_{j=1}^n \frac{(d_j / d_i)^2}{l_{ij}}$
IDD4	Martin and Ek (1984)	$\sum_{j=1}^n \frac{d_j}{d_i} \frac{1}{(l_{ij} + 1)}$
IDD5	Staebler (1951)	$\sum_{j=1}^n l_{ij}$
IDD6	Moore et al. (1973)	$\sum_{j=1}^n \frac{d_i^2}{d_i^2 + d_j^2} l_{ij}$
Distance-independent indices (IDI)		
IDI1	Daniels et al. (1986)	$\left(d_i^2 n \right) / \sum_{j=1}^n d_j^2$
IDI2	Mugasha (1989)	$\frac{\sum_{j=1}^n (d_j / d_i)}{n}$
IDI3	Lorimer (1983)	$\sum_{j=1}^n d_j / d_i$
IDI4	Looney et al. (2018)	$\sum_{j=1}^n d_j$
IDI5	Corona and Ferrara (1989)	$\sum_{j=1}^n (d_j^2 / d_i^2)$
IDI6	Tomé and Burkhart (1989)	d_i / d_{\max}
IDI7	Glover and Hool (1979)	d_i^2 / \bar{d}^2
IDI8	Stage (1973)	d_i / d_q
IDI9	Pedersen et al. (2013)	d_q / d_i
IDI10	Stage (1973)	SA_i^2 / SA_q^2
IDI11	Stage (1973)	BAL_i
Semi-distance-independent indices (ISI)		
ISI1	Stage (1973)	$SA_i^2 / SA_{q_n}^2$
ISI2	Glover and Hool (1979)	d_i^2 / \bar{d}_n^2

Where: d_i = diameter of the i -th subject tree, measured at 1.30 m – dbh (cm); d_j = diameter of the j -th competitor tree, measured at 1.30 m – dbh (cm); l_{ij} = distance between the subject tree i and its competitor j (m); n = number of competitor trees; d_{\max} = maximum dbh of the trees in the sample plot (cm); \bar{d} = arithmetic mean of dbh for trees in sample plot (cm); d_q = quadratic mean diameter (q) of sample plot (cm); SA_i = sectional area of the i -th subject tree (m²); SA_q = sectional area corresponding to the quadratic mean diameter (q) of the boles in sample plot (m²); BAL_i = sum of sectional areas of neighbor trees with larger boles than the subject tree i ; SA_{q_n} = sectional area corresponding to the quadratic mean diameter (q) of the n competing trees of the subject tree; \bar{d}_n = arithmetic mean of the dbh of the n competing trees of the subject tree.

2.2.3 Indices and metrics analysis

We tested the inter-tree competition metrics (complex network metrics and classical competition indices) to figure out trends of both methods in the growth patterns. The resulting values might lead to two directions which include similar or distinct gradients between indices and metrics. The cluster analysis investigated the similarity between the attributes of the competition metrics by using a dendrogram (LEMENKOVA, 2020) based on Euclidean distance and Ward's method. This dendrogram verified the similarity of competition metrics based on their correlation strength with the dependent variable. Additionally, we evaluated the relation between the dependent variable and competition metrics using Pearson's correlation coefficient. Even though they cover similar information, we maintain these variables for further analysis in our tree growth modeling process. This step is justified since all categories of competition metrics should be tested in accuracy and interpretability terms as inputs of growth models.

2.3 Tree diameter increment modeling

Several growth models have been applied to model tree diameter increment over the years. To predict forest development accurately, forest managers are interested in an improved understanding of how to quantify competition properly and its effects on individual tree growth (CONTRERAS; AFFLECK; CHUNG, 2011). The inter-tree competition indices may predict the growth rate, as discussed in many published research. The majority of them

applied classic competition indices with high accuracy (KUEHNE; WEISKITTEL; WASKIEWICZ, 2019; MALEKI; KIVISTE; KORJUS, 2015). The novel approach measures the inter-tree competition in social-ecological behavior. The modeling process involves the periodic annual diameter increment (PAId) as a dependent variable in our model. Hence, a range of independent variables was tested such as (i) individual tree level (TL): diameter at breast height (DBH) (cm), sectional area (SA) (cm²), geographic coordinates (X and Y); (ii) classical competition indices, and (iii) topological metrics. We also transformed DBH and SA into the square root and second power as potential predictors. We selected a total of 456 trees (*C. langsdorffii* – 154 and *X. brasiliensis* – 302) within the 2010-2017 period for our four modeling strategies (S): S₁ – distance-dependent indices (DD), S₂ – distance-independent indices (DI), S₃ – semi-distance-independent indices (SI), and S₄ – topological metrics (TM) from complex network approach. Besides, all strategies involve initial attributes at the individual tree level (TL) for each tested species separately. The modeling process considered the use of the Genetic algorithm (GA) and Random Forest (RF). In this context, we applied a GA to select a set of variables in the RF training. The GA has a stochastic searching inspired by the biological and genetic theories for solving several problems (CERRADA et al., 2015). The advantage of the algorithms' association (GA-RF) is related to the shrinkage of the problem complexity of the model (HONG et al., 2018; JADHAV; HE; JENKINS, 2018). Previously, we performance the algorithms tuning parameters (GA - population size (400); generation (10); tournament operator, crossover (0.5); mutation (0.1); stop criteria (10 generations); and RF – ntree (1500); mtry (1); and nodesize (5)) for better performance. The data processing and computational code were developed in R software, version 3.5.1 (R CORE TEAM, 2018), and *randomForest* package (LIAW; WIENER, 2002). As the problem arises at a multi-objective optimization level, the fitness function (EQUATION 12) denotes a minimum number of variables and high precision as a desirable goal. The normalized solution assumes the sum of two terms: 1) the ratio between the OOB error and the maximum value found with all variables, and 2) the ratio between the number of selected variables (n) and the total (N). Finally, the importance of the predictor variables was measured based on Increment in Mean Square Error (%IncMSE). Predictor variables with a high value of %IncMSE are considered the most important. Their omission implies a reduction in the predictive power of the model in terms of mean square error – MSE (MIAO et al., 2018).

$$fitness_i = \left(\frac{OOB\ error_i}{\max(OOB\ error)} + \frac{n_i}{N} \right) \quad (12)$$

2.4 Goodness-of-fit metrics for modeling strategies evaluation

The predictive performance of each tested strategy was evaluated according to the following criteria: R^2 - Coefficient of determination (EQUATION 13), MSE - mean squared error (EQUATION 14), RMSE - root mean square error (EQUATION 15), MBE - mean bias error (EQUATION 16), and MAE - mean absolute error (EQUATION 17). Where y_i is the measured value of PAI_d in the i observation and \hat{y}_i is the predicted value of the i observation within n observations. We have applied a score ranking procedure to guide the better model strategy selection due to a set of criteria. These integer values scores range between (1-4) for better (1) and worst (4) accuracy order. The final index is the overall of all scores reached by each method. Hence, the lower value is the most accurate strategy (THOMAS et al., 2006). Later, we verified estimates' quality using the residual plots and histograms.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (14)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (15)$$

$$MBE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n} \quad (16)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (17)$$

Furthermore, we examined the biologic consistency of tree diameter increment curves applying the derivate form of Chapman-Richards function (EQUATION 18). This growth function is well consolidated in forest management field for high precision and biological interpretation of the parameters. The benchmarking analysis is crucial to validate our strategy. We used the GA package from R (SCRUCCA, 2021) to optimize the initial parameters of the Chapman-Richards function and the minpack.lm R package (ELZHOV et al., 2016) within Levenberg–Marquardt method to fit this function. Where y' : diameter growth rate (periodic annual diameter increment – PAId), β_i : function parameters, x : dbh.

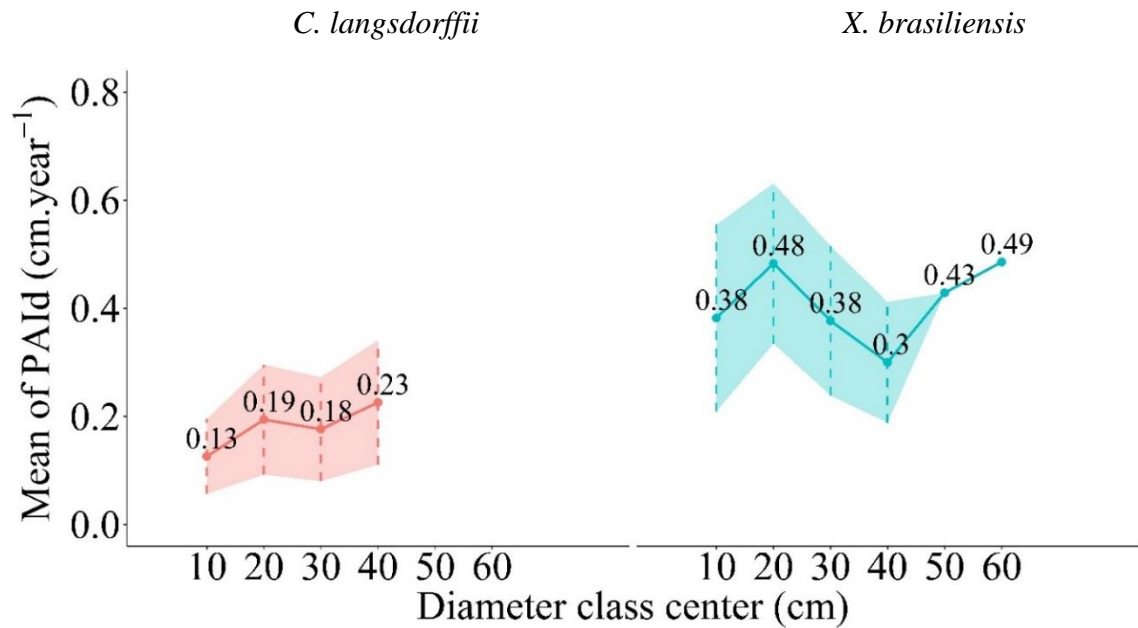
$$y' = \beta_1 \beta_2 \beta_3 \exp(-\beta_2 x) \{1 - \exp(-\beta_2 x)\}^{\beta_3 - 1} \quad (18)$$

3 RESULTS

3.1 Forest structure and inter-tree competition

The studied area had no disturbance or any ecological sustainability risk management in the evaluated period. We noted a resilience tendency of stand structure due to the diameter distribution frequencies in the uneven-aged forest type, evidenced by the reverse J-shaped curve. This stand showed a positive balance between ingrowth (11.12%) and mortality rate (3.37%), regardless of tree species. In this context, *X. brasiliensis* obtained a recruitment rate about three times higher (9.84%) than *C. langsdorffii*, and a lower mortality rate (1.75%). Initially, we may affirm that this forest reflects the natural dynamic and structure over years. The tree studied species had their density changes by a positive population gain with an increase of *X. brasiliensis* 8% higher than *C. langsdorffii*. Nevertheless, morphometric characteristics as basal area led to a higher dominance of *C. langsdorffii* with low differences (4.2%) from *X. brasiliensis*. As expected, the dominant trees have a higher growth rate than suppressed trees with an inferior growth pattern over diameter classes for *C. langsdorffii* versus *X. brasiliensis* (FIGURE 2). Although the overall data variance of the growth rate was 26% higher for *C. langsdorffii*, growth rate deviation was higher for *X. brasiliensis* when analyzed at most diameter classes. However, we have no evidence to enlighten the lower growth rate at intermediate classes of *X. brasiliensis*.

Figure 2 - Mean increment at diameter classes for each studied species.



The inter-tree competition indices/metrics presented two contrasting tendencies according to the diameter classes and mathematical formulation. They presented a clear increase/decrease trend pattern across the diameter classes (TABLE 2). Regardless of the methods, the increasing trends were more predominant, especially for *C. langsdorffii*. The set of distance-dependent indices (IDD) has similar behavior except for IDD1 and IDD4 indices. Higher values of IDD2 suggest the negative effect of competitor tree size and inverse distance towards subject trees. Individuals trees of *C. langsdorffii* (dbh<35cm) are subject to a greater competitive influence than *X. brasiliensis* individuals. The IDD5 and IDD6 indices indicate an increasing competition rate as the classes become larger for both species. Conceptually, the lower these indices, the greater the competition. Therefore, *X. brasiliensis* trees suffered less competitive pressure from competitor trees. The only exception occurred to its individuals within diameter interval 15-25 cm. In general, we have noted a superior competition capacity for *X. brasiliensis* individuals than observed in *C. langsdorffii*. The semi-distance-independent indices (ISI) capture this similar tendency with values increasing over diameter classes (Stage index, ISI2) and sectional area (Glover and Hool index, ISI1). Considering distance-independent indices (IDI), the Mugasha index (IDI2) revealed that individuals (dbh>15cm) have competitors with smaller size. Pedersen index (IDI9) demonstrated the same behavior according to the quadratic mean diameter of the plot. Looney index (IDI4) supports our findings

by denoting competitive advantages to *X. brasiliensis* individuals even with their competitors on average 70% larger (size or number) than *C. langsdorffii* competitors.

The complex network metrics were useful to quantify the inter-tree competition mapping the spatial relationship between trees (APPENDIX A). The topological metrics (TM) point out more spatial patterns for *C. langsdorffii* over diameter classes than *X. brasiliensis*. There are fewer network connections as competitor trees for individuals with $dbh > 35$ cm in both species. The same hypothesis was confirmed by the clustering coefficient (CC) since its values indicated that bigger individuals participate less in the groups' formation than smaller ones. The coreness (CR) assumes that the pivot trees of the ecological system are those with $25 \leq dbh < 45$ cm, which support the connection of all trees. Dominant trees are spatial distant from other individuals in the network as defined by the Closeness centrality (CL) metric. The Eigenvector centrality (EC) reinforces that *Copaifera langsdorffii* individuals at the inferior stratum of the forest have neighboring trees with more competitors than dominant trees. The highest values of the betweenness centrality (BC) suggest more proximity of *Xylopia brasiliensis* to other individuals than *Copaifera langsdorffii* in our system. The PageRank (PR) metric values expressed that *Copaifera langsdorffii* trees have a similar probability of competition at diameter classes. Conversely, higher competition probability was found for smaller individuals of *Xylopia brasiliensis*. The studied species showed distinct behavior of nearest neighbors degree ($K_{m,i}^w$) values since their individuals with nearest neighbors more connected had intermediate (*Copaifera langsdorffii*) and small size (*Xylopia brasiliensis*).

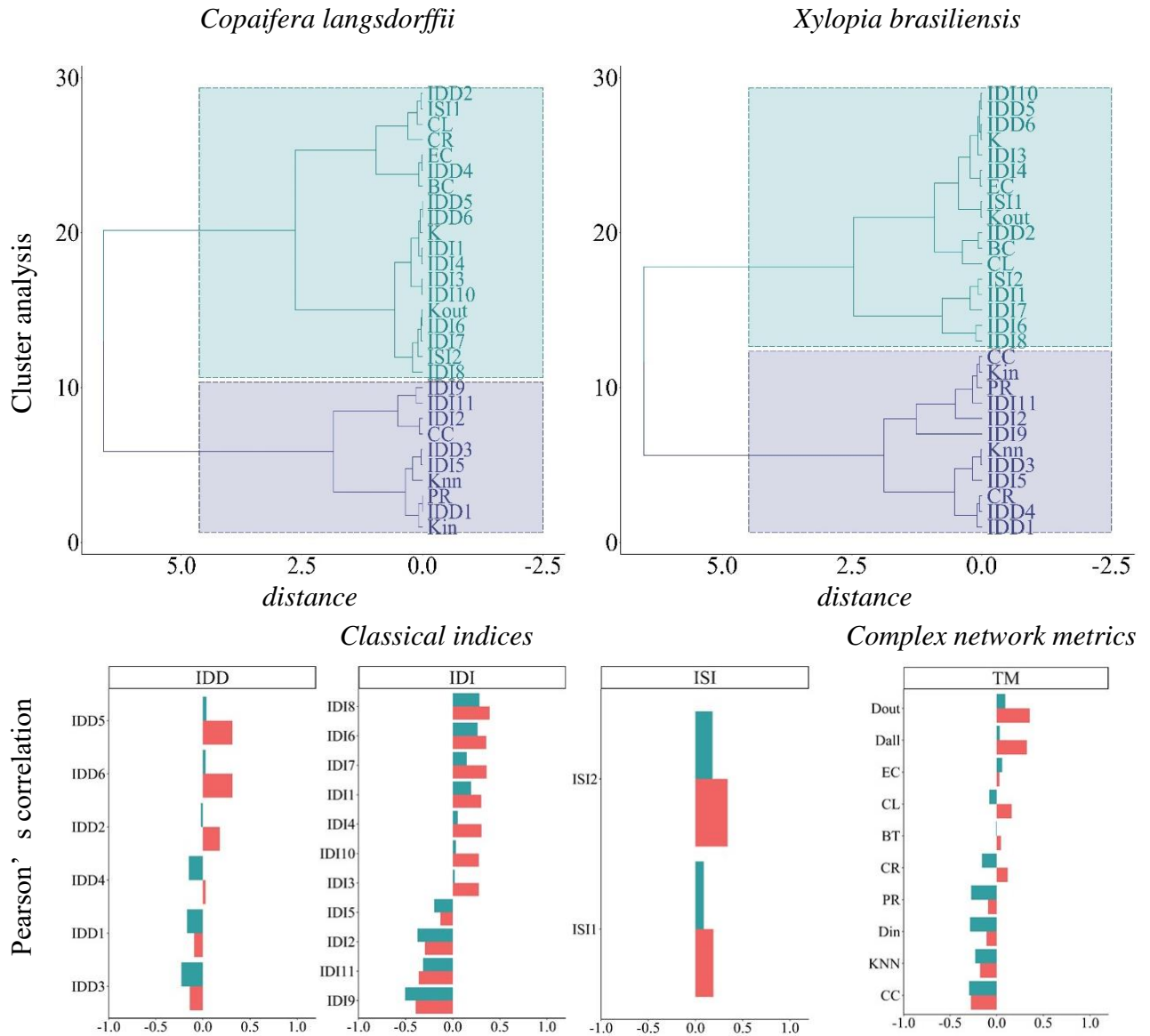
Table 2 - Competition metrics/indices average values within the subject species' diameter classes.

Indices/metrics	<i>Copaifera langsdorffii</i>				<i>Xylopia brasiliensis</i>					
	Diameter class center (cm)									
	10	20	30	40	10	20	30	40	50	60
IDD1	3.50	2.32	2.21	1.92	2.38	1.64	1.77	1.95	0.92	1.45
IDD2	31.95	50.85	66.35	78.33	28.20	36.18	54.77	84.27	46.02	86.42
IDD3	7.74	2.12	1.60	1.05	4.36	1.35	1.07	0.84	0.23	0.47
IDD4	1.45	1.57	1.58	1.55	1.28	1.16	1.33	1.55	0.76	1.25
IDD5	5.19	41.65	84.66	203.18	8.32	35.13	95.96	255.88	217.86	474.63
IDD6	2.79	31.93	68.49	174.63	4.94	26.71	80.05	232.71	204.24	438.27
IDI1	0.81	2.48	3.23	5.07	1.16	2.53	4.20	8.72	15.63	11.37
IDI2	1.49	0.62	0.52	0.39	1.25	0.62	0.45	0.31	0.22	0.25
IDI3	3.55	6.40	8.11	10.93	3.69	5.17	7.56	10.54	6.16	12.23
IDI4	37.10	143.07	246.86	446.08	47.53	115.01	237.09	460.65	308.21	728.66
IDI5	6.88	5.37	5.75	6.02	6.10	4.26	4.68	4.36	1.79	4.22

IDI6	0.26	0.56	0.71	0.87	0.32	0.60	0.86	0.96	1.00	1.00
IDI7	0.63	2.63	4.23	6.61	0.92	2.68	4.86	10.30	13.50	9.46
IDI8	0.64	1.36	1.70	2.13	0.79	1.40	1.87	2.64	2.93	2.55
IDI9	1.75	0.77	0.61	0.48	1.41	0.74	0.55	0.39	0.34	0.39
IDI10	0.31	4.14	10.34	23.81	0.61	4.48	14.08	54.74	73.23	42.07
IDI11	0.69	0.47	0.37	0.18	0.71	0.39	0.17	0.07	0.00	0.00
ISI1	1.23	8.03	12.64	28.91	2.91	8.37	21.94	85.83	244.22	129.31
ISI2	0.93	3.20	4.46	7.19	1.29	3.26	5.63	11.70	20.63	15.41
Kout	2.88	10.90	16.72	29.10	3.48	8.87	17.52	35.75	28.00	48.00
Kin	2.33	2.42	3.26	1.80	2.67	2.18	2.32	1.88	0.00	1.00
K	5.20	13.31	19.98	30.90	6.15	11.05	19.84	37.63	28.00	49.00
CC	0.47	0.21	0.16	0.09	0.37	0.21	0.11	0.08	0.06	0.04
CR	3.37	4.42	5.28	5.00	3.78	4.10	4.68	5.00	4.00	4.00
EC	2E-02	3E-02	3E-02	3E-03	1E-02	1E-02	3E-03	0E+00	0E+00	0E+00
BC	140.93	184.44	757.91	431.20	168.52	425.59	1292.82	434.88	0.00	39.00
PR	4E-04	3E-04	4E-04	3E-04	5E-04	4E-04	4E-04	3E-04	2E-04	3E-04
Knn	8.64	7.87	9.25	7.80	8.85	7.30	6.82	7.29	4.31	4.97
CL	47.22	108.53	215.03	309.97	47.71	103.66	228.59	284.92	234.90	479.97

Our cluster analysis revealed similar patterns for most of the competition indices/metrics in both species (FIGURE 3). The threshold point split the set into two groups within a heterogeneous composition. Our findings define a certain level of similarity between complex network metrics and classical competition indices. In general, they are sharing the same information to measure the inter-tree competition. Only CR and IDD4 indices changed their group for each species. Therefore, our results strongly indicate that complex networks metrics may be a very useful way as the classical competition indices corroborating with the information of the relationship between trees. Therefore, they are also appropriate to describe growth rate and the periodic annual increment of diameter (PAId). In general, the Pearson's correlations presented positive values (FIGURE 3). The Stage index (IDI8) showed the greatest positive correlation for both species, followed by IDI7, IDI6, K^{out} for *Copaifera langsdorffii* and IDI6, IDI1, ISI2 for *Xylopia brasiliensis*. Conversely, the indices of Pedersen (IDI9), Stage – *BAL* (IDI11), Mugasha (IDI2), and the clustering coefficient (CC) had higher negative values for both species.

Figure 3 - Cluster analysis of grouping the set of competition indices/metrics and Pearson's correlation of periodic annual diameter increment (PAI_d) for *Copaifera langsdorffii* ■ and *Xylopia brasiliensis* ■ .



Where: IDI= distance-independent indices, IDD= distance-dependent indices, ISI= semi-independent-distance indices, TM= topological metrics of network, ISI1= Stage (1973), ISI2= Glover and Hool (1979), IDI1= Daniels et al. (1986), IDI2= Mugasha (1989), IDI3= Lorimer (1983), IDI4= Looney et al. (2018), IDI5= Corona and Ferrara (1989), IDI6= Tomé and Burkhart (1989), IDI7= Glover and Hool (1979), IDI8= Stage (1973) based on quadratic mean diameter, IDI9= Pedersen et al. (2013), IDI10= Stage (1973) based on sectional area, IDI11= Stage (1973) based on BAL, IDD1= Hegyi (1974), IDD2= Rouvinen and Kuuluvainen (symmetric, 1997), IDD3=, Rouvinen and Kuuluvainen (asymmetric, 1997), IDD4= Martin and Ek (1984), IDD5= Staebler (1951), IDD6= Moore et al. (1973), PR= PageRank, Knn= weighted average nearest neighbors degree, EC= Eigenvector centrality, Kout= out-degree, Kin= in-degree, K= total degree of the node, CR= Coreness, CL= Closeness centrality, CC= clustering coefficient and BC= Betweenness centrality.

3.2 Individual tree diameter growth modeling performance

Variable selection under Random Forest is boosted after applying a multi-objective genetic algorithm by finding an optimized number of variables with minimum error. Our findings highlight that the subset of selected variables changed for each study species due to the diverse growth pattern of them. Under this circumstance, we reported slight precision differences for each set of tested variables. This means that all inter-tree competition methods are suitable to explain the individual tree diameter growth with acceptable limitations (TABLE 3). The species growth rate and spatial interaction of trees drove the better indices/metrics selection for each species. Therefore, we noted that complex network metrics have superior advances facing classical indices only for *C. langsdorffii* and distance independent competition indices for *X. brasiliensis*.

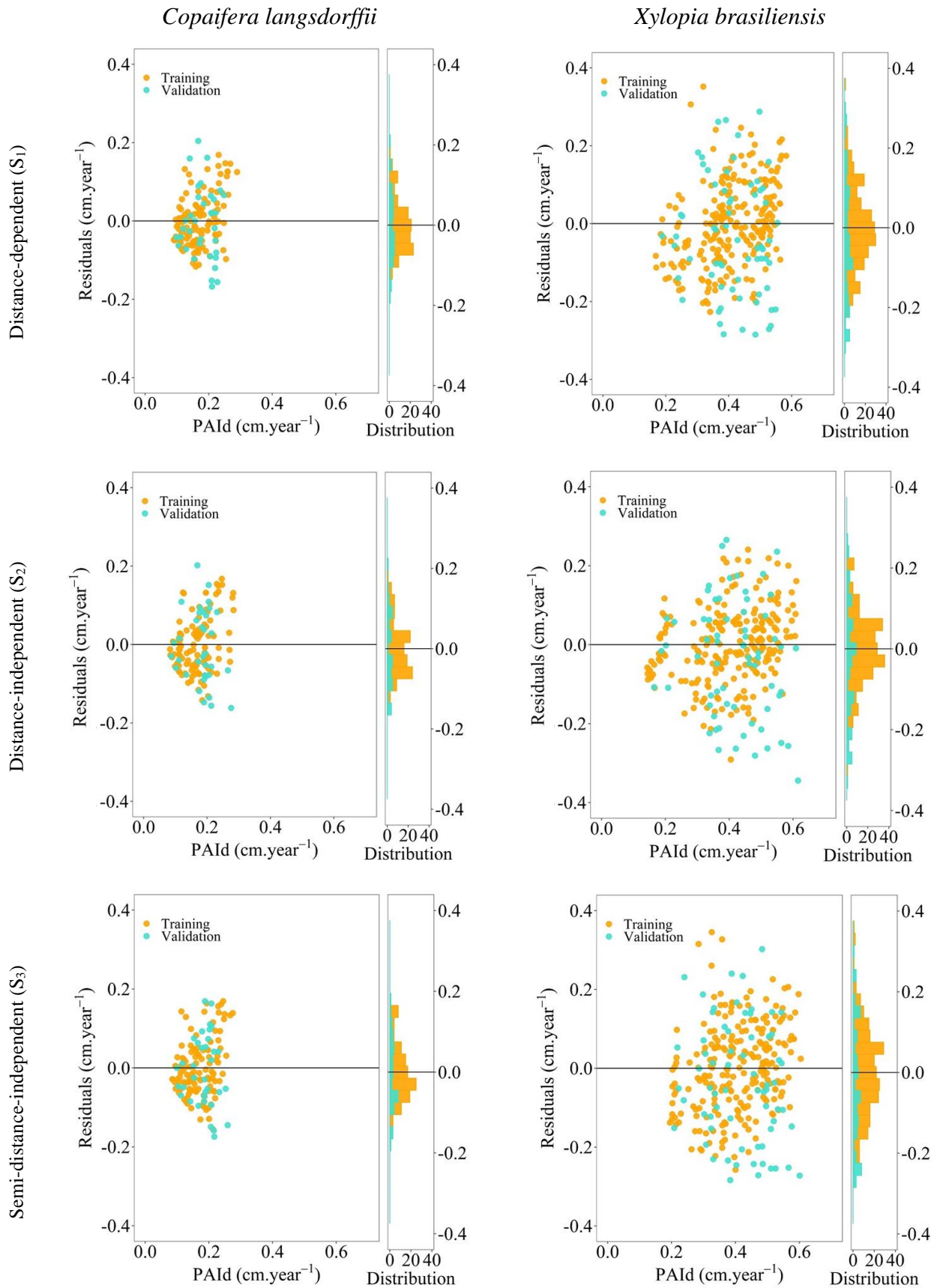
Table 3 - Statistics analysis of the periodic annual diameter increment (PAI_d) modeling strategies for *Copaifera langsdorffii* and *Xylopiya brasiliensis*.

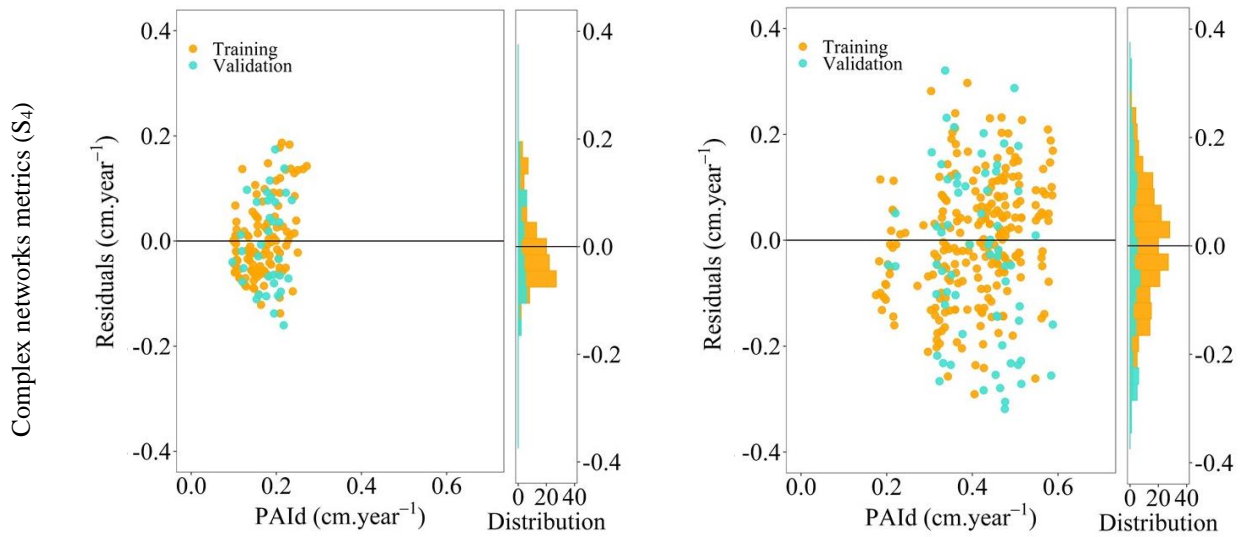
SP	Statistics	Classical competition indices						Complex networks metrics (S ₄)	
		Distance-dependent (S ₁)		Distance-independent (S ₂)		Semi-distance-independent (S ₃)		Training	Validation
		Training	Validation	Training	Validation	Training	Validation		
<i>Copaifera langsdorffii</i>	MSE	0.004	0.008 ²	0.005	0.008 ⁴	0.005	0.008 ³	0.005	0.007 ¹
	RMSE	0.065	0.089 ²	0.068	0.092 ⁴	0.071	0.090 ³	0.071	0.086 ¹
	MBE	-0.001	-0.007 ²	-0.001	-0.007 ³	-0.001	-0.010 ⁴	0.000	-0.005 ¹
	MAE	0.052	0.073 ¹	0.055	0.080 ⁴	0.057	0.077 ³	0.056	0.076 ²
	R ²	0.785	0.366 ²	0.740	0.313 ⁴	0.716	0.364 ³	0.730	0.443 ¹
	Scores	9 ^b		19 ^d		16 ^c		6 ^a	
<i>Xylopiya brasiliensis</i>	MSE	0.011	0.024 ²	0.009	0.023 ¹	0.012	0.024 ³	0.012	0.027 ⁴
	RMSE	0.104	0.154 ²	0.092	0.151 ¹	0.110	0.156 ³	0.111	0.165 ⁴
	MBE	-0.001	-0.038 ³	0.000	-0.038 ²	0.001	-0.039 ⁴	0.000	-0.034 ¹
	MAE	0.083	0.130 ²	0.073	0.122 ¹	0.089	0.133 ³	0.090	0.138 ⁴
	R ²	0.813	0.449 ²	0.848	0.493 ¹	0.778	0.433 ³	0.768	0.301 ⁴
	Scores	11 ^b		6 ^a		16 ^c		17 ^d	

Where: MSE – mean squared error, RMSE – root mean square error, MBE – mean bias error, MAE – mean absolute error, R² – Coefficient of determination, and scores ^{n, 1} The sequence order of numbers and letters defines the ranking scale and the better modeling variable strategy, respectively.

Given the residual plot analyzes (FIGURE 4), there is a slight tendency to overestimate the PAI_d for *C. langsdorffii* in all strategies by histogram distribution. It is evident that most of the strategies showed a normal distribution of residuals for *X. brasiliensis*. However, we highlighted that the distance-independent indices use (S₂) resulted in a higher concentration of residuals at lower error classes for this species.

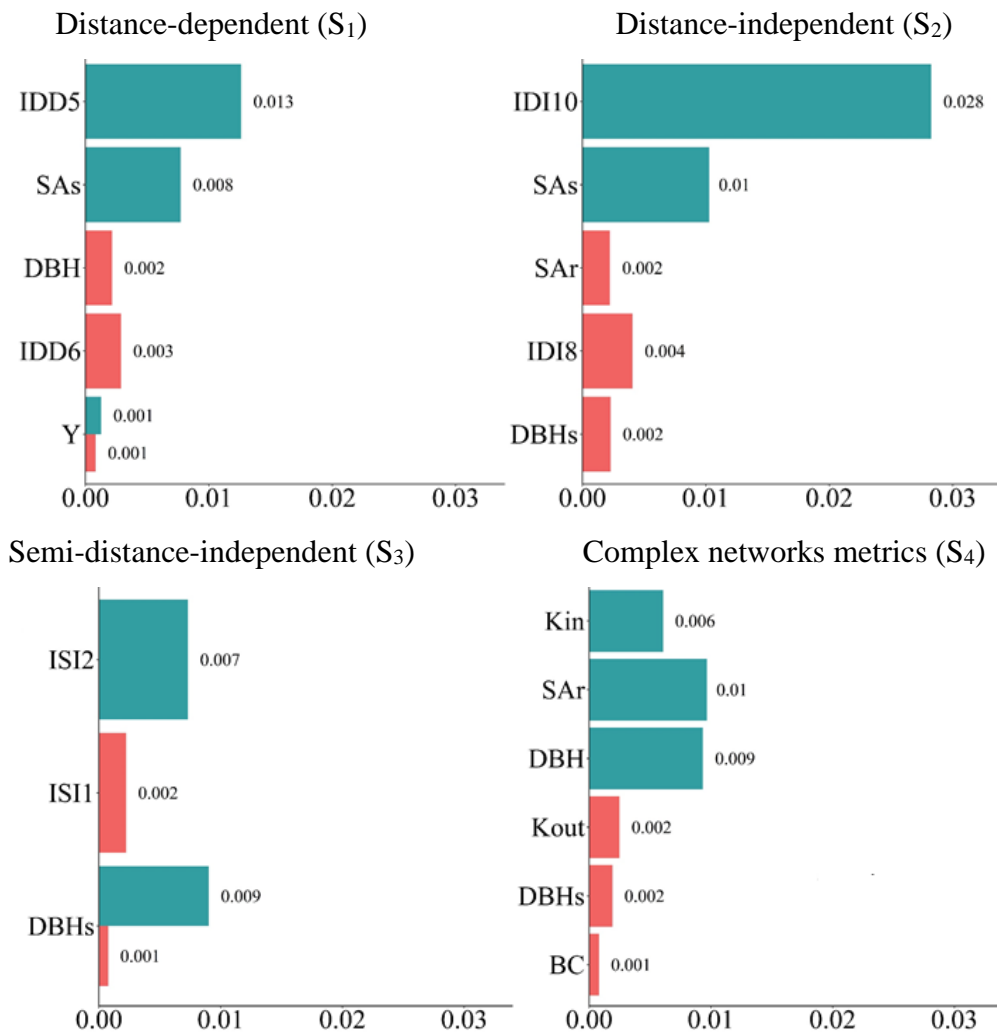
Figure 4 - Residuals plot with marginal histograms considering the modeling strategies for *Copaifera langsdorffii* and *Xylopia brasiliensis*.





The variable selection approach carries on the advanced strategy to increase the final accuracy and avoid noises (FIGURE 5). Generally, the most relevant variables are chosen to predictions' improvement in a multi-objective form. Our findings achieved an average variables' reduction of 83%. The local site may influence the growth rate since the latitude (Y coordinate) was selected for distance-dependent strategy (S_1) in both species. The DBH is usually a key source of a range of model and biological processes. For *C. langsdorffii*, beyond this variable was selected in all modeling strategies, it indirectly incorporated the Moore index (IDD6) and Stage index (IDI8) with a slight difference in importance between S_1 (distance-dependent) and S_2 (distance-independent). However, the topological metrics of network such as K_{out} and BC were selected instead of the others resulting in better estimates for this species. These metrics captured the negative effect of competitors over the subject trees by associating their spatial distribution and neighborhood density with tree growth. Concerning *X. brasiliensis*, the Stage (IDI10) and Staebler (IDD5) indices were more relevant than ISI2 and K_{in} . It means a superior influence of area occupation and neighborhood radius on the growth pattern. Nevertheless, we should note that the low performance of semi-distance-independent (S_3) and topological metrics of Complex Networks (S_4) modeling strategies are not conditioned only by the indices ISI2 and K_{in} but also by the interaction of the entire set of variables.

Figure 5 - Graphical analysis of the variables selected for *Copaifera langsdorffii* (■) and *Xylopia brasiliensis* (■) models.

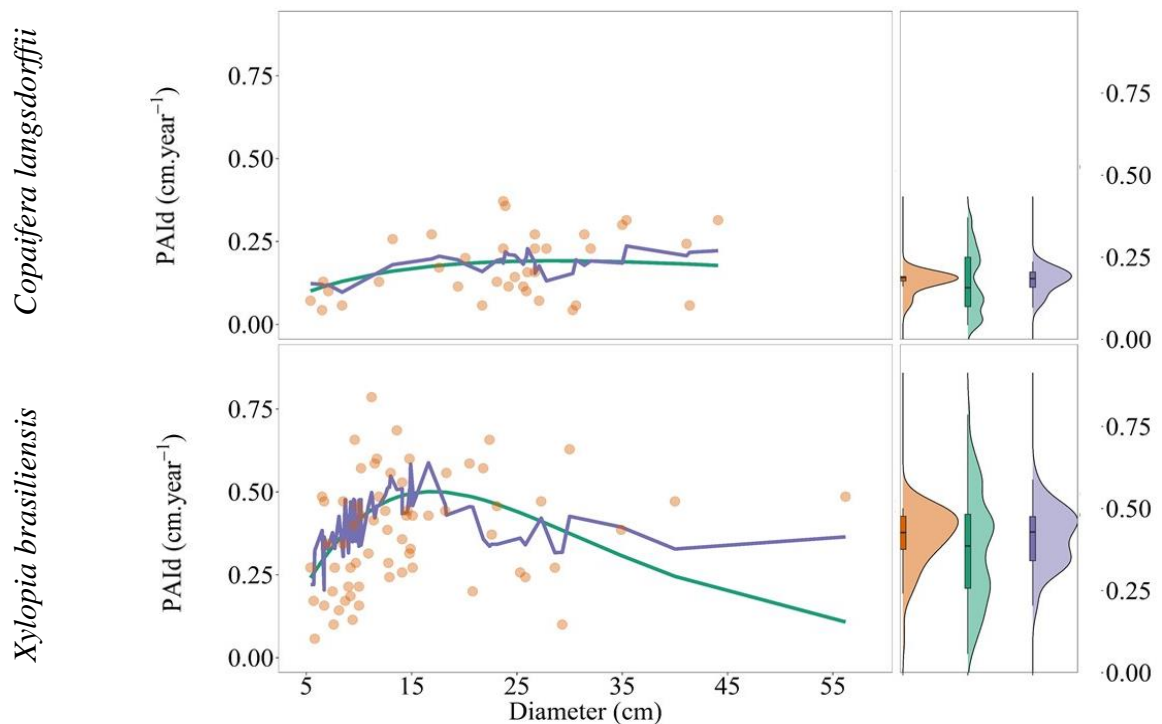


Where: Y= geographic coordinate (latitude), SAr= square root of sectional area, SAs= second power of sectional area, DBHs= square of diameter at breast height, DBH= diameter at breast height, IDD5= Staebler (1951), IDD6= Moore et al. (1973), IDI8= Stage (1973) based on quadratic mean diameter, IDI10= Stage (1973) based on sectional area, ISI1= Stage (1973), ISI2= Glover and Hool (1979), BC= Betweenness centrality, Kin= in-degree, Kout= out-degree.

The benchmarking analysis of accuracy and biological consistency suggested a positive response from the use of complex network metrics (FIGURE 6). Adding these variables, the graphical results were visually superior than Chapman-Richards function. This model parameters were significant ($p < 0.001$) for *Xylopia brasiliensis* ($\beta_1 = 19.094216$, $\beta_2 = 0.057337$, $\beta_3 = 2.645241$), and only the β_3 for *Copaifera langsdorffii* ($\beta_1 = 18.57332$, $\beta_2 = 0.01952$, $\beta_3 = 1.75503$). The individual tree growth predictions of observed data (validation set) denote a significant effect of K_{in}/SAr (*Copaifera langsdorffii*) and K_{out}/BC (*Xylopia brasiliensis*). These

curves validate the inter-tree competition as additional effects to explain these biological patterns confirming the variables selected robustness. The competition effects by these complex network metrics approximated the behavior of these curves more consistently with the natural distribution and orientation of the growth data than the Chapman-Richards function.

Figure 6 - Individual tree diameter increment predicted by Chapman-Richards function (—) and Genetic algorithm with Random forest – (GA-RF —) using complex network metrics (S_4) in validation dataset (●).



4 DISCUSSION

Individual tree growth models depend on competition indices widely used in the literature. Different competitive abilities, growth rates, and shade tolerance levels lead the species to develop specific adaptation mechanisms, varying their behavior pattern (KUEHNE et al., 2020). Defining the competitive abilities of individual trees for natural resources (light, water, and minerals) has been a constant challenge in the forestry area. It is especially attributed to the high diversity of a tropical forest. The dimensional characteristics of trees are the result of factors that conditioned their development in the forest. Then, they cannot be neglected, as they reflect the growth potential of individuals. Notably, we observed in our study that DBH and SA were the variables with the greatest effective participation in the growth. In addition to these variables, crown attributes and competition have been recognized by their association

with diameter growth (CUNHA et al., 2016). These are considered as modifying factors of potential growth that allow incorporating the nature of the interactions between trees. Therefore, it is essential to insert the variables that denote competition in the modeling of individual trees, as they can improve the prediction of the dimensions and dynamics of tree growth (CARRIJO et al., 2020). Our results show an evident difference in growth pattern of the species under study. In average terms of diametric increment, *C. langsdorffii* showed slower growth. Such behavior is given to its greater responsiveness to the competition. This can be confirmed by its higher values of linear association between most categories of indices/metrics of competition with PAI_d , in comparison with *X. brasiliensis*. For both species, the indices with the greatest positive and negative linear association were, respectively, Stage index (IDI8) and Pedersen index (IDI9). Although they are structurally different indices, both are commonly formed by the quadratic mean diameter of the plot. This whole-stand variable, when incorporated into competition measurements distance-independent, corresponds to the hierarchical position of the subject tree within the plot (MORENO et al., 2017; SAUD et al., 2016). The same relative dimension (ratio) used by IDI8 was evaluated in the study of Sharma et al. (2019) that observed the increase in the diameter increment of the species *Fagus sylvatica* L. with the reduction of competition, expressed by the increase in the ratio between the DBH and the quadratic mean diameter.

The variation of the H-D relation over time affects the diameter increment, then, it is essential to comprehend it as part of forest development. Based on this context, it should be noted that most individuals of *X. brasiliensis* are established in the emergent layer of the forest canopy, as they have plausible growth rates (SCOLFORO et al., 2017). Therefore, these individuals reach the emergent layer of the canopy more quickly because they invest more in height to the detriment of the diameter. This behavior can lead to less susceptibility to competition (SHARMA; BRUNNER, 2017). Additionally, its crowns were formed by sparse orthotropic branches emerging from the main stem (TERRA et al., 2018), enabling greater access to light. These factors may explain a lesser dependence of this species, classified as early secondary, in relation to the interactions with its neighborhood. This is the opposite of what happens with *C. langsdorffii* individuals, since this species is classified as late secondary and climax. Thus, they are subject to greater competition for light and soil depth, as they are still looking for canopy dominance through investment in height and the formation of symmetrical crown (COSTA et al., 2012). This context corroborates the study developed by Stadt et al. (2007) in mature boreal mixed forests that reported poorer fits of competition indices for shade-

intolerant species than shade-tolerant ones. These authors credited this fact to the frequent occupation of intolerant surviving trees in dominant positions in the canopy, which consequently suffered less competitive intensity. On the other hand, as most trees of tolerant species occur in the sub-canopy, they are subject to a greater variety of competition. Thus, stratification by competing species or even by ecosite is advisable to improve the model's performance.

In this sense, our comparative approach of different competition indices/metrics demonstrated that there is no consensus of the most suitable index/metric to define the inter-tree competition for all species. This fact is explained since their application becomes vulnerable due to the local and genetic conditions inherent to the tree (CONTRERAS; AFFLECK; CHUNG, 2011; SHARMA; BRUNNER, 2017). Therefore, the complex network emerges as a design that can better represent competition patterns. Our findings suggest that the network structure dealt with inter-tree competition in a more naturally interpretive way. This technique provided a slight superiority in addressing the effect of competition on the periodic annual increment in diameter of *C. langsdorffii*. Thus, the categories of competition metrics/indices that showed improvements in obtaining growth estimates for each species were distance-independent indices and topological metrics, respectively, for *X. brasiliensis* and *C. langsdorffii*. This behavior has been observed by several studies in the literature that have comparatively addressed categories of indices or even indices belonging to the same category. As examples of these studies, the following can be cited: Ledermann (2010), Castro et al. (2014), Maleki et al. (2015), Kuehne et al. (2019) and Curto et al. (2020). In this way, our findings indicate that the quality of fit of competition indices may be considered species-specific (FUKUMOTO et al., 2020). Different indices resulted in the best adjustments for the *X. brasiliensis* and *C. langsdorffii* in each modeling strategy. About this latest species, the Stage index (IDI8) and Moore index (IDD6) stood out within their respective categories, distance-independent and distance-dependent, as the best predictors of PAI_d . Both were also reported by Curto et al. (2020) as the best competition indicators among the different indices tested in their categories for an overstocked stand of *Araucaria angustifolia* (Bertol.) Kuntze. In the study by Tavares Júnior et al. (2020) only indices semi-independent of distance were applied to assess the increment in diameter of individual trees in different fragments of the Atlantic Forest. These authors observed that there was no superiority of a single index. The variation in performance of indices in response to fragment type supports this assertion. Our study also

confirmed that the most appropriate index depends on the type of species, being the Stage index (ISI1) for the species *C. langsdorffii* and the Glover and Hool index (ISI2) for *X. brasiliensis*.

Another relevant result revealed a notable pattern, through the positioning of the strategy based on distance-dependent indices (S_1), as the second-best strategy for both species. This result is supported by the effect of neighborhood interactions on growth in diameter, which can confer the efficiency of spatial indices as predictors of growth (MALEKI; KIVISTE; KORJUS, 2015). In this sense, the complex network is also a promising alternative to represent distance dependence. This allows examining the spatial distribution patterns of individual trees (MONGUS et al., 2018). The statistical properties of a competition network structure help us to understand plant population dynamics (NAKAGAWA; YOKOZAWA; HARA, 2016). Mongus et al. (2018) related that clusters and betweenness centralities have more influence on the development of each tree than the parameters commonly used, such as the number of a tree's competitors and distances between them. The betweenness centrality was one of the variables selected as a predictor of the diameter increment in *C. langsdorffii* individuals, which certainly contributed to the greater precision of the S_4 strategy. This metric denotes greater importance for trees with greater participation in the set of competitors of other trees. On the contrary, the *X. brasiliensis* diameter increment estimates presented lower precision. It can be attributed to the choice of the metric, in-degree of the nodes, as the only modifying factor for potential growth. Therefore, our results reinforced that only the number of competitors for a certain individual is an insufficient parameter to express the effect of competition on their pattern of development within a forest.

Modeling the individual tree periodic annual diameter increment (PAI_d) has been an arduous task due to the interaction of several factors. Both reasons that drive advances in this matter are: i) the need for more accurate quantification of competition since it is required as an input in the development of growth and production models at the level of individual trees (CONTRERAS; AFFLECK; CHUNG, 2011) and ii) the difficulty in modeling the complex and non-linear nature of individual tree growth (VIEIRA et al., 2018). Therefore, we built in the current study modeling strategies based on the Random Forest (RF) regression method. This model stands out for its ability to deal with the non-linear relationship between the predictor variables and the response variable without statistical assumptions (OU; LEI; SHEN, 2019). The application of RF has increased due to its efficiency in providing reliable estimates. The studies by Ou et al. (2019) and Tavares Júnior et al. (2020) are examples of this statement. Additionally, the use of Genetic Algorithm (GA) applied together with the RF, called optimized

RF, offers the opportunity to generate better results by selecting the optimal combination of variables (HONG et al., 2018). In this context, the findings of this study showed that the adequacy of strategies was different for each species. This fact reveals the importance of selecting an optimized set of competition indices /metrics and variables that enable its assertive use to improve growth estimates. In this context, it is noteworthy that the machine learning technique was not sensitive to the local structure of the data. Besides that, these algorithms did not choose the ideal set of variables (JADHAV; HE; JENKINS, 2018). Thus, the estimates provided by the RF regression in each modeling strategy may improve or not, as they depend on the preliminary task of selection of variables carried out through a random search made by GA. Another way to improve estimates is to explore better the effect of variables that can really contribute to understanding the response variable. Although the complex networks have not configured an expressive accuracy gain over to the classical competition indices, their flexible structure offers an advantage to allow including factors that drive the interaction between trees in the weighting of connections. Future research can broaden the scope of this study by incorporating shade tolerance, aspects of soil, climate, water, light and water availability, and other attributes to compile large data sets as characteristic weights of interactions into the complex network. This approach allows us the analysis of several scenarios corresponding to silvicultural treatments (MONGUS et al., 2018). For this reason, this study motivates scientific progress compared to the existing literature to meet the needs of developing models of growth and yield at the level of the tree by using more accurate simulations, collaborating with more assertive decision-making for forest management.

5 CONCLUSION

The tropical forest resilience has high dependence on the diametric structure over years, and each individual tree has adaptive mechanisms to surpass the negative effects of competition. In general, tree species also have a wide range of interaction between them, which drives the diameter growth patterns as observed in *X. brasiliensis* and *C. langsdorffi*. The inter-tree competition impacts in our studied species are heterogeneous at the same ecological site. The periodic annual increment of tree diameter is highly associated with the neighborhood size, spatial distribution of competitors' trees, tree size or their position of the canopy stratum, and their connections into a network structure. All inter-tree competition indices/metrics categories are suitable to model our dependent variable. Considering the expansion of concepts'

formulation, some distance-independent indices and complex network metrics were more accurate for *X. brasiliensis* and *C. langsdorffii*, respectively. Both the complex network metrics, K_{out} and BC, added ecological meaning to growth modeling by considering the density of neighboring competitors and the proximity of the subject tree to an arrangement of trees under competition relationship. Even though the Chapman-Richards growth function is highlighted in forest management, the genetic algorithm/random forest associated with complex network metrics were superior to describe the individual growth rate of tree diameter. The applied model provides biologically reasonable estimates of diameter growth within the environmental conditions of the validation dataset. This finding denotes the applicability of using complex networks metrics to encompass ecological meaning and growth models' generalization improvements. Finally, we hope that our work will encourage the scientific community to apply complex network theory to describe the relationship between trees with valuable insights for forest management.

ACKNOWLEDGEMENTS

The authors are especially grateful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brazil) for the financial support under Finance Code 001, Federal University of Lavras (UFLA – Brazil) for providing the data, Forest Management Laboratory (LEMAF – Brazil) for helping in field campaigns. Angélica S. Mata thanks the support from FAPEMIG (Grant N°. APQ-01294-21) and CNPq (Grant N°. 423185/2018-7).

REFERENCES

- AAKALA, T. et al. Influence of competition and age on tree growth in structurally complex old-growth forests in northern Minnesota, USA. **Forest Ecology and Management**, Amsterdam, v. 308, p. 128–135, 2013.
- ABDO, M. T. V. N. et al. Pioneer tree responses to variation of soil attributes in a tropical semi-deciduous forest in Brazil. **Journal of Sustainable Forestry**, Philadelphia, v. 36, n. 2, p. 134–147, 2016.
- AHMED, A.; THOMO, A. Computing source-to-target shortest paths for complex networks in RDBMS. **Journal of Computer and System Sciences**, San Diego, v. 89, p. 114–129, 2017.
- ALBERT, R.; BARABÁSI, A. L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, College Park, v. 74, n. 1, p. 47–97, 2002.
- ALBUQUERQUE, R. P. et al. Tree-ring formation, radial increment and climate–growth relationship: assessing two potential tree species used in Brazilian Atlantic forest restoration projects. **Trees**, New York, v. 33, p. 877–892, 2019.
- ALVARES, C. A. et al. Köppen’s climate classification map for Brazil. **Meteorologische Zeitschrift**, Stuttgart, v. 22, n. 6, p. 711–728, 2013.
- AVILA, A. L. de et al. Recruitment, growth and recovery of commercial tree species over 30 years following logging and thinning in a tropical rain forest. **Forest Ecology and Management**, Amsterdam, v. 385, p. 225–235, 2017.
- BARRAT, A. et al. The architecture of complex weighted networks. **Proceedings of the National Academy of Sciences**, Washington, v. 101, n. 11, p. 3747–3752, 2004.
- BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics Reports**, Amsterdam, v. 424, n. 4–5, p. 175–308, 2006.
- BOECK, A. et al. Predicting tree mortality for European beech in southern Germany using spatially explicit competition indices. **Forest Science**, Bethesda, v. 60, n. 4, p. 613–622, 2014.
- BOERS, N. et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. **Nature**, London, v. 566, n. 7744, p. 373–377, 2019.
- CAILLERET, M. et al. A synthesis of radial growth patterns preceding tree mortality. **Global Change Biology**, Oxford, v. 23, n. 4, p. 1675–1690, 2016.
- CAMPOE, O. C. et al. Meteorological seasonality affecting individual tree growth in forest plantations in Brazil. **Forest Ecology and Management**, Amsterdam, v. 380, p. 149–160, 2016.
- CARRIJO, J. V. N. et al. The growth and production modeling of individual trees of *Eucalyptus urophylla* plantations. **Journal of Forestry Research**, Harbin, v. 31, n. 5, p. 1663–1672, 2020.

- CASTRO, R. et al. Competição em Nível de Árvore Individual em uma Floresta Estacional Semidecidual. **Silva Lusitana**, Oeiras, v. 22, n. 1, p. 43–66, 2014.
- CERRADA, M. et al. Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. **Sensors**, Basel, v. 15, n. 9, p. 23903–23926, 2015.
- CONTRERAS, M. A.; AFFLECK, D.; CHUNG, W. Evaluating tree competition indices as predictors of basal area increment in western Montana forests. **Forest Ecology and Management**, Amsterdam, v. 262, n. 11, p. 1939–1949, 2011.
- COSTA, L. da F.; RODRIGUES, F. A.; CRISTINO, A. S. Complex networks: The key to systems biology. **Genetics and Molecular Biology**, Ribeirão Preto, v. 31, n. 3, p. 591–601, 2008.
- COSTA, M. do P. et al. Allometry and architecture of *Copaifera langsdorffii* (Desf.) Kuntze (fabaceae) in neotropical physiognomies in southeastern Brazil. **Ciência Florestal**, Santa Maria, v. 22, n. 2, p. 223–240, 2012.
- CRUZ, M.; LIEBERMAN, D.; LIEBERMAN, M. Tropical Tree Growth and Longevity: Validation of Growth Simulation, a Bootstrapping Model. **Journal of Sustainable Forestry**, Philadelphia, v. 39, n. 7, p. 674–691, 2020.
- CSARDI, G. **Package ‘igraph’**: Network Analysis and Visualization. Version 1.0.0, p. 1-431, 2015.
- CUNHA, T. A. DA; FINGER, C. A. G.; HASENAUER, H. Tree basal area increment models for Cedrela, Amburana, Copaifera and Swietenia growing in the Amazon rain forests. **Forest Ecology and Management**, Amsterdam, v. 365, p. 174–183, 2016.
- CURTO, R. D. A. et al. Effectiveness of competition indices for understanding growth in an overstocked stand. **Forest Ecology and Management**, Amsterdam, v. 477, p. 118472, 2020.
- ELZHOV, T. V et al. **Package “minpack.lm”**: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. Version 1.2-1, p. 1–14, 2016.
- FERNÁNDEZ-TSCHIEDER, E.; BINKLEY, D. Linking competition with Growth Dominance and production ecology. **Forest Ecology and Management**, Amsterdam, v. 414, p. 99–107, 2018.
- FIEN, E. K. P. et al. Drivers of individual tree growth and mortality in an uneven-aged, mixed-species conifer forest. **Forest Ecology and Management**, Amsterdam, v. 449, p. 117446, 2019.
- FUKUMOTO, K. et al. Evaluation of individual distance-independent diameter growth models for Japanese cedar (*Cryptomeria japonica*) trees under multiple thinning treatments. **Forests**, Basel, v. 11, n. 344, p. 1–13, 2020.
- GAO, L. et al. Coreness variation rule and fast updating algorithm for dynamic networks. **Symmetry**, Basel, v. 11, n. 477, p. 1–10, 2019.

- GHANBARI, R.; JALILI, M.; YU, X. Correlation of cascade failures and centrality measures in complex networks. **Future Generation Computer Systems**, Amsterdam, v. 83, p. 390–400, 2018.
- GONZAGA, A. P. D. et al. Brazilian Deciduous Tropical Forest enclaves: floristic, structural and environmental variations. **Revista Brasileira de Botânica**, São Paulo, v. 40, n. 2, p. 417–426, 2017.
- HENNI, K.; MEZGHANI, N.; GOUIN-VALLERAND, C. Unsupervised graph-based feature selection via subspace and pagerank centrality. **Expert Systems with Applications**, Oxford, v. 114, p. 46–53, 2018.
- HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, Amsterdam, v. 630, p. 1044–1056, 2018.
- HUI, G. et al. A novel approach for assessing the neighborhood competition in two different aged forests. **Forest Ecology and Management**, Amsterdam, v. 422, p. 49–58, 2018.
- JADHAV, S.; HE, H.; JENKINS, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. **Applied Soft Computing**, Amsterdam, v. 69, p. 541–553, 2018.
- KUEHNE, C. et al. Comparing strategies for representing individual-tree secondary growth in mixed-species stands in the Acadian Forest region. **Forest Ecology and Management**, Amsterdam, v. 459, p. 117823, 2020.
- KUEHNE, C.; WEISKITTEL, A. R.; WASKIEWICZ, J. Comparing performance of contrasting distance-independent and distance-dependent competition metrics in predicting individual tree diameter increment and survival within structurally-heterogeneous, mixed-species forests of Northeastern United States. **Forest Ecology and Management**, Amsterdam, v. 433, p. 205–216, 2019.
- LATORA, V.; NICOSIA, V.; RUSSO, G. **Complex networks: Principles, Methods and Applications**. Cambridge: Cambridge University Press, 2017.
- LEDERMANN, T. Evaluating the performance of semi-distance-independent competition indices in predicting the basal area growth of individual trees. **Canadian Journal of Forest Research**, Ottawa, v. 40, p. 796–805, 2010.
- LEMENKOVA, P. R Libraries {dendextend} and {magrittr} and Clustering Package scipy.cluster of Python For Modelling Diagrams of Dendrogram Trees. **Carpathian Journal of Electronic and Computer Engineering**, Baia Mare, v. 13, n. 1, p. 5–12, 2020.
- LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R news**, Wien, v. 2, p. 18–22, 2002.
- LIU, Y. et al. Identify influential spreaders in complex networks, the role of neighborhood. **Physica A - Statistical Mechanics and its Applications**, Amsterdam, v. 452, p. 289–298, 2016.

- LOPES, I. L. e et al. A comparative approach of methods to estimate machine productivity in wood cutting. **International Journal of Forest Engineering**, Philadelphia, v. 33, n.1, p. 43–55, 2022.
- LÜ, L. et al. The H-index of a network node and its relation to degree and coreness. **Nature Communications**, London, v. 7, p. 1–7, 2016.
- MAGLARAS, L. A. et al. Social internet of vehicles for smart cities. **Journal of Sensor and Actuator Networks**, Basel, v. 5, n. 3, p. 1–22, 2016.
- MALEKI, K.; KIVISTE, A.; KORJUS, H. Analysis of individual tree competition effect on diameter growth of silver birch in Estonia. **Forest Systems**, Madrid, v. 24, n. 2, p. 1–13, 2015.
- MATA, A. S. da. Complex Networks: a Mini-review. **Brazilian Journal of Physics**, São Paulo, v. 50, p. 658–672, 2020.
- MIAO, S. et al. Random Forest Algorithm for the Relationship between Negative Air Ions and Environmental Factors in an Urban Park. **Atmosphere**, Basel, v. 9, n. 463, p. 1–13, 2018.
- MO, H.; DENG, Y. Identifying node importance based on evidence theory in complex networks. **Physica A: Statistical Mechanics and its Applications**, Amsterdam, v. 529, p. 121538, 2019.
- MOGHADAM, H. E. et al. Complex networks analysis in Iran stock market: The application of centrality. **Physica A: Statistical Mechanics and its Applications**, Amsterdam, v. 531, p. 121800, 2019.
- MONGUS, D. et al. Predictive analytics of tree growth based on complex networks of tree competition. **Forest Ecology and Management**, Amsterdam, v. 425, p. 164–176, 2018.
- MORENO, P. C. et al. Individual-tree diameter growth models for mixed *Nothofagus* second growth forests in southern Chile. **Forests**, Basel, v. 8, n. 12, p. 1–19, 2017.
- MOTTER, A. E. et al. Spontaneous synchrony in power-grid networks. **Nature Physics**, London, v. 9, p. 191–197, 2013.
- MYERS, N. et al. Biodiversity hotspots for conservation priorities. **Nature**, London, v. 403, p. 853–858, 2000.
- NAKAGAWA, Y.; YOKOZAWA, M.; HARA, T. Complex network analysis reveals novel essential properties of competition among individuals in an even-aged plant population. **Ecological Complexity**, Amsterdam, v. 26, p. 95–116, 2016.
- NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, Philadelphia, v. 45, n. 2, p. 167–256, 2003.
- OHEIMB, G. VON et al. Individual-tree radial growth in a subtropical broad-leaved forest: The role of local neighbourhood competition. **Forest Ecology and Management**, Amsterdam, v. 261, n. 3, p. 499–507, 2011.
- OU, Q.; LEI, X.; SHEN, C. Individual tree diameter growth models of Larch-Spruce-Fir

mixed forests based on machine learning algorithms. **Forests**, Basel, v. 10, n. 187, p. 1–20, 2019.

OUYANG, S. et al. Effects of stand age, richness and density on productivity in subtropical forests in China. **Journal of Ecology**, Malden, v. 107, n. 5, p. 2266–2277, 2019.

PEDERSEN, R. Ø. et al. On the evaluation of competition indices - The problem of overlapping samples. **Forest Ecology and Management**, Amsterdam, v. 310, p. 120–133, 2013.

PEREIRA, A. P. de A. et al. Nitrogen-fixing trees in mixed forest systems regulate the ecology of fungal community and phosphorus cycling. **Science of the Total Environment**, Amsterdam, v. 758, 2021.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2018.

REZENDE, C. L. et al. From hotspot to hopespot: An opportunity for the Brazilian Atlantic Forest. **Perspectives in Ecology and Conservation**, Oxford, v. 16, n. 4, p. 208–214, 2018.

RÍO, M. del et al. Tree allometry variation in response to intra- and inter-specific competitions. **Trees - Structure and Function**, New York, v. 33, n. 1, p. 121–138, 2019.

SABATIA, C. O.; BURKHART, H. E. Competition among loblolly pine trees: Does genetic variability of the trees in a stand matter? **Forest Ecology and Management**, Amsterdam, v. 263, p. 122–130, 2012.

SAUD, P. et al. Using quadratic mean diameter and relative spacing index to enhance height-diameter and crown ratio models fitted to longitudinal data. **Forestry**, Oxford, v. 89, p. 215–229, 2016.

SCHNEIDER, M. K.; LAW, R.; ILLIAN, J. B. Quantification of neighbourhood-dependent plant growth by Bayesian hierarchical modelling. **Journal of Ecology**, Malden, v. 94, n. 2, p. 310–321, 2006.

SCHOLTEN, T. et al. On the combined effect of soil fertility and topography on tree growth in subtropical forest ecosystems—a study from SE China. **Journal of Plant Ecology**, Oxford, v. 10, n. 1, p. 111–127, 2017.

SCOLFORO, H. F. et al. A new model of tropical tree diameter growth rate and its application to identify fast-growing native tree species. **Forest Ecology and Management**, Amsterdam, v. 400, p. 578–586, 2017.

SCRUCCA, L. **Package “GA”**: Genetic Algorithms. Version 3.2.2, p. 1–45, 2021.

SHARMA, R. P. et al. Generalized nonlinear mixed-effects individual tree diameter increment models for beech forests in Slovakia. **Forests**, Basel, v. 10, n. 451, p. 1–24, 2019.

SHARMA, R. P.; BRUNNER, A. Modeling individual tree height growth of Norway spruce and Scots pine from national forest inventory data in Norway. **Scandinavian Journal of Forest Research**, Oslo, v. 32, n. 6, p. 501–514, 2017.

SILVA, J. L. A.; SOUZA, A. F.; VITÓRIA, A. P. Historical and current environmental selection on functional traits of trees in the Atlantic Forest biodiversity hotspot. **Journal of Vegetation Science**, Malden, v. 32, p. e13049, 2021.

SOARES, A. A. V. et al. Development of stand structural heterogeneity and growth dominance in thinned Eucalyptus stands in Brazil. **Forest Ecology and Management**, Amsterdam, v. 384, p. 339–346, 2017.

STADT, K. J. et al. Evaluation of competition and light estimation indices for predicting diameter growth in mature boreal mixed forests. **Annals of Forest Science**, Les Ulis, v. 64, p. 477–490, 2007.

SUN, M.; WANG, Y.; GAO, C. Visibility graph network analysis of natural gas price: The case of North American market. **Physica A - Statistical Mechanics and its Applications**, Amsterdam, v. 462, p. 1–11, 2016.

SUN, S.; CAO, Q. V.; CAO, T. Evaluation of distance-independent competition indices in predicting tree survival and diameter growth. **Canadian Journal of Forest Research**, Ottawa, v. 49, n. 5, p. 440–446, 2018.

TANG, H.; DUBAYAH, R. Erratum: Light-driven growth in Amazon evergreen forests explained by seasonal variations of vertical canopy structure (Proceedings of the National Academy of Sciences, v. 114, n.10, p. 2640-2644, 2017). **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 116, n. 18, p. 9137, 2019.

TAVARES JÚNIOR, I. da S. et al. Machine learning: Modeling increment in diameter of individual trees on Atlantic Forest fragments. **Ecological Indicators**, Amsterdam, v. 117, p. 106685, 2020.

TÉO, S. J.; FILHO, A. F.; LINGNAU, C. Análise espacial do estresse competitivo, incremento diamétrico e estrutura de uma floresta ombrófila mista, Irati, PR. **Floresta**, Curitiba, v. 45, n. 4, p. 681–694, 2015.

TERRA, M. DE C. N. S. et al. Stemflow in a neotropical forest remnant: vegetative determinants, spatial distribution and correlation with soil moisture. **Trees**, New York, v. 32, p. 323–335, 2018.

THOMAS, C. et al. Pinus taeda. **Ciência Florestal**, Santa Maria, v. 16, n. 3, p. 319–327, 2006.

TRIGUERO-OCAÑA, R. et al. Dynamic network of interactions in the wildlife-livestock interface in mediterranean Spain: An epidemiological point of view. **Pathogens**, Basel, v. 9, n. 120, p. 1–16, 2020.

TSIOTAS, D.; CHARAKOPOULOS, A. Visibility in the topology of complex networks. **Physica A - Statistical Mechanics and its Applications**, Amsterdam, v. 505, p. 280–292, 2018.

VANCLAY, J. K. et al. Spatially explicit competition in a mixed planting of *Araucaria cunninghamii* and *Flindersia brayleyana*. **Annals of Forest Science**, Les Ulis, v. 70, p. 611–

619, 2013.

VIEIRA, G. C. et al. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment**, Amsterdam, v. 619–620, p. 1473–1481, 2018.

WANG, S.; ZHENG, L.; YU, D. The improved degree of urban road traffic network: A case study of Xiamen, China. **Physica A - Statistical Mechanics and its Applications**, Amsterdam, v. 469, p. 256–264, 2017.

ZEMP, D. C. et al. Self-amplified Amazon forest loss due to vegetation-atmosphere feedbacks. **Nature Communications**, London, v. 8, p. 1–10, 2017.


ZHANG, J.; HUANG, S.; HE, F. Half-century evidence from western Canada shows forest dynamics are primarily driven by competition followed by climate. **Proceedings of the National Academy of Sciences**, Washington, v. 112, n. 13, p. 4009–4014, 2015.


ZHANG, R. Global dynamic analysis of a model for vector-borne diseases on bipartite networks. **Physica A - Statistical Mechanics and its Applications**, Amsterdam, v. 545, p. 1–16, 2020.


ZHANG, Z. et al. The effect of tree size, neighborhood competition and environment on tree growth in an old-growth temperate forest. **Journal of Plant Ecology**, Oxford, v. 10, n. 6, p. 970–980, 2016.

APPENDIX A. SUPPLEMENTARY DATA

Appendix A. The representation of the complex network structure with nodes colored according to each species and their sizes scaled according to the values of the topological metrics.

 *C. langsdorffii*

 *X. brasiliensis*

 Others

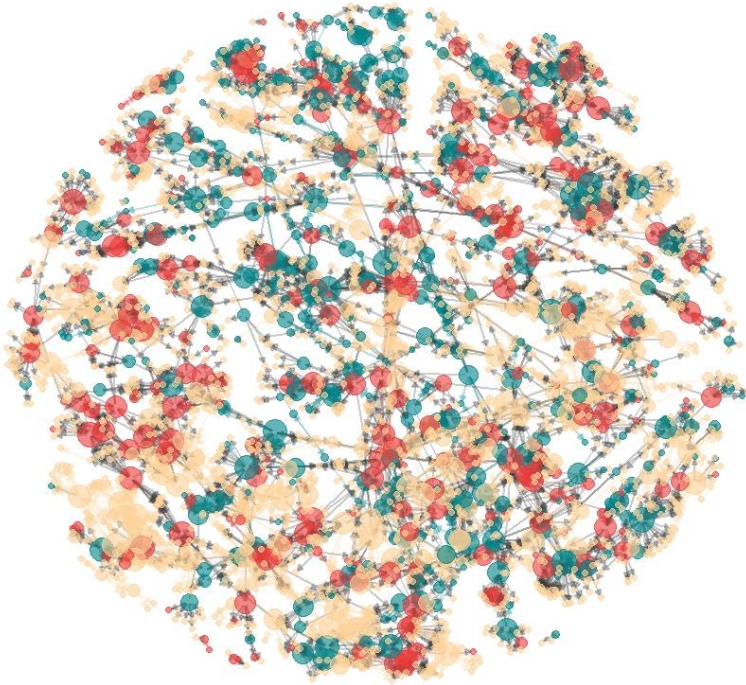
1) In-degree



Values



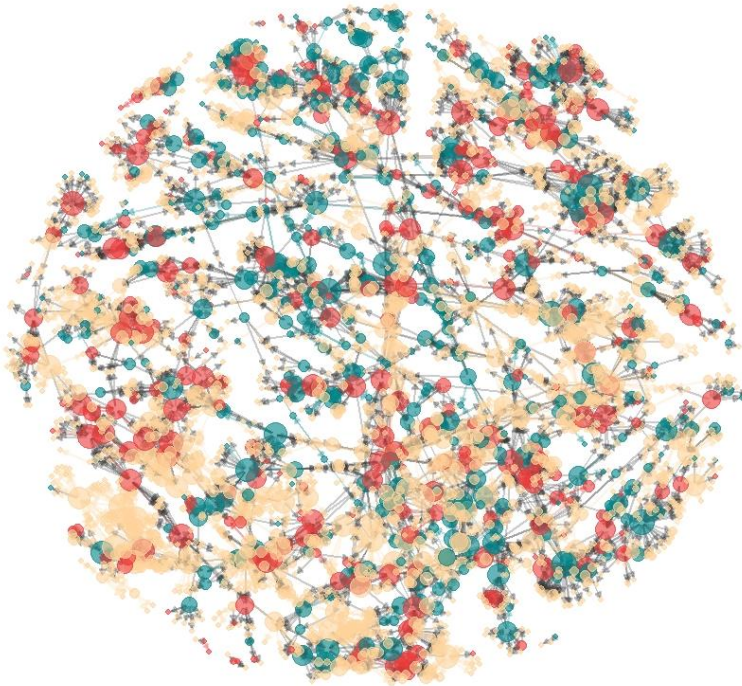
2) Out-degree



Values

- [0, 1]
- (1, 2]
- (2, 3]
- (3, 4]
- (4, 5]
- (5, 6]
- (6, 7]
- (7, 8]
- (8, 9]
- (9, 10]
- (10, 11]
- (11, 13]
- (13, 14]
- (14, 16]
- (16, 18]
- (18, 22]
- (22, 28]
- (28, 55]

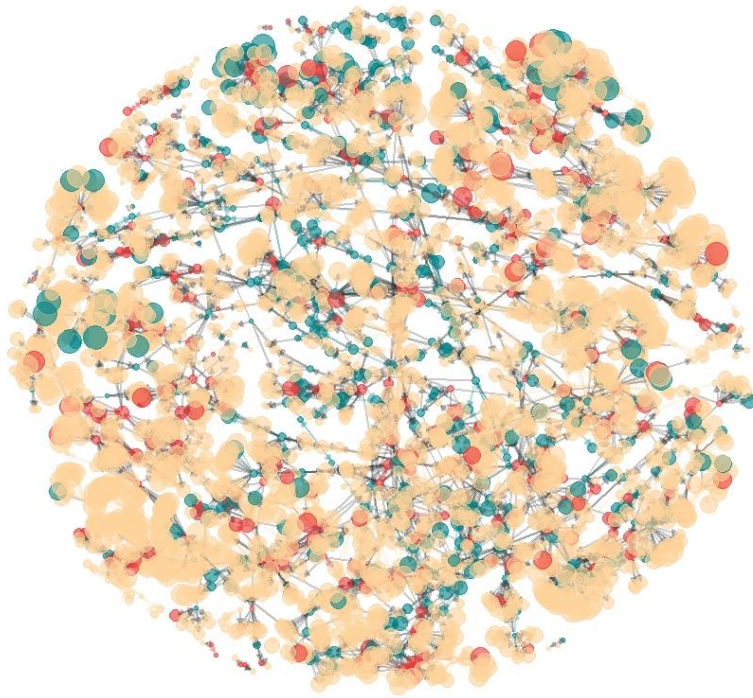
3) Total degree



Values

- [1, 2]
- (2, 3]
- (3, 4]
- (4, 5]
- (5, 6]
- (6, 7]
- (7, 8]
- (8, 9]
- (9, 9, 10]
- (10, 11]
- (11, 12]
- (12, 13]
- (13, 14]
- (14, 15]
- (15, 16]
- (16, 17]
- (17, 18]
- (18, 19]
- (19, 21]
- (21, 23]
- (23, 26]
- (26, 29]
- (29, 35]
- (35, 56]

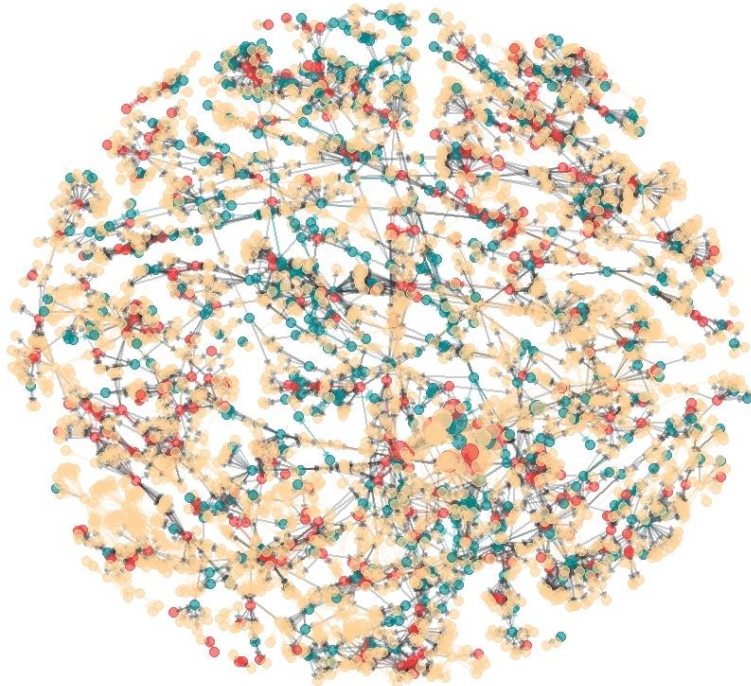
4) Nearest neighbors degree



Values

- [1,2] ● (21,5,22]
- (2,3] ● (22,23]
- (3,4] ● (23,24]
- (4,5] ● (24,25]
- (5,6] ● (25,26]
- (6,7] ● (26,27]
- (7,8] ● (27,28]
- (8,9] ● (28,29]
- (9,10] ● (29,30]
- (10,11] ● (30,32]
- (11,12] ● (32,33]
- (12,13] ● (33,34]
- (13,14] ● (34,35]
- (14,15] ● (35,37]
- (15,16] ● (37,39]
- (16,17] ● (39,42]
- (17,18] ● (42,43]
- (18,19] ● (43,49]
- (19,20] ● (49,56]
- (20,21] ●

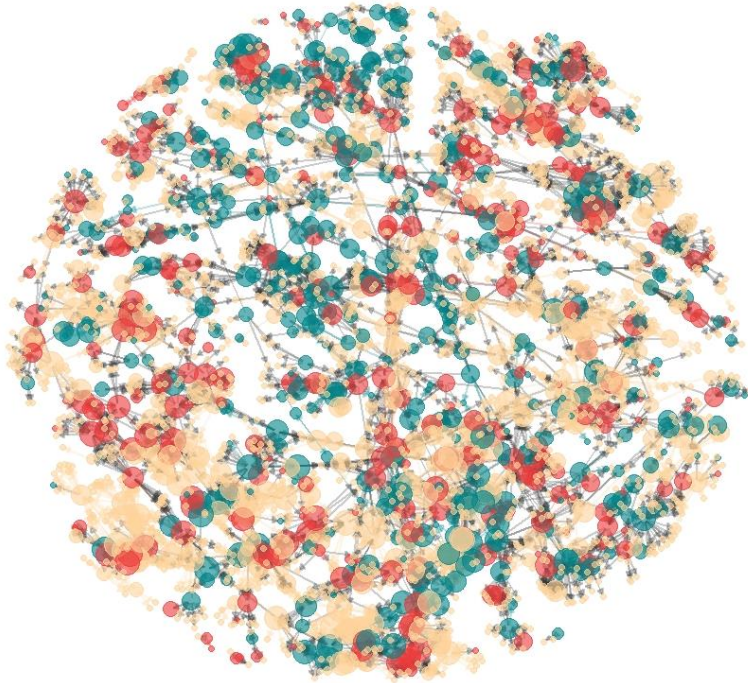
5) Eigenvector centrality



Values

- [0,0.1]
- (0.1,0.2]
- (0.2,0.3]
- (0.3,0.4]
- (0.4,0.691]
- (0.691,1]

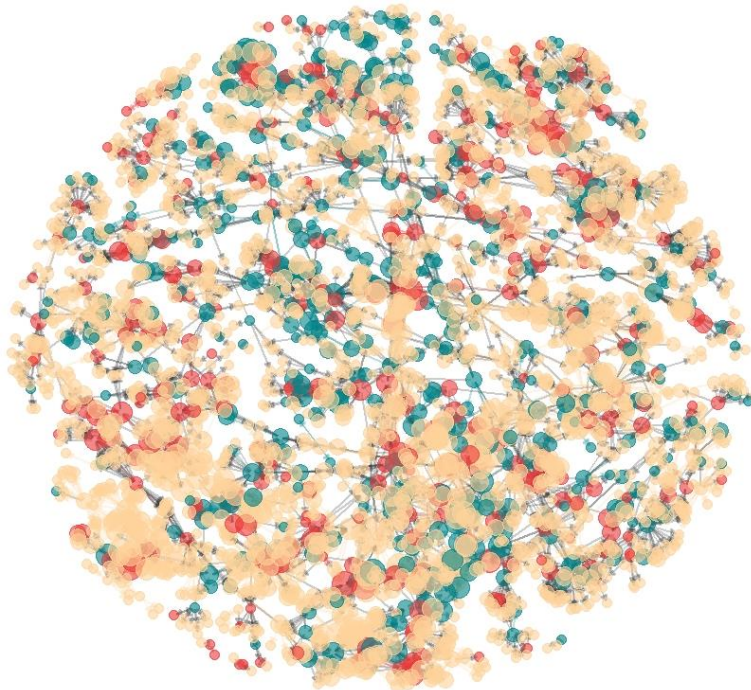
6) Closeness centrality



Values

- [0,3]
- (3,4]
- (4,7]
- (7,12]
- (12,18]
- (18,25]
- (25,36]
- (36,48]
- (48,63.6]
- (63.6,85]
- (85,108]
- (108,129]
- (129,159]
- (159,199]
- (199,252]
- (252,336]
- (336,622]

7) Coreness



Values

- [1,2]
- (2,3]
- (3,4]
- (4,5]
- (5,6]
- (6,7]
- (7,9]

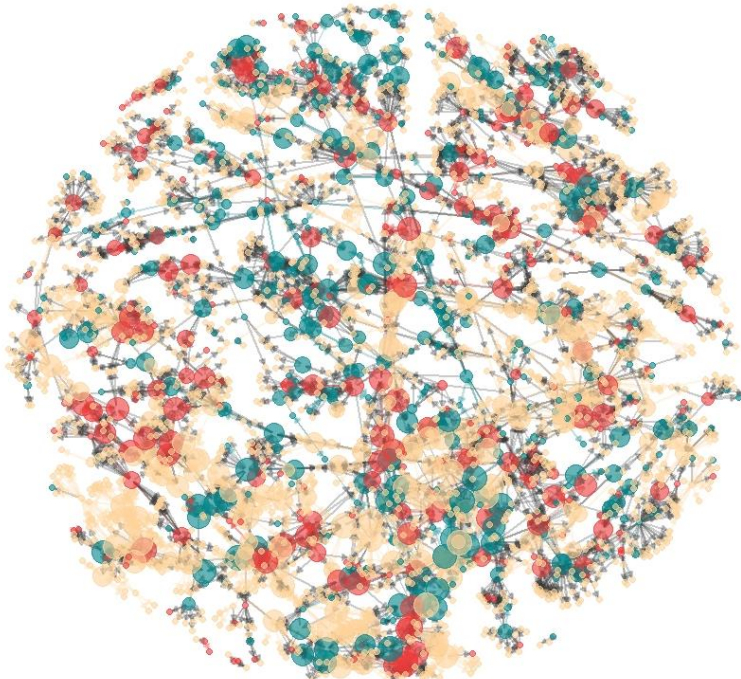
8) Clustering coefficient



Values

- [0,0.2]
- (0.2,0.3]
- (0.3,0.5]
- (0.5,0.7]
- (0.7,1]

9) Betweenness centrality



Values

- [0,1]
- (1,2]
- (2,4]
- (4,6]
- (6,9]
- (9,12]
- (12,21]
- (21,30.2]
- (30.2,36.3]
- (36.3,48.5]
- (48.5,66]
- (66,89.8]
- (89.8,136]
- (136,190]
- (190,269]
- (269,426]
- (426,736]
- (736,1.73e+03]
- (1.73e+03,4.69e+03]
- (4.69e+03,1.58e+04]

10) PageRank



Values

- [0.0002,0.0003]
- (0.0003,0.0004]
- (0.0004,0.0005]
- (0.0005,0.0006]
- (0.0006,0.0007]
- (0.0007,0.0028]