UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM GEOCIÊNCIAS APLICADAS E GEODINÂMICA

ESTIMATIVA DE BIOMASSA NA REGIÃO AMAZÔNICA UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

TESE DE DOUTORADO Nº 68

CARLOS ALBERTO PIRES DE CASTRO FILHO

Brasília – DF
2021

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM GEOCIÊNCIAS APLICADAS E GEODINÂMICA

ESTIMATIVA DE BIOMASSA NA REGIÃO AMAZÔNICA UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Carlos Alberto Pires de Castro Filho

Orientador: Prof. Dr. Edilson de Souza Bias

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Geociências Aplicadas e Geodinâmica, na Área de Concentração Geoprocessamento e Análise Ambiental, do Instituto de Geociências da Universidade de Brasília (UnB), como requisito para obtenção do Título de Doutor.

Brasília – DF
2021

Universidade de Brasília – UnB
Instituto de Geociências
Programa de Pós-Graduação em Geociências Aplicadas e Geodinâmica
Área de Concentração Geoprocessamento e Análise Ambiental

Estimativa de Biomassa na Região Amazônica Utilizando Técnicas de Aprendizado de Máquina

Carlos Alberto Pires de Castro Filho

Banca Examinadora:

Prof. Dr. Edilson de Souza Bias – IG/UnB (Presidente)

Prof. Dr. Eraldo Aparecido Trondoli Matricardi (Dpto. Eng. Florestal/UnB)

Prof. Dr. Gilson Alexandre Ostwald Pedro da Costa (UERJ)

Dr. Wagner Barreto da Silva (Diretoria de Serv. Geográfico/Exército Brasileiro)

Aprovado pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de Doutor.

Brasília, 29 de novembro de 2021

Castro-Filho, Carlos Alberto Pires de
Estimativa de Biomassa na Região Amazônica Utilizando Técnicas de Aprendizado de Máquina / Carlos Alberto Pires de Castro Filho; Orientador Edilson de Souza Bias. -- Brasília, 2021. 170 p.

Tese (Doutorado - Doutorado em Geociências Aplicadas) – Universidade de Brasília, 2021.

1. Sensoriamento Remoto. 2. Biomassa. 3. SAR. 4. Aprendizado de Máquina.
I. Bias, Edilson de Souza, orient. II. Título

REFERÊNCIA BIBLIOGRÁFICA
Castro-Filho, Carlos Alberto Pires de. Estimativa de Biomassa na Região Amazônica Utilizando Técnicas de Aprendizado de Máquina. Tese de Doutorado. Brasília, Instituto de Geociências, Universidade de Brasília, UnB, 170 p., 2021.

Carlos Alberto Pires de Castro Filho
carlos,pires.1976@gmail.com

“If the future's looking dark
We're the ones who have to shine
If there's no one in control
We're the ones who draw the line
Though we live in trying times
We're the ones who have to try
Though we know that time has wings
We're the ones who have to fly”

(Everyday Glory - Rush)

Agradeço a todos da minha família por terem cedido tantos momentos em prol deste trabalho. Meus pais Carlos e Denise pela educação que me deram; meus avós e tios pela formação do meu caráter; minha irmã e meus primos pela parceria; e minha linda esposa Simone pelo amor. Compartilho minha alegria com minhas queridas Júlia e Lívia e com meu filho Benjamin, "filho da felicidade".

Agradeço à UnB e ao meu orientador, Dr. Edilson de Souza Bias, por me dar a oportunidade de colocar em prática as ideias inovadoras e desafiadoras desta tese, orientando sempre com palavras de incentivo. Nada disso seria possível sem o profissionalismo e o senso de humanidade deste pesquisador.

Agradeço ao Exército Brasileiro, mais especificamente aos integrantes da Diretoria de Serviço Geográfico, pelo incentivo. Serei eternamente grato ao Sr Gen Div Pedro Paulo Levi Mateus Canazio e à Cel Linda Soraya Issmael por todo o apoio e pela amizade.

Agradeço aos pesquisadores e colegas do INPE que também contribuíram para a absorção do conhecimento necessário: Professora Corina, Professor João Roberto, Dr Sidnei Sant'Anna, Dr Rogerio Negri, Dr Eliana Pantaleão, Ms Luciana Pereira ... entre tantos outros.

Obrigado a todos que torceram por mim e que contribuíram, direta ou indiretamente, para este trabalho!

Cada parágrafo da presente tese foi elaborado em feriados, fins de semana e em momentos que seriam destinados ao descanso após um árduo dia de trabalho. Logo, dedico ela a todos os brasileiros que pagam seus impostos e que, apesar de raramente utilizarem de forma direta os serviços públicos, acreditam que as instituições estatais trabalham para o crescimento do nosso Brasil. Eis um singelo retorno!

O presente trabalho foi realizado com o Apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

No ano de 2016 mais de 190 países participaram da 21ª Conferência das Partes das Nações Unidas sobre Mudança Climática, realizada em Paris. Apesar de intensos trabalhos visando elaborar um tratado, os resultados não atenderam às expectativas devido à falta de metodologias que medem com precisão a quantidade de biomassa florestal. Imagens de sensoriamento remoto podem ser usadas para que seja realizada uma quantificação mais precisa e viável da biomassa existente em regiões de difícil acesso, como a região amazônica, com destaque para as imagens na faixa do micro-ondas, mais especificamente as de radares. Em função da grande quantidade de dados de sensoriamento remoto disponíveis, faz-se necessário o desenvolvimento de técnicas e ferramentas que visem organizá-los e analisá-los de forma inteligente e automática, como as técnicas de Aprendizado de Máquina (*Machine Learning*). A presente tese tem por objetivo geral desenvolver e aplicar uma metodologia para estimar a quantidade de biomassa arbórea em uma área da região amazônica, a partir de dados de SAR, utilizando técnicas de Aprendizado de Máquina. As etapas metodológicas de tese encontra-se divididas em três artigos técnicos sequenciais que cobrem os objetivos propostos. O primeiro artigo possui como hipótese a possibilidade de ajuste da altura interferométrica, atributos de InSAR, a partir da identificação de áreas de solo exposto, isto é, onde o valor é teoricamente igual a 0 (zero). Além de inovadora, a hipótese previa o ajuste do modelo digital da região visando aprimorar a modelagem referente à estimativa de biomassa. Entretanto, como resultado, o método proposto no primeiro artigo não possibilitou a melhora significativa da estimativa de biomassa florestal, não sendo adotado nas próximas etapas do trabalho. O segundo artigo dá continuidade ao primeiro e apresenta a aplicação de técnicas de Aprendizado de Máquina sobre os atributos de SAR extraídos dos dados disponíveis. De forma inédita avalia e compara modelos de estimativa de biomassa baseados em atributos qualitativos e quantitativos. O segundo artigo conclui que as diferentes regiões da Floresta Amazônica e suas respectivas características demandam modelos e técnicas específicas, não se enquadrando em um único padrão. Neste caso não foi possível identificar uma única técnica de Aprendizado de Máquina que se mostrasse como a mais adequada ao objetivo, apesar dos melhores resultados apontarem para o uso das redes neurais artificiais. O terceiro e último artigo conclui o trabalho da presente tese por meio da análise e construção de produtos temáticos de biomassa. Neste último artigo é apresentado um sistema computacional desenvolvido que visa otimizar o processo de categorização, necessário à representação visual da geoinformação. Os resultados obtidos no terceiro artigo mostram que o algoritmo de Otimização de Categorização proposto demonstrou capacidade de encontrar novos subintervalos de categorias que aumentaram o índice de concordância Kappa. Como resultado, foram construídos produtos temáticos que apresentaram acurácia temática superior aos obtidos pelos métodos clássicos de categorização. Juntamente, do ponto de vista computacional, a heurística proposta no algoritmo possibilitou a identificação de resultados de forma eficiente, evitando os altos custos de processamento. A hipótese proposta na tese, isto é, de que a aplicação de técnicas de aprendizado de máquina sobre dados de SAR permitem obter a estimativa de biomassa da região amazônica com erros abaixo de 20%, atendendo os padrões preceituados por organismos internacionais, não foi confirmada. Os resultados obtidos nos modelos elaborados são classificados somente como *moderados*. Dentre os fatores que podem ter contribuído para este resultado, está a quantidade reduzida de amostras de biomassa, com pequena variação de valores, o que prejudicou o ajuste dos modelos gerados e o acesso restrito aos dados de SAR das bandas X e P, não sendo possível gerar novos atributos coerentes.

Palavras-chave: Biomassa florestal; Amazônia; Sensoriamento remoto; Radar de abertura sintética; modelo digital; aprendizado de máquina.

ABSTRACT

In 2016, more than 190 countries participated in the 21st United Nations Conference of Parties on Climate Change, held in Paris. Despite the intense work aiming at preparing a treaty, the results did not meet expectations due to the lack of methodologies that accurately measures the amount of forest biomass. Remote sensing images can be used to make a more accurate and viable quantification of the existing biomass in regions with difficult access, such as the Amazon region, with emphasis on images in the microwave range, more specifically those from radar. Due to the large amount of remote sensing data available, it is necessary to develop techniques and tools that aims to organize and analyze them in an intelligent and automatic way, such as Machine Learning techniques. The present thesis has as general objective to develop and apply a methodology to estimate the amount of arboreal biomass in an area of the Amazon region, using SAR data and Machine Learning techniques. The thesis methodological steps are divided into three sequential technical articles that covers the proposed objectives. The first article hypothesizes the possibility of adjusting the interferometric height, InSAR feature, using the exposed soil areas identified in the image, that is, where the value is theoretically equal to 0 (zero). In addition to being innovative, the hypothesis predicted the adjustment of the region digital model in order to improve the biomass estimation modeling. However, as a result, the method proposed in the first article did not present a significant improvement in the estimation of forest biomass and was not adopted in the next stages of the work. The second article gives sequence for the first and presents the application of Machine Learning techniques over SAR features extracted from the available data. In an unprecedented way, it presents a methodology that evaluates and compares biomass estimation models based on qualitative and quantitative features. The second article concludes that the different Amazon Forest regions and their respective characteristics demands specific models and techniques, not fitting into a single pattern. In this case, it was not possible to identify a single Machine Learning technique that proved to be the most adequate for the purpose, despite the best results pointing to the use of artificial neural networks. The third and last article concludes the work of this thesis through the analysis and construction of thematic biomass products. In this last article, a computational system that aims to optimize the categorization process was developed, necessary for the visual representation of geoinformation. The results obtained in the third article shows that the proposed Categorization Optimization algorithm demonstrated the ability to find new subintervals of categories that increased the Kappa agreement index. As a result, thematic products were constructed and presented thematic accuracy superior to those obtained by the classical categorization methods. Besides that, from a computational point of view, the heuristic proposed in the algorithm enabled the identification of results in an efficient way, avoiding high processing costs. The hypothesis proposed in the thesis, that is, that the application of machine learning techniques over SAR data allows to obtain an estimate of biomass in the Amazon region with errors below 20%, attending to the standards established by international organizations, was not confirmed. The results obtained in the constructed models were classified only *moderate*. Among the factors that may have contributed to this result, there is the reduced amount of biomass samples, with little variation in values, which impaired the adjustment of the generated models and the restricted access to the X and P bands SAR data, not being possible to generate new coherent features.

Keywords: Forest biomass; Amazon; Remote sensing; Synthetic aperture radar; digital model; machine learning.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AC	altura comercial
AD	Árvore de Decisão
AGB	*above ground biomass*
AIRSAR	*Airborne* SAR
ALOS	*Advanced Land Observing Satellite*
AT	altura total
BMI	*biomass index* (índice de biomassa)
CF	*calibration factor*
COP	Conferência das Partes das Nações Unidas sobre Mudança Climática
CSI	*canopy structure index* (estrutura do dossel)
DAP	diâmetro na altura do peito
DEM	*Digital Elevation Model* (modelos digitais de elevação)
DLR	*Deutsche Zentrum für Luft- und Raumfahrt* (Centro Aeroespacial Alemão)
DN	*Digital Number*
DSG	Diretoria de Serviço Geográfico
ENVI	*Environment for Visualizing Images*
ESA	*European Space Agency* (Agência Espacial Européia)
ET-PCDG	Especificação Técnica para Produtos de Conjunto de Dados Geoespaciais
FLONA	Floresta Nacional
GIS	*Geographic Information System*
GLCM	*Gray Level Co-occurrence Matrix* (matriz de co-ocorrência de níveis de cinza)
GRG	Gradiente Reduzido Generalizado
H_{int}	altura interferométrica
IBAMA	Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis

IBGE	Instituto Brasileiro de Geografia e Estatística
IDL	*Interactive Data Language*
INPA	Instituto Nacional de Pesquisas da Amazônia
INPE	Instituto Nacional de Pesquisas Espaciais
IPCC	*Intergovernmental Painel for Climate Change*
ITI	*interaction type index* (índice de tipo de interação)
JAROS	*Japan Resources Observation System Organization*
JAXA	*Japan Aerospace Exploration Agency*
JPL	*Jet Propulsion Laboratory*
INPA	Instituto Nacional de Pesquisas da Amazônia
InSAR	*Interferometric Synthetic Aperture Radar (Radar de Abertura Sintética Interferométrico)*
LiDAR	*Light Detection And Ranging*
LR	*logistic statistical regression*
MCT	Ministério da Ciência e Tecnologia
MDE	Modelo Digital de Elevação
MDS	Modelos Digitais da Superfície
MDT	Modelos Digitais do Terreno
ML	*machine learning* (Aprendizado de Máquina)
MLP	*Multilayer perceptron* (perceptron multicamadas)
MR	*Multiple statistical regression*
NASA	*National Aeronautics and Space Administration*
OBCT	*Ordinary Binary Classification Tree* (árvore de decisão univariada)
PALSAR	*Phased Array type L-band Synthetic Aperture Radar*
PolSAR	*Polarimetric Synthetic Aperture Radar (Radar de Abertura Sintética Polarimético)*
PIB	Produto Interno Bruto
PolSARPro	*Polarimetric SAR data Processing and Educational Tool*

PRODES	Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite
P_d	*double bounce*
P_s	espalhamento superficial
P_t	potência total
P_v	espalhamento volumétrico
RAT	*Radar Tools*
REDD	*Reduce Emissions for Deforestation and Degradation*
R_c	razão de polarização cruzada
RF	*Random Forest*
RNA	Redes Neurais Artificiais
ROI	*region of interest*
R_p	razão de polarização paralela
SAR	*Synthetic Aperture Radar* (Radar de Abertura Sintética)
SGC	São Gabriel da Cachoeira
SLC	*Single Look Complex*
SNAP	*Sentinel Application Plataform*
SR	*Simple statistical regression*
SRTM	*Shuttle Radar Topography Mission*
SVM	*Suport Vector Machine*
TEEB	*The Economics of Ecosystems and Biodiversity*
UNFCCC	*United Nations Framework Convention on Climate Change*
UTM	Universal Tranversa de Mercator
VSI	*volume scattering index* (espalhamento volumétrico)
WEKA	*Waikato Environment for Knowledge*

1 INTRODUÇÃO

1.1 Contextualização

No ano de 2016 mais de 190 países participaram da 21ª Conferência das Partes das Nações Unidas sobre Mudança Climática (COP-21), realizada em Paris. Esta conferência visou dar prosseguimento ao Protocolo de Kyoto, que expirou em 2012, e, consequentemente, definir metas a serem buscadas referentes a emissões de gases poluentes na atmosfera. Apesar de intensos trabalhos, um tratado legalmente vinculante, capaz de obrigar a comunidade internacional a cortar emissões de gases responsáveis pelo efeito estufa, não foi produzido. Dentre os motivos deste insucesso, destaca-se a falta de metodologias que medem com precisão estes cortes e estabelecem mecanismos para esta redução.

Segundo a *United Nations Framework Convention on Climate Change* – UNFCCC (2008) – o artigo 3.4 do Protocolo de Kyoto determina que os países devam informar anualmente as mudanças nos estoques de carbono associados à biomassa florestal. O *Intergovernmental Painel for Climate Change* – IPCC (2003), órgão subordinado à Organização Mundial de Meteorologia e Programa das Nações Unidas para o Meio Ambiente, afirma que os relatórios com estas informações devem seguir uma metodologia baseada nos princípios de transparência, consistência, comparabilidade, completude e acurácia.

No entanto, Malhi et al. (1999) e Saatchi e Moghaddam (2000) afirmam que estudos quantificando o ciclo de carbono entre a atmosfera e as florestas ainda são necessários. Os modelos referentes ao fluxo de gases poluentes para a atmosfera levam em consideração as taxas referentes às mudanças de uso do terreno, mas sua principal fonte de erro é a incerteza nos parâmetros de entrada referentes à quantidade de biomassa existente nestas regiões.

No Brasil a degradação de florestas ocorre principalmente em função da transformação destas áreas em áreas de pastagem, em projetos de colonização, em atividades madeireiras e em redefinição de fronteiras agrícolas (NARVAES, 2010). Segundo Beaudoin et al. (1994), a biomassa florestal de superfície é um parâmetro importante do ecossistema terrestre, já que entre 80% e 92% da biomassa terrestre, e consequentemente da quantidade de carbono, encontra-se nos ecossistemas florestais.

De acordo com o Ministério da Ciência e Tecnologia – MCT (2010), o desmatamento de ecosistemas florestais, ocasionado em função de mudanças no uso do solo, representa 61% do total de gases do efeito estufa emitidos pelo país, sendo o principal o CO_2. Segundo o PROJETO PRODES (2019), o desmatamento entre os meses de agosto de 2017 a 2019 cresceu em comparação aos anos anteriores, sendo de 7.900 km^2. Desde o início das medições

realizadas pelo referido projeto, o total dos gases emitidos pelo Brasil, segundo o Programa das Nações Unidas para o Meio Ambiente, representa 52% do emitido na América Latina, tornando o país o 5º maior emissor do mundo.

Diante deste quadro, o governo brasileiro, por meio do Plano Nacional de Mudanças Climáticas, definiu a meta de redução de 80% até 2022, tendo como parâmetro o índice de desmatamento anual de 2005. O motivo desta meta vem em função de que, segundo o *The Economics of Ecosystems and Biodiversity* – TEEB (2008), cerca de 11% do Produto Interno Bruto (PIB) brasileiro depende de recursos naturais fornecidos diretamente pelo meio ambiente, como nutrientes do solo e água. Além disso, concomitantemente à redução do desmatamento, objetiva-se reduzir as emissões de gases causadores do efeito estufa em 39% até 2022, tendo também como parâmetro o ano de 2005 onde foram emitidos na atmosfera gases do efeito estufa equivalentes a cerca de 2,19 bilhões de toneladas de CO_2 (MCT, 2010).

Silva (2007) afirma que para uma região florestal desmatada por queimada pode-se calcular a quantidade de carbono que foi lançado na atmosfera, caso haja uma estimativa de biomassa anteriormente existente naquele local. Afirma, ainda, que na região amazônica 48,5% da biomassa seca é composta de carbono, viabilizando a transformação entre essas medidas.

Gonçalves (2007) e Ghasemi et al. (2011) afirmam que a estimativa de biomassa visa subsidiar os planos de preservação de ecossistemas, além de propiciar informação para o manejo sustentável florestal. Permite, ainda, que durante o monitoramento se possa verificar o tipo, a direção, a intensidade e a extensão da degradação de diversas áreas, causadas por influência humana ou por incêndios florestais de causas naturais (SAATCHI et al., 2007a).

Santos et al. (2003) observam que imagens de sensoriamento remoto são necessárias para que seja feita uma quantificação mais precisa e viável da biomassa existente em uma determinada região. Mais especificamente na Amazônia brasileira, informações de estimativa de biomassa em grandes escalas, derivadas de sensoriamento remoto, proporcionariam simulações valiosas referentes aos fluxos de carbono (BELLASSEN et al., 2011). Nestas regiões, apesar de já existirem mais de 300 planos de manejo florestal, sua grande maioria é somente qualitativo (GONÇALVES, 2007).

Entretanto, Ghasemi (2011) afirma que imagens ópticas de sensoriamento remoto possuem capacidade limitada para estimar biomassa florestal. Nestes casos, a interação da radiação solar ocorre principalmente sobre o dossel das árvores, perdendo sensibilidade com o tronco e os galhos. Imhoff (1995) e Kasischke et al. (1997) complementam afirmando que os sensores remotos mais adequados na quantificação de biomassa em regiões de florestas

tropicais são os que trabalham na faixa do micro-ondas, mas especificamente os radares. Estes sensores têm diversas vantagens sobre os ópticos, dentre os quais se destacam: (a) a capacidade de seus sinais penetrarem em nuvens, comuns em regiões tropicais; (b) o controle na potência de emissão dos sinais, sendo independente das variações da radiação solar; e (c) a possibilidade de configuração do comprimento de onda e respectiva penetrabilidade no dossel florestal, obtendo assim informação a respeito da estrutura florestal.

Diversos autores destacam a potencialidade no uso de radares de abertura sintética (SAR) polarimétricos (PolSAR) de baixa frequência para a estimativa de biomassa (BEAUDOIN et al., 1994; POPE et al., 1994; IMHOFF, 1995; KASISCHKE et al., 1997; SAATCHI e MOGHADDAM, 2000; ASKNE et al., 2003; SANTOS et al., 2003; e SAATCHI et al., 2007a). Enquanto a estimativa de biomassa com o uso de bandas de SAR de alta frequência, como C e X, costumam saturar a níveis próximos de 20 e 40 toneladas por hectare (t/ha) respectivamente, o uso de bandas L e P são capazes de registrar valores acima de 100 t/ha.

Autores utilizam, ainda, a potencialidade interferométrica dos dados de SAR (NEEFF et al., 2005; GAMA, 2007; TREUHAFT et al., 2009; WILLIAMS et al., 2009; NI et al., 2010). Nestes casos os níveis de saturação dos atributos extraídos dos dados dos SAR interferométricos (InSAR) podem chegar até a 300 t/ha e o coeficiente de determinação que relacionam estes dados à biomassa atinge valores de cerca de $r^2 = 0,86$.

Em função dos bons resultados obtidos por pesquisadores, novos projetos que visam utilizar dados de SAR para estimar biomassa encontram-se em execução ou em etapa de planejamento. Como exemplos, podem ser citados os projetos ALOS PALSAR, da Agência Espacial Japonesa (Japan Aerospace Exploration Agency – JAXA), e o BIOMASS, da Agência Espacial Européia (ESA).

O sensor Radar de Abertura Sintética em Fase na Banda L (Phased Array type L-band Synthetic Aperture Radar – PALSAR), a bordo do Satélite Avançado de Observação da Terra (Advanced Land Observing Satellite – ALOS), foi desenvolvido em um projeto conjunto entre a JAXA e a Organização do Sistema de Observação de Recursos Japones (Japan Resources Observation System Organization – JAROS) (JAXA, 2021). Possui frequência de 1,2 GHz (comprimento de onda de 27 cm) resolução espacial máxima de 10 metros e encontra-se operacional desde 2006, possuindo diversos objetivos como o de monitoramento de (a) desastres naturais ou antrópicos; (b) uso e cobertura do solo; (c) plantações agrícolas; (d) cobertura florestal (JAXA, 2021). Devido ao sucesso da missão, em 2014 foi lançado o ALOS 2 cujo sensor PALSAR possui resolução espacial máxima de 1 metro.

Já, o projeto BIOMASS possui o objetivo específico de mapear toda a biomassa aérea mundial com resolução espacial de 200 metros, compatível com as necessidades de inventários de escalas nacionais e com os cálculos de fluxo de carbono (SCIPAL, 2010). Para que o objetivo seja atingido pretende-se lançar em órbita, no ano de 2022, um veículo orbital equipado com um sensor remoto de SAR operando na banda P que imageará toda a Terra no mínimo duas vezes ao ano (ESA, 2021). O projeto ainda se encontra em fase de análise e atualmente um dos principais desafios científicos é o desenvolvimento de um algoritmo de faça a associação entre os dados de SAR que serão gerados e a biomassa global, tanto para florestas boreais com para tropicais (DUBOIS-FERNANDEZ et al., 2010).

No Brasil, entre os projetos que visam gerar imagens de SAR e que poderão ser utilizadas na estimativa de biomassa destaca-se o de Cartografia da Amazônia, mais especificamente o Subprojeto Cartografia Terrestre, também conhecido como Radiografia da Amazônia. Nele, até 2022, será recoberta uma área total de 1.700.000 km^2 da região amazônica com sensores aerotransportados nas bandas X e P, visando a futura construção de cartas na escala 1:50.000 pela Diretoria de Serviço Geográfico (DSG) do Exército Brasileiro (DSG, 2008). Além de realizar este mapeamento, o projeto visa também gerar dados necessários ao suporte de projetos de infraestrutura e exploração sustentável de recursos naturais da região.

Em função da grande quantidade de dados que podem ser originados dos sensores SAR disponíveis, faz-se necessária a aplicação de técnicas que visam organizar e analisar recursos de forma inteligente e automatizada (Del Frate e Solimini, 2004; Englhart et al., 2012; Camargo et al., 2019; Wylie et al., 2019; Yuan et al., 2020; Martinez-Alvarez e Bui, 2020). As técnicas de aprendizado de máquina – ML (oriunda do inglês *machine learning*) são capazes de modelar conhecimento e fazer associações entre diferentes tipos de informações, podendo ser quantitativas ou qualitativas (Quinlan, 1993; NG, 2018). De acordo com Brink et al. (2016) e Faceli et al. (2021) as principais vantagens do ML são a precisão, uma vez que o algoritmo ótimo é selecionado a partir das características dos dados e do problema a ser resolvido; automação na aprendizagem, que ajusta os modelos de acordo com o sucesso ou fracasso dos resultados; velocidade de processamento; personalização, sendo adequado em qualquer tipo de problema; e escalabilidade, por serem processos que se adaptam ao crescimento dos dados.

1.2 Problema

A Floresta Amazônica representa um dos biomas mais importantes de todo o nosso planeta, constituindo um grande celeiro de produção de biomassa. A par da existência de diversos estudos abordando a quantificação de biomassa, inclusive nesta região, são necessárias metodologias que permitam medir esta biomassa de forma consistente e com nível de acurácia compatível com as demandas ecológicas e ambientais.

1.3 Hipótese

A aplicação de técnicas de aprendizado de máquina sobre dados de SAR permitirá obter a estimativa de biomassa da região amazônica com erros abaixo de 20%, atendendo os padrões preceituados por organismos internacionais.

1.4 Objetivo

1.4.1 Objetivo Geral

Desenvolver e aplicar uma metodologia para estimar a quantidade de biomassa arbórea, a partir de dados de SAR, utilizando técnicas de Aprendizado de Máquina.

1.4.2 Objetivos Específicos

Visando o objetivo geral, podem ser relacionados os seguintes objetivos específicos:

- Analisar técnicas não-paramétricas de aprendizado de máquinas aplicando-as na construção de modelos preditivos para estimativa de biomassa arbórea a partir de variáveis extraídas de dados de SAR;
- Implementar, analisar e propor modelos preditivos para estimativa de biomassa arbórea utilizando métodos estatísticos;
- Avaliar os modelos preditivos construídos visando apontar o mais adequado aos dados e região em trabalho;
- Analisar técnica de ajuste da altura interferométrica para o desenvolvimento de modelo de estimativa de biomassa; e

- Implementar, analisar e propor metodologia para construção de produto temático que englobe as representações cartográficas referentes às categorias de biomassa e suas características.

1.5 Justificativa

Para uma região florestal desmatada por queimada pode-se calcular a quantidade de carbono que foi lançada na atmosfera por meio da estimativa de biomassa anteriormente existente naquele local. Os modelos referentes ao fluxo de gases poluentes para a atmosfera levam em consideração as taxas referentes às mudanças de uso do terreno, mas sua principal fonte de erro é a incerteza nos parâmetros de entrada referentes à quantidade de biomassa existente nestas regiões. Logo, o desenvolvimento de um modelo inovador de estimativa de biomassa que vise minimizar as incertezas é fundamental para que análises do fluxo climático sejam realizadas.

A estimativa de biomassa também visa subsidiar: planos de preservação de ecossistemas; informações para o manejo sustentável florestal; análise sobre o tipo, a direção, a intensidade e a extensão da degradação de diversas áreas, causadas por influência humana ou por incêndios florestais de causas naturais; informações relativas ao processo de sucessão secundária em áreas que foram abandonadas após um período de exploração agropecuária ou madeireira.

Outra questão que justifica a pesquisa referente a estimativa de biomassa são os mecanismos que buscam a redução de gases poluentes através da conservação de florestas, chamados de programas *Reduce Emissions for Deforestation and Degradation* – REDD. Por meio destes programas, serão criados valores econômicos que serão pagos por países poluidores aos que evitarem o desmatamento, isto é, que mantiverem a "floresta de pé". Logo, tal programa possibilita a sustentabilidade de regiões pouco desenvolvidas, como a amazônica, sem que haja degradação ambiental.

Múltiplas reuniões e debates foram realizados internacionalmente visando adotar um método científico para implantação do projeto REDD, não chegando a um consenso. Análises sobre a implantação do REDD concluem que cada país poderá definir um método específico de mensuração e monitoramento de suas florestas, em função de suas capacidades técnicas e condições financeiras, desde que seja garantida a transparência científica, eficiência ambiental e apoio político.

Ainda existe a necessidade de definição de um algoritmo que estime a quantidade de biomassa aérea para grandes regiões. Neste sentido, as técnicas de aprendizado de máquina buscaram construir modelos adequados na estimativa de biomassa para regiões da Amazônia. A aplicação destas técnicas não se restringiu a uma única abordagem, mas, de forma inovadora, é feita uma comparação entre algumas disponíveis, como as de redes neurais artificiais e árvores de decisão.

Poucos autores buscaram utilizar e comparar diferentes técnicas para estimar biomassa. A grande maioria dos trabalhos publicados utilizam um único método de mineração aplicada a uma pequena quantidade de atributos gerados por sensores ópticos. Juntamente a isto, a maioria destes trabalhos tem como foco a classificação de uso do solo em florestas boreais, cuja complexidade difere da floresta amazônica que é a região de estudo do presente trabalho.

Finalmente, a presente tese não visa apenas apresentar um modelo de estimativa de biomassa, mas também propor uma metodologia para representar o produto temático em questão, constituindo uma ferramenta analítica, que permitirá ações de gestão na tomada de decisão. A metodologia visa minimizar a perda de acurácia que ocorre devido à etapa de categorização dos dados numéricos, fundamental para a construção de um produto temático.

Por todos os aspectos apresentados acima, a presente pesquisa propõe uma nova abordagem que visa agregar conhecimento a cada uma das etapas de construção de um produto temático de estimativa de biomassa florestal: extração, avaliação e seleção de atributos de SAR, incluindo o ajuste da altura interferométrica; desenvolvimento e análise de modelos de estimativa de biomassa por meio de técnicas paramétricas e não paramétricas; e apresentação de proposta de categorização para produtos temáticos de biomassa.

2 REVISÃO BIBLIOGRÁFICA

2.1 Estimativa de Biomassa

Biomassa florestal ou fitomassa é a quantidade, em unidade de massa, do material lenhoso contido em uma unidade de área de floresta (ARAÚJO et al., 1999). Segundo o IPCC (2003), a biomassa viva é estimada em diferentes componentes: a biomassa acima do solo (ou aérea), referente aos troncos, galhos grossos e finos, casca, frutos, flores e folhas; e a biomassa abaixo do solo, onde são consideradas somente raízes com diâmetro de base acima de 2mm. Além disso a biomassa pode ser também classificada como fresca (ou simplesmente biomassa, termo que será adotado neste trabalho), quando a respectiva unidade arbórea está de pé (e viva) ou recém derrubada, ou seca, após sofrer um processo de secagem onde é extraído todo e qualquer líquido dos componentes da árvore (SILVA, 2007).

Segundo Araújo et al. (1999) as estimativas de biomassa podem ser feitas utilizando métodos diretos ou indiretos. Nos métodos diretos, também conhecidos como métodos destrutivos, a biomassa é medida através de um processo que envolve o corte e a pesagem de todos os indivíduos arbóreos ou arbustivos de uma determinada área, neste caso denominada de parcela. Já, o método indireto é realizado através de inventários florestais com medições *in loco* das características físicas de cada indivíduo arbóreo, como a espécie, a altura do diâmetro na altura do peito (DAP), a área basal, a altura comercial (AC), que é a medição até o aparecimento dos primeiros galhos da árvore, e a altura total (AT). A partir destas medições são então utilizadas equações alométricas para estimar a biomassa.

Nos estudos que envolvem inventários florestais, equações alométricas são equações de regressão que relacionam características de parte de um organismo com o seu todo (SILVA, 2007). Para efeito deste trabalho, a alometria é o estudo da biomassa, representando o todo, em função de características do indivíduo como o DAP e a AC ou AT.

Silva (2007) afirma ainda que o DAP, comumente medido a 1,3m de altura do solo, é uma variável que em diversos estudos (UHL et al., 1988; BROWN et al., 1989; WILLIAMS et al., 2005) apresentou correlação positiva e de grande significância para os modelos de estimativa de biomassa. Por outro lado a AC ou AT são variáveis que caracterizam o sítio em que a floresta se desenvolve, sendo fundamentais para estimar a biomassa em regiões afastadas dos dados de origem. A diferença entre ambas é que no método indireto de medição a AC é mais fácil de ser estimada e, portanto, mais precisa. Já, para os métodos diretos, a AT se torna mais adequada por representar a totalidade do eixo vertical do indivíduo arbóreo.

Dentre as equações alométricas mais utilizadas, as de Brown et al. (1989) e Chambers et al. (2001) são vastamente aplicadas para estimar a quantidade de biomassa aérea em florestas tropicais primárias (respectivamente Equação 2.1 e 2.2).

$$Biomassa\ Seca = 0,044\left(DAP^2\,AT\right)^{0,9719} \tag{2.1}$$

$$\ln\left(Biomassa\ Seca\right) = -0,370 + 0,333\ln\left(DAP\right) + 0,933\left[\ln\left(DAP\right)\right]^2 - 0,122\left[\ln\left(DAP\right]^3\right. \tag{2.2}$$

Já, para as florestas tropicais secundárias, a biomassa aérea é comumente estimada pelas equações de Uhl et al. (1988) e Nelson et al. (1999), respectivamente Equação 2.3 e 2.4. Nas Equações 2.1, 2.2, 2.3 e 2.4 as unidades para a Biomassa Seca, o diâmetro na altura do peito (DAP) e a altura total (AT) são, respectivamente, quilogramas (kg), centímetros (cm) e metros (m).

$$\ln\left(Biomassa\ Seca\right) = -2,17 + 1,02\left[\ln\left(DAP\right)\right]^2 + 0,39\ln\left(AT\right) \tag{2.3}$$

$$\ln\left(Biomassa\ Seca\right) = -1,9968 + 2,4128\ln\left(DAP\right) \tag{2.4}$$

As equações desenvolvidas por Uhl et al. (1988), Brown et al. (1989), Nelson et al. (1999) e Chambers et al. (2001) são equações genéricas para regiões tropicais que, segundo Araújo et al. (1999) podem variar em função: (a) do conjunto de dados referentes às diferentes espécimes, (b) dos diferentes sítios e tipos de florestas e (c) dos diferentes componentes de biomassa levados em consideração para o cálculo da biomassa total.

No entanto, Williams et al. (2005) compararam o cálculo de biomassa utilizando equações alométricas específicas, para 14 espécies arbóreas em 11 diferentes sítios na Austrália, com uma única equação alométrica genérica desenvolvida a partir destes dados. A conclusão foi de que a diferença obtida nas avaliações foi insignificantemente pequena, além de que a equação genérica, em função do seu poder de generalização, poderia ser aplicada em regiões florestais daquele país sem que um inventário florístico detalhado fosse necessário. Williams et al. (2005) afirmam ainda que a única ressalva reside no fato de as equações alométricas genéricas serem aplicadas a um único tipo florestal, conforme realizaram em seu estudo.

Conclusões semelhantes foram feitas por Lima et al. (2005). Neste caso Lima et al. (2005) aplicaram a equação alométrica desenvolvida por Higuchi et al. (1998) em Manaus-AM sobre dados inventariados em duas reservas extrativistas da Amazônia ocidental, a cerca de 700 e 1000 quilômetros de distância. Os resultados obtidos indicam que com a equação utilizada, mesmo sendo em função somente do DAP, os resultados obtidos são qualitativamente válidos. Lima et al. (2005) destacam, ainda, sobre a possibilidade da existência de uma equação genérica para um tipo específico de floresta, independente das diferenças de espécies arbóreas.

Aplicando este conceito de uma equação genérica para um tipo florestal específico, Silva (2007) desenvolveu novas equações alométricas baseadas em um inventário florestal realizado em Manaus-AM. Neste caso, o peso da biomassa fresca total (biomassa aérea somada às raízes grossas) pode ser estimado utilizando o modelo alométrico para florestas primárias conforme a Equação 2.5.

$$Biomassa\ Total = 2,7179\, DAP^{1,8774} \tag{2.5}$$

Caso o sítio não seja em Manaus-AM, Silva (2007) orienta no uso da Equação 2.6 para florestas primárias na região amazônica. Nas equações 2.5 e 2.6 a biomassa total, o diâmetro na altura do peito (DAP) e a altura total (AT) encontram-se com as unidades em kg, cm e m, respectivamente.

$$Biomassa\ Total = 0,5521\, DAP^{1,6629}\, AT^{0,7224} \tag{2.6}$$

Juntamente a estas equações alométricas específicas para a região amazônica, Silva (2007) conclui que para esta região o teor médio de água existente na biomassa fresca é de 41,6% e que o teor de carbono na biomassa seca (referente à 58,4% da biomassa fresca) é de 48,5%. Por meio destas porcentagens é possível obter a quantidade de carbono existente na biomassa fresca da região amazônica.

Recentemente Chave et al. (2014) desenvolveram de uma equação alométrica genérica para o cálculo de biomassa aérea (AGB, do inglês *Above Ground Biomass*) visando atender regiões de florestas tropicais de diversas áreas do globo terrestre, isto é, um modelo pantropical. O trabalho levanta questões tais como: (i) qual é o melhor modelo pantropical que considera as variáveis referentes à densidade da madeira (ρ), o DAP e a AT; (ii) como o modelo pantropical se compara aos modelos desenvolvidos para regiões específicas, locais; e

(iii) se apenas o DAP e a densidade da madeira são variáveis suficientes no desenvolvimento de um modelo. Os autores concluem que a equação alométrica pantropical desenvolvida atende à precisão demandada, havendo pequena discrepância com comparação aos modelos locais. Ademais, para modelo pantropical desenvolvido, o tipo de vegetação explicou 0,6% da variância residual relativa, a localidade explicou 21,4% e a variação de espécimes arbóreas dentro da localidade explicou 78%.

A equação 2.7 apresenta o modelo desenvolvido por Chave et al. (2014) onde a AGB, a densidade da madeira (ρ), o diâmetro na altura do peito (DAP) e a altura total (AT) encontram-se com as unidades em kg, g cm^{-3}, cm e m, respectivamente.

$$AGB = 0,0673\,(\rho\, DAP^2\, AT)^{0,976} \tag{2.7}$$

Além dos métodos diretos e indiretos de estimativa de biomassa, outros métodos são realizados com o uso de sensores remotos, não havendo a necessidade de extensos trabalhos de campo, o que viabiliza financeiramente o levantamento de grandes áreas. Nestes casos busca-se relacionar as características representadas pelas imagens às características físicas medidas das árvores ou, diretamente, à biomassa existente na região imageada.

Diversos autores vêm buscando utilizar dados de SAR para modelar e, consequentemente, estimar a quantidade de biomassa em diversos tipos de florestas. Beaudoin et al. (1994), Dobson et al. (1995), Pope et al. (1994), Santos et al. (2003), Saatchi et al. (2007a), Collins et al. (2009) e Narvaes (2010) utilizaram dados de PolSAR em seus trabalhos. Os radares com estas características são capazes de gerar imagens complexas, com informação de fase e de amplitude, e nas polarizações conforme características técnicas do sensor.

Beaudoin et al. (1994) buscaram associar parâmetros biofísicos medidos em inventários de florestas boreais com respostas de retroespalhamento obtidas por sistemas de SAR nas bandas P, L e C do sensor *Airborne* SAR (AIRSAR) da *Jet Propulsion Laboratory* (JPL) da NASA. Em seu trabalho, além de afirmarem que a quantidade de biomassa de um indivíduo arbóreo está linearmente relacionada com sua idade, concluíram ainda, através de regressões lineares, que a banda polarimétrica mais adequadas para a estimativa de biomassa é a P na polarização HH. Esta banda se mostrou a mais adequada para cálculos de biomassa independentes do tipo arbóreo, não sendo afetada pela presença de água e pela orientação angular dos troncos e galhos dos indivíduos no terreno.

Resultados semelhantes foram obtidos por Santos et al. (2002) que analisaram a relação entre os dados de PolSAR da banda P, do sensor aerotransportado desenvolvido pela empresa alemã *Aerosensing RadarSysteme GmbH*, com a área basal média de árvores em transectos da Floresta Nacional (FLONA) do Tapajós. Os autores concluíram que a maior correlação foi obtida com o uso da banda P na polarização HH, porém, neste caso, utilizando um modelo de regressão exponencial.

Estudos realizados por Dobson et al. (1995) utilizaram dados polarimétricos nas bandas L e C do sensor SIR-C para estimar a biomassa em uma floresta boreal americana Neste caso foi utilizada uma metodologia que se iniciava com a classificação do uso do solo e, posteriormente, buscava-se estimar parâmetros biofísicos separadamente para cada parte de uma árvore (troncos, copas, galhos e folhagem). Os valores obtidos por estes parâmetros biofísicos foram então inseridos em equações alométricas específicas para cada parte da árvore e somadas para que se obtivesse o valor total da biomassa para cada tipo de classe de uso do solo. Por meio de avaliações, concluíram que a banda L na polarização HV apresentou os melhores resultados, reduzindo os efeitos de saturação observados em outros estudos. Conclusões semelhantes foram realizadas por Collins et al. (2009) que utilizaram dados do PolSAR nas bandas L e P para gerar modelos para estimativa de biomassa nas florestas de savanas na Austrália.

Outros estudos também mostraram resultados semelhantes. Os comprimentos de onda mais longos (banda L e P) e as polarizações cruzadas (HV e VH) têm maior sensibilidade à estimativa da biomassa aérea (Luckman et al. 1997, Kurvonen et al. 1999, Sun et al. 2002). Por outro lado as polarizações HH e VV se mostram mais sensíveis às mudanças nas condições da superfície (Le Toan et al. 1992, Dobson et al. 1995).

Segundo van der Sanden (1997), a grande vantagem dos dados de PolSAR é que a informação complexa polarimétrica possibilita gerar novos parâmetros que identificam os mecanismos de espalhamento ocorrentes na cena imageada, caracterizando os alvos. Como exemplo, Pope et al. (1994) geraram a partir dos dados de SAR polarimétricos nas bandas P, L e C do sensor AIRSAR os seguintes índices biofísicos: índice de biomassa, de estrutura do dossel, de espalhamento volumétrico e o índice de tipo de interação. Estes índices têm como vantagens o fato de: (a) serem baseados em razões ou normalizações de bandas, tornando-os independentes das variações de declividade no terreno; (b) são associados a mecanismos de retroespalhamento e; (c) são lineares, facilitando as operações estatísticas. Maiores detalhes sobre cada um destes índices serão apresentados no subitem referente aos atributos oriundos de dados de SAR polarimétricos e interferométricos.

Posteriormente, Saatchi e Moghaddam (2000) analisaram o uso de imagens de PolSAR nas bandas C, L e P do sensor AIRSAR para estimar a biomassa em florestas boreais utilizando um algoritmo que estratifica cada indivíduo arbóreo em dossel e tronco e que analisa os atributos biométricos e estruturais de cada um desses extratos. A acurácia obtida na estimativa de biomassa foi de 91% utilizando todas as polarizações nas bandas L e P. Os autores afirmam que apesar da simplicidade devido a pequena quantidade de espécies arbóreas existentes na região imageada, o algoritmo por eles criado independe do nível de umidade da floresta, do tipo de floresta e do tipo de radar utilizado. No entanto, afirmam que é necessária a medição de características dendométricas de amostras *in loco*, o que inviabiliza o método desenvolvido para grandes regiões.

Santos et al. (2003) buscaram avaliar a capacidade de dados de PolSAR nas bandas X e P do sensor *Aerosensing RadarSysteme* para caracterizar florestas primárias e secundárias e analisar suas respectivas relações com a biomassa. Para isto, realizaram um inventário florestal onde as áreas de regeneração foram classificadas em três classes, em função dos estágios de sucessões secundárias. Cada uma das quatro classes (uma de floresta primária e três de secundária) foi avaliada com relação à quantidade de biomassa existente utilizando modelos de regressão logarítmicos e polinomiais. Os autores concluíram que independentemente do modelo de regressão e da classe, os melhores resultados foram obtidos com os dados da banda P nas polarizações HH e HV.

Resultados semelhantes também são observados por Saatchi et al. (2007a) que indicam as imagens de SAR com comprimento de onda da banda P como sendo os mais indicados para trabalhos que visam caracterizar biomassa florestal ou estruturas arbóreas. Segundo os autores, isto acontece porque em bandas com comprimentos de onda menor existe muita saturação do sinal de radar devido a atenuação que ocorre nas folhagens do dossel.

Em trabalho mais recente, Narvaes (2010) sugere que além dos atributos gerados a partir de dados polarimétricos, o potencial da informação interferométrica de um SAR para modelagem de biomassa em ambientes tropicais ainda deverá ser analisada. Gama (2007), ao buscar uma modelagem para estimar a biomassa em um povoamento de *Eucalyptus*, concluiu que os atributos que obtiveram melhores resultados foram provenientes dos produtos interferométricos. A partir da técnica de interferometria por radar é possível a geração de modelos de elevação do terreno, também chamados de modelos digitais de elevação (DEM), de uma determinada região utilizando uma ou mais antenas cuja linha-base é conhecida (MURA, 2000). O módulo do coeficiente de correlação complexo entre as bandas utilizadas

na construção do DEM é outro dado gerado no processo de interferometria, sendo chamado de coerência interferométrica.

Dentre os produtos gerados pela interferometria de radar, a altura interferométrica (H_{int}) é caracterizada por ser a diferença entre os DEMs obtidos por bandas com diferentes comprimentos de onda (NEEFF et al., 2005). Para efeito do presente trabalho os DEMs serão chamados de Modelos Digitais do Terreno (MDT) quando referentes às bandas que geram um modelo do solo, ou chamados de Modelos Digitais da Superfície (MDS) quando referentes às bandas que geram um modelo do dossel (DUTRA et al., 2002). Segundo Dutra et al. (2002), Neeff et al. (2003), Neeff et al. (2005), Gama (2007), Williams et al. (2009) e Ni et al. (2010) o valor da H_{int} obtida entre o MDT e o MDS tem relação com os parâmetros dendrométricos, bem como o de biomassa florestal.

Dutra et al. (2002) em uma campanha na FLONA do Tapajós compararam a H_{int} obtida através da diferença dos DEMs das bandas X e P do sensor *Aerosensing RadarSysteme* com a altura média das árvores medidas em parcelas tanto de florestas primárias como de regeneração. Observaram que o valor da H_{int} encontra-se acima da altura média de todas as árvores, mas abaixo da altura média das árvores que possuíam mais de 20m de altura. Neste sentido concluíram que uma avaliação criteriosa dos DEMs obtidos pela interferometria é necessária para validar o uso da H_{int} em modelos de estimativa de biomassa.

Na mesma região de estudo de Dutra et al. (2002), Neeff et al. (2003) utilizaram a distribuição de Weibull para desenvolver um modelo de estimativa do DAP médio a partir de dados de PolInSAR nas bandas X e P. O melhor resultado obtido com o modelo de regressão linear foi de $r^2 = 0,89$ de coeficiente de determinação utilizando os atributos de H_{int} entre as bandas X e P e de retroespalhamento da banda P de polarização HH.

Neeff et al. (2005) deram continuidade aos trabalhos de Neeff et al. (2003) na região da FLONA do Tapajós utilizando dados PolInSAR das bandas X e P. Neste caso os autores analisaram a correlação entre a altura florestal e a H_{int}. Igualmente, analisaram a correlação entre a biomassa, a área basal, o retroespalhamento obtido pelas bandas polarimétricas e a H_{int}.

Na primeira análise de Neeff et al. (2005), observou-se que a correlação entre a altura florestal e a H_{int} é alta. No entanto, a H_{int} superestima a média das alturas em florestas secundárias, enquanto a mesma subestima a média das alturas em florestas primárias, fato também verificado por Dutra et al. (2002).

Já, nas outras análises, Neeff et al. (2005) observaram que houve somente baixas correlações entre o retroespalhamento da banda P com a área basal (r^2=0,19) e com a biomassa (r^2=0,34). Para o cálculo da biomassa, no entanto, foram utilizadas as equações 2.4

(NELSON et al., 1999) e 2.2 (CHAMBERS et al., 2001), para as parcelas de floresta secundária e primária respectivamente, e os autores atingiram um coeficiente de determinação em torno de $r^2 = 0,84$ ao desenvolverem o modelo da Equação 2.8. Neste modelo, a biomassa é dada em toneladas por hectare (t/ha), a H_{int} está em metros (m) e σ^0_{HH} é o valor do coeficiente retroespalhamento na polarização HH em unidade de decibéis (dB). Maiores detalhes sobre este parâmetro serão apresentados no subitem 2.2 a seguir.

$$Biomassa = 44,965 + 13,87\, H_{\text{int}} + 10,566\, \sigma^0_{HH} \tag{2.8}$$

Gama (2007) afirma que a dificuldade na geração de modelos para estimativa de variáveis biofísicas nos trabalhos realizados por Neeff et al. (2003) e Neeff et al. (2005) reside no fato de existir grande variabilidade de espécies na região da FLONA do Tapajós. Brown et al. (1989) afirmam que a estimativa de biomassa em uma determinada região se torna mais difícil quanto maior for a diversidade de espécies arbóreas. Neste sentido Gama (2007) buscou desenvolver um modelo de estimativa de biomassa em uma região composta por talhões de *Eucalyptus* e com características do terreno controladas utilizando atributos extraídos do mesmo sensor PolInSAR, nas bandas X e P, utilizado por Neeff et al. (2005). O modelo desenvolvido que obteve o melhor resultado ($r^2 = 0,86$ de coeficiente de determinação) foi o que utilizou os atributos de H_{int} e de índice de estrutura do dossel, este último atributo descrito por POPE et al. (1994).

Posteriormente Ni et al. (2010) geraram a H_{int} a partir de bandas SAR de comprimentos de onda diferentes. Neste caso o MDT foi obtido através do modelo ASTER *Global Digital Elevation Model* (ASTER-GDEM) e o MDS foi obtido através do SRTM-DEM derivado da banda InSAR de comprimento de onda C. Comparando esta nova H_{int} com a altura da floresta boreal obtida através de dados de *Light Detection And Ranging* (LiDAR), os autores obtiveram, para um modelo de regressão linear, um coeficiente de determinação de $r^2 = 0,62$. Concluíram que erros advindos dos dados de LiDAR e das diferentes declividades no terreno, dentre outros, podem ter afetado os resultados.

Pesquisas também buscam analisar outros métodos de estimativa de biomassa florestal. Treuhaft et al. (2009) e Hajnsek et al. (2009) compararam atributos interferométricos de SAR com dados de LiDAR. Já, Williams et al. (2009) realizaram a estimativa a partir de bandas segmentadas de atributos.

Treuhaft et al. (2009) buscaram realizar a estimativa da altura de uma floresta tropical utilizando dados de InSAR derivados de múltiplas linhas-base na banda C, de LiDAR e

medições visuais em campo. Além disso, analisaram as fontes de erros em cada método. Chegaram a conclusão de que o erro médio quadrático para a altura das árvores para os três métodos eram bastante próximos, variando em torno de 3m, 3,2m e 3,4m para as medições visuais, por LiDAR e por InSAR.

Resultados semelhantes foram obtidos também por Hajnsek et al. (2009) ao comparar dados de PoInSAR do sistema de SAR aerotransportado experimental (E-SAR) do Centro Aeroespacial Alemão (DLR), nas bandas L e P, e dados de LiDAR para estimar a altura de uma floresta tropical. O coeficiente de determinação, em um modelo de regressão linear, entre ambas as alturas estimadas superou 0,91, validando o uso de sensores de SAR de baixa frequência sobre florestas densas.

Williams et al. (2009) utilizaram dados de InSAR nas bandas X e P inicialmente para extrair o atributo de H_{int}. A partir deste atributo, juntamente com a banda polarimétrica P-HH, foi feita uma segmentação da imagem. Sobre esta segmentação realizou-se uma classificação de uso do solo a qual foi utilizada para identificar a classe de floresta e, sobre ela, estimar a quantidade de biomassa no segmento utilizando a formulação apresentada em Neeff et al. (2005). Os autores analisaram também o uso de um modelo de regressão utilizando apenas a H_{int} e indicando que este atributo, sozinho, já é capaz de estimar a biomassa com precisão adequada para áreas com valores superiores a 150ton/ha. Eles concluem, ainda, que o uso da segmentação permite avaliar a altura média e variância em áreas grandes e homogêneas (acima de 1ha), reduzindo assim os efeitos do ruído *speckle* e, consequentemente, o erro na estimativa de biomassa. Conclusões semelhantes foram feitas por Thiel et al. (2009) que compararam a classificação pixel-a-pixel com a orientada a segmentos sobre dados de SAR para classes de uso do solo.

2.2 Características e Atributos Extraídos de Dados SAR

De acordo com Henderson e Lewis (1998), o termo radar significa “detecção e medição de distância via rádio”, referente à frequência de transmissão dos pulsos eletromagnéticos que são utilizados. O próprio termo também especifica o objetivo do sistema radar que, de forma simplificada, mede as características do sinal de retorno obtido pela reflexão do pulso eletromagnético no alvo. Através destas medições e de suas respectivas análises é possível caracterizar o alvo e medir a distância até o mesmo.

Um dos parâmetros de um sistema radar que influencia na caracterização e precisão na medição de distâncias até os alvos é a frequência do pulso utilizado e o consequente

comprimento de onda. Estes comprimentos de onda, por sua vez, são divididos em faixas de bandas que possuem características conforme a Tabela 2.1.

Tabela 2.1 Faixas de bandas de radar.

Banda	Faixa de Frequência (GHz)	Faixa de Comprimento de Onda (cm)
X	12,5 – 8,0	2,4 – 3,75
L	2,0 – 1,0	15 – 30
P	1,0 – 0,3	30 – 100

Fonte: Adaptado de Henderson e Lewis (1998).

Além dos atributos obtidos diretamente dos dados de SAR polarimétricos e interferométricos, outros também serão gerados nesta tese. Segundo Theodoridis e Koutroumbas (2006), a etapa de extração de atributos visa gerar novos atributos que poderão ser mais adequados na caracterização das classes de tipo de uso do solo em estudo.

Por sua vez, as classes de tipo de uso do solo, durante o processo de interação com as ondas eletromagnéticas de um sistema de SAR polarimétrico, são caracterizadas em função dos tipos de espalhadores (WOODHOUSE, 2006). Espalhadores determinísticos, ou coerentes, são aqueles cujo comportamento da onda incidente e refletida é conhecido. Tais casos são característicos de alvos urbanos. Por outro lado, os espalhadores não determinísticos, ou incoerentes, por possuírem diversos centros de espalhamento, geram a despolarização da onda. Neste caso a captação do retorno da onda incidente pelo sistema de SAR é visto como uma sobreposição de diversas ondas com polarização variável. Estes casos ocorrem principalmente em regiões de floresta.

A partir da característica polarimétrica dos dados de SAR é possível extrair diversos atributos derivados da matriz de espalhamento complexa. Esta matriz é representada na imagem polarimétrica por elementos complexos referentes à amplitude e à fase da onda transmitida e recebida. Segundo Woodhouse (2006) esta matriz é dada pela Equação 2.9 onde S_{pq} representa o sinal retroespalhado, emitido na polarização p e recebido na polarização q, podendo p e q serem horizontais (H) ou verticais (V) independentemente.

$$[S]=\begin{pmatrix} S_{VV} & S_{VH} \\ S_{HV} & S_{HH} \end{pmatrix} \tag{2.9}$$

O valor de S_{pq}, por sua vez, é obtido pela Equação 2.10 onde $A_{pq}=|S_{pq}|$ e Φ_{pq} representam, respectivamente, a amplitude e a fase do sinal.

$$S_{pq} = A_{pq}e^{i\Phi pq} \tag{2.10}$$

Já, o valor do retroespalhamento calibrado radiometricamente, usualmente chamado de sigma-zero (σ^0_{pq}) é calculado para cada polarização por meio da Equação 2.11. Nesta equação o valor do número digital (DN, do inglês *digital number*) é o valor do pixel na imagem e o fator de calibração (FC) é específico para cada sensor. Determinado valor tem importância nos processos de extração de atributos, na identificação de alvos e/ou na classificação de áreas de interesse por ser utilizado como valor de referência, calibrado, isento de fatores externos que modificam os valores radiométricos da imagem.

$$\sigma^0_{pq} = 10 * \log_{10} <DN^2> + FC \quad (2.11)$$

Os dados de SAR observados podem ainda ser representados por matrizes que relacionam o campo elétrico incidente e o espalhado (WOODHOUSE, 2006), como é o caso das matrizes de Müller [M] e de Kennaugh [K], ou por matrizes de potência de covariância [C] e de coerência [T], segundo Cloude e Pottier (1996). Estas matrizes são fundamentais na aplicação das técnicas de decomposição de alvos descritas por Cloude e Pottier (1996), por Freeman e Durden (1998) e por Touzi (2007), gerando uma grande quantidade de atributos que poderão melhor caracterizar os alvos. Neste caso os atributos podem ser divididos em incoerentes e coerentes.

Os atributos incoerentes independem dos dados de fase da matriz de espalhamento complexa, sendo extraídos somente a partir dos dados de amplitude. Dentre os atributos incoerentes destacam-se o coeficiente de retroespalhamento; as razões de polarização paralela e cruzada, descritas por Henderson e Lewis (1998); a potência total, descrita por Boerner et al. (1991); e os parâmetros apresentados por Pope et al. (1994) de índice de biomassa, de estrutura do dossel e de espalhamento volumétrico.

O coeficiente de retroespalhamento $(\sigma^0_{pq})_{dB}$ representa a quantidade de energia que retorna ao sistema radar após interagir com o alvo. A partir dos σ^0_{pq} os demais atributos incoerentes são calculados.

Conforme Henderson e Lewis (1998) o parâmetro de razão de polarização paralela (R_p) está associado à orientação e forma dos elementos espalhadores na floresta enquanto a razão de polarização cruzada (R_c) se refere ao espalhamento volumétrico do alvo. Já, a potência total (P_t) representa a soma de todos os mecanismos de espalhamento ocorrentes na floresta.

O parâmetro incoerente de Pope et al. (1994) referente ao índice de biomassa (BMI, *biomass index*) é um indicador da quantidade de estrutura lenhosa na floresta. Já, o de estrutura do dossel (CSI, *canopy structure index*) compara a estrutura vertical com a horizontal da vegetação. O de espalhamento volumétrico (VSI, *volume scattering index*) está relacionado com a densidade do dossel, sendo diretamente proporcional à quantidade de elementos que provocam espalhamento do tipo múltiplo.

Os atributos coerentes baseiam-se não somente nos dados de amplitude e intensidade, mas também nos de fase. Neste caso, para o cálculo, são utilizadas as matrizes de espalhamento complexa [S], de coerência [T] e de covariância [C]. Alguns dos atributos coerentes mais utilizados são a diferença de fase entre as imagens nas polarizações HH e VV, equivalente ao índice de tipo de interação (ITI, *interaction type index*) desenvolvido por Pope et al. (1994); a coerência polarimétrica entre as imagens nas polarizações HH e VV (HENDERSON e LEWIS, 1998); a entropia, a anisotropia e o ângulo alfa descrito pelas técnicas de decomposição de alvos de Cloude e Pottier (1996); as componentes de espalhamento volumétrico, dupla reflexão (*double bounce*) e superficial, segundo Freeman e Durden (1998); e a magnitude de Touzi, a fase de Touzi, o ângulo de orientação e o ângulo de helicidade, gerados pelas técnicas de decomposição de alvos de Touzi (2007).

A diferença de fase ($\Delta\varphi$) e a coerência polarimétrica (γ) entre as polarizações HH e VV são descritas por Henderson e Lewis (1998) em função dos termos de [S]. Enquanto o valor de $\Delta\varphi$ está relacionado ao tipo de espalhamento dominante, o γ determina o grau de correlação da informação de fase entre as imagens nas polarizações HH e VV. Henderson e Lewis (1998) afirmam ainda que as diferenças de fase e coerências polarimétricas envolvendo as polarizações cruzadas (HV e VH) não apresentam informação relevante, já que as mesmas apresentam-se ruidosas devido aos espalhamentos volumétricos que ocorrem em regiões de floresta.

Os parâmetros de decomposição de alvos desenvolvidos por Cloude e Pottier (1996) são calculados a partir de autovalores extraídos da matriz de coerência [T]. A entropia (H) determina o grau de aleatoriedade do processo de espalhamento e a equivalência nas contribuições dos espalhadores. Em casos que H apresenta valores intermediários, a anisotropia (A) representa a importância dos mecanismos secundários de espalhamento. Já, o ângulo alfa (α) descreve o tipo de mecanismo de espalhamento dominante no pixel.

A partir dos termos da matriz de espalhamento [S], Freeman e Durden (1998) desenvolveram três parâmetros que são diretamente associados ao tipo de espalhamento que cada alvo contribui para a potência total retroespalhada. O parâmetro de espalhamento

volumétrico (P_v) mede a contribuição deste tipo de espalhamento, simulando o dossel florestal. Já, o *double bounce* (P_d) é resultado de um conjunto de refletores de canto diédricos. Finalizando, o parâmetro de espalhamento superficial (P_s) mede o quanto este tipo de espalhamento contribui para a potência total.

Dentre os atributos obtidos por técnicas de decomposição de alvos, os de Touzi (2007) são os mais atuais e, segundo o autor, complementam os desenvolvidos por Cloude e Pottier (1996). Os parâmetros de Touzi são obtidos através dos autovalores da matriz [S] e têm as seguintes características: a magnitude de Touzi (α_s) fornece o tipo de simetria referente ao tipo de espalhamento do alvo; a fase de Touzi ($\Phi_{\alpha s}$) representa uma caracterização mais completa do tipo de espalhamento do alvo; o ângulo de orientação (ψ) está associado ao ângulo de inclinação do alvo; e o ângulo de helicidade (τ) permite a medida do grau de simetria de espalhamento do alvo, distinguindo espalhadores simétricos e assimétricos.

Outros atributos de SAR que poderão ser extraídos são os fornecidos nas medições interferométricas. Através dos DEM gerados pela interferometria é possível extrair a altura interferométrica (H_{int}) de cada pixel da imagem. Gama (2007) afirma ainda que os atributos de $LogH_{int}$ e de H_{int}^2 são ainda mais adequados na modelagem matemática para estimativa de biomassa.

Além destes atributos extraídos dos DEM, outro dado fornecido pela interferometria é a banda de coerência interferométrica, conforme citado na seção anterior. Esta banda produz uma melhora considerável no processo de caracterização dos alvos (GABOARDI, 2002), sendo, portanto, um importante atributo a ser utilizado. Thiel et al. (2009) mostram que além de melhorar a classificação pixel-a-pixel de imagens de SAR, a banda de coerência interferométrica ajuda no processo de segmentação destes dados, possibilitando uma posterior classificação orientada a segmentos que obterá melhores resultados.

Outro tipo de atributo que não é único de dados de SAR e que também pode ser extraído dos mesmos é o textural. Atributos deste tipo serviram como dados de entrada nos trabalhos de Simard et al. (2000), Kuplich et al. (2005) e Saatchi et al. (2007b) e obtiverem resultados de destaque.

Simard et al. (2000) gerou bandas de textura multiresolução a partir de imagens de SAR e concluiu que estas bandas foram úteis para distinguir diferentes classes de vegetação alagada. Maiores detalhes sobre o método de classificação utilizado neste trabalho serão apresentados no próximo item.

Já, Kuplich et al. (2005) utilizou o atributo referente à matriz de co-ocorrência de níveis de cinza (GLCM) de dados de SAR na banda L para estimar a biomassa aérea em

região de floresta amazônica. Os resultados obtidos mostraram que a inserção deste atributo no modelo gerado melhorou significativamente a associação entre dados de SAR e biomassa, obtendo um coeficiente de determinação máximo de $r^2 = 0,82$. Segundo Saatchi et al. (2007b), a partir dos atributos de textura se obtém informação a respeito da rugosidade da vegetação e da distribuição do tamanho das copas das árvores, ambos relacionados com a variação de biomassa florestal.

Concluindo esta seção, foi apresentada uma grande quantidade de atributos que podem ser extraídos de dados oriundos de sistemas de SAR, sendo específicos, ou não, dos mesmos. Cada um destes atributos extraídos foi utilizado em trabalhos anteriores visando caracterizar o alvo de interesse e, na maioria das vezes, foram avaliados separadamente com relação a esta capacidade. Em cada artigo descrito os respectivos autores ressaltam a relação que existe entre os atributos trabalhados e a quantidade de biomassa existente na região imageada. No entanto, para casos em que diversos atributos são utilizados simultaneamente, é necessário o uso de técnicas de aprendizado de máquina visando identificar quais atributos são os mais relevantes e se existe redundância entre eles. Algumas destas técnicas serão apresentadas na seção seguinte.

2.3 Aprendizado de Máquina

Michalski et al. (1983) apresentam a definição clássica de aprendizado de máquina – ML como o "estudo e a modelagem computacional dos processos de aprendizagem, em suas múltiplas manifestações". Originalmente um dos objetivos descritos era o de simular a forma humana de aprendizagem.

Porém, atualmente, com o desenvolvimento da ciência de dados, Faceli et al. (2021) descrevem ML como uma forma de análise de dados que orientam os computadores a aprenderem de modo autônomo, aprimorando o desempenho com relação a um assunto específico. Neste caso a "máquina" aprende de forma dinâmica, com novas experiências que podem continuamente alimentar o sistema e aprimorar a modelagem.

De forma semelhante, Brink et al . (2019) afirmam que são técnicas que se baseiam na análise de dados para encontrar padrões que podem ser definidos por modelos matemáticos. Neste caso, o ser humano desenvolvedor não escreve as instruções em linguagem computacional, não constrói os modelos, do que a máquina deve fazer. Quanto mais dados, mais o algoritmo de ML aprimora a modelagem para a execução de processos de forma eficiente, isto é, para aprender como obter o resultado esperado.

Segundo Russel e Norvig (2020), ML é um dos pilares da Inteligência Artificial – IA, sendo considerada uma subárea desta. Neste caso, IA é a capacidade de uma maquina em imitar características humanas, isto é, fazer o que o ser humano já faz, mas de forma digital: raciocínio, reconhecimento de fala, tomada de decisão, tradução de idioma.

Já Witten et al. (2016) diferenciam Mineração de Dados de ML, termos que são confundidos como sinônimos. A Mineração de Dados é realizada utilizando técnicas de ML aplicadas sobre grandes volumes de dados, os Big Data, extraindo informações úteis sobre bancos de dados que muitas vezes são caóticos: desorganizados, incompletos e de difícil compreensão.

Faceli et al. (2021) classificam as técnicas de ML entre supervisionado, não supervisionado, semi supervisionado (nem todos os dados são rotulados) e Aprendizado de reforço.

De acordo com Faceli et al. (2021) as principais vantagens de ML são:

- Análise de dados com eficiência, isto é, com o tempo reduzido, auxiliando na tomada de decisão oportuna;
- Adaptabilidade para atender ao problema proposto, aceitando novas configurações para atender as novas necessidades;
- Capacidade de analisar de forma aprofundada problemas específicos; e
- Capacidade na identificação de classes não previstas, não observadas anteriormente, realizando o agrupamento destes casos e identificando novas possibilidades de análises.

Diferentemente dos métodos paramétricos de modelagem, as técnicas de ML não utilizam as características estatísticas dos dados de entrada, tornando-se independente de suas distribuições. No âmbito do sensoriamento remoto, estas técnicas são as mais adequadas a serem aplicados sobre dados de SAR devido ao fato de não dependerem de padrões estatísticos para seu processamento.

A seguir encontram-se apresentadas as técnicas de ML aplicadas na presente tese.

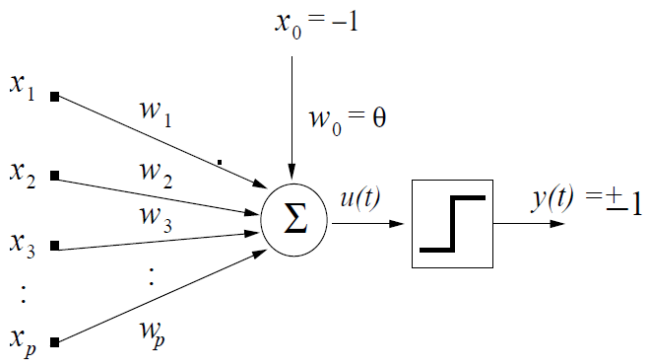
2.3.1 Redes Neurais Artificiais

Uma das abordagens mais utilizadas para o aprendizado de máquinas são os algoritmos de Redes Neurais Artificiais – RNAs. Segundo Bishop (1995), tais algoritmos são baseados em simulações simplificadas de neurônios reais, processando vários valores de entrada e apresentando um de saída. Para isto são utilizadas unidades de processamento simples, chamadas de nodos, que computam funções matemáticas. Cada nodo recebe valores

dos nodos pertencentes à camada anterior, ponderados por um peso que é calculado. No final do processamento, que envolve várias iterações, o valor de saída pode ser numérico ou uma determinada classe, rotulando assim o dado de entrada. Quando se deseja que o resultado do processo seja a classificação do dado de entrada, deve-se levar em consideração o limiar que será existente na camada de saída. Caso o valor calculado seja superior a este limiar (cujo valor é definido durante o processo inicial de aprendizagem), determinada amostra pertencerá, ou não, a uma classe.

Na Figura 2.1 pode-se observar um nodo de RNA onde x_i (i =1, …, p) são os valores dos dados de entrada, w_i (i =1, …, p) os pesos aplicados a cada atributo de entrada, θ o valor do limiar, *u(t)* o valor da função de entrada na interação *t* e *y(t)* o valor da função de saída na iteração *t*, isto é, da classificação, que neste caso de rede simples pode ser de +1 (pertence à classe) ou -1 (não pertence à classe).

Figura 2.1 – Nodo de Rede Neural Artificial
Fonte: Adaptada de Bishop (1995).

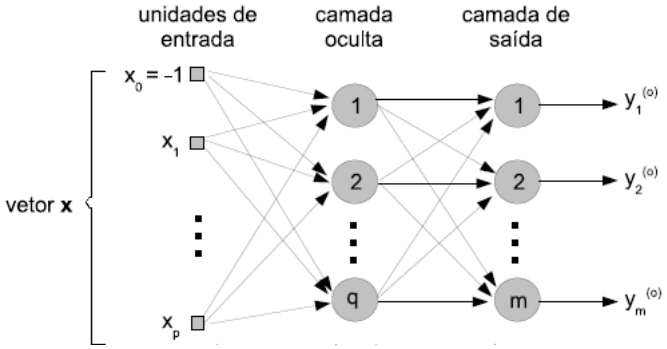


Visando solucionar problemas complexos de classificação, Bishop (1995) desenvolveu o perceptron de múltiplas camadas (MLP, do inglês *Multilayer Perceptron*), ilustrado na Figura 2.2. Além do vetor de entrada $\boldsymbol{x}$ e dos resultados classificados $\boldsymbol{y}$ na camada de saída (neste caso específico para m diferentes classes), a RNA possui camadas ocultas que modificam o espaço de atributos, deixando de atender somente a problemas lineares de classificação. Durante o processo de cálculo dos pesos $\boldsymbol{w}$ é utilizado o algoritmo de *backpropagation* buscando distribuir o erro calculado na camada de saída por todos os nodos de todas as camadas da rede. Maiores detalhes sobre o algoritmo são descritos por Samui et al. (2017).

Segundo Atkinson e Tatnall (1997) a quantidade de nodos e de camadas a serem utilizados na construção de um MLP depende de fatores relacionados ao problema a ser analisado tais como: a quantidade de atributos e do tipo e quantidade de dados de treinamento;

a quantidade e características das classes de saída; e do nível de complexidade do problema. Ao mesmo tempo, deve-se observar o nível de especialização da RNA construída, atentando-se para o *overfitting* e a consequente perda na capacidade de generalização do modelo.

Figura 2.2 – Perceptron de Múltiplas Camadas
Fonte: Adaptada de Bishop (1995).



O uso de RNA no sensoriamento remoto, mais especificamente da técnica de MLP, tem se mostrado adequado em inúmeros trabalhos. Na grande maioria dos trabalhos publicados a RNA desenvolvida buscou realizar a classificação de uso do solo utilizando dados de sensores ópticos. Em alguns casos foram aplicados sobre atributos radiométricos de dados multiespectrais (BISCHOF et al., 1992; PAOLA e SCHOWENGERDT, 1995; FOODY et al., 2003; MUUKKONEN e HEISKANEN, 2005), de dados hiperespectrais (LOGAN et al., 1997) ou sobre atributos de forma e contextuais (ANDRADE et al., 2003). No entanto, Del Frate e Solimini (2004) afirmam que o uso das RNAs independe do tipo de dado que irá ser utilizado, podendo inclusive ser dados de SAR. Isto se deve ao fato desta técnica poder lidar com o mapeamento não-linear de um espaço de entrada multidimensional se adequando a diferentes distribuições estatísticas.

Resultados nos trabalhos publicados mostram que uma RNA baseada em MLP com uma única camada oculta e com funções de ativação não-lineares é suficiente para gerar uma rede neural voltada a dados de SAR (JIN e LIU, 1997; TZENG e CHEN, 1998; DUTRA e HUBER, 1999; FOODY et al., 2003; KUPLICH, 2006). Outros trabalhos buscaram regras mais específicas de classificação utilizando duas camadas ocultas (DEL FRATE e SOLIMINI, 2004; MUUKKONEN e HEISKANEN, 2005), embora perdessem em poder de generalização (PAOLA e SCHOWENGERDT, 1995).

Outra vantagem das RNAs é que estas são capazes de realizar o processo de classificação supervisionada de dados utilizando uma quantidade menor de amostras de

treinamento. Paola e Schowengerdt (1995) afirmam que isto ocorre porque as regras de associação a uma classe são baseadas não somente nas características particulares dos dados de treinamento daquela classe, mas também nas das demais classes.

Por ocasião da presente pesquisa, o primeiro trabalho encontrado que buscou estimar biomassa com uso de dados de SAR e utilizando RNAs foi o de Jin e Liu (1997). Neste trabalho os autores estimaram características biofísicas de uma plantação de trigo, dentre as quais a biomassa, a partir de dados de PolSAR na banda X. As estimativas geradas pelo modelo desenvolvido foram comparadas através de gráficos com levantamentos *in loco* e analisadas visualmente, apresentando resultados compatíveis.

Posteriormente Foody et al. (2003) buscaram estimar a biomassa aérea de florestas tropicais a partir de índices obtidos pelas razões entre as bandas B_n, sendo n={1, 2, 3, 4, 5 ,7}, do sensor *Thematic Mapper*. O método utilizado constou inicialmente de um processo de seleção de atributos por ranqueamento onde as dez bandas com melhores resultados foram inseridas como entrada de uma RNA. Os melhores resultados foram alcançados com os índices B_4/B_3, $(B_4-B_3)/(B_4+B_3)$, $B_4/(B_1+B_2)$, $(B_1B_2)/B_3$, $B_4/B_1B_2B_3B_5B_7$ e $((B_4-(B_1+B_2))/((B_4+(B_1+B_2))$, obtendo um coeficiente de correlação de r = 0,71 para um modelo de regressão linear entre a biomassa medida *in loco* e a estimada.

Ao comparar as técnicas de estimativa de biomassa aérea por regressão linear e por RNAs, Muukkonen e Heiskanen (2005) obtiveram um resultado melhor para o primeiro caso. Os coeficientes de determinação obtidos por modelo de regressão linear entre a biomassa medida in loco e a estimada utilizando cada uma das técnicas foi de r^2=0,59 e 0,55, respectivamente. Neste trabalho os autores utilizaram dados ópticos ASTER sobre uma floresta boreal na Finlândia.

Como resultado da presente pesquisa, o único trabalho observado que buscou utilizar um algoritmo de RNA para estimar a biomassa florestal a partir de dados de SAR é o publicado por Del Frate e Solimini (2004). Os autores desenvolveram uma RNA com duas camadas ocultas que buscou modelar a relação entre os valores de retroespalhamento de dados de PolSAR nas bandas L e P com a biomassa aérea em regiões de floresta boreal. O coeficiente de correlação obtido com modelo de regressão linear entre os valores de biomassa estimada pelo algoritmo desenvolvido e o levantado *in loco* foi de r = 0,84, havendo destaque para a banda de polarização P-HV.

Atualmente, as RNAs evoluíram para o *Deep Learning*, capaz de realizar múltiplas interações onde são aplicados filtros convolucionais com parâmetros adaptativos e técnicas de *pooling*, visando o aprendizado aprofundado e computacionalmente eficiente das

características que associam os dados de entrada com os de saída (BALL et al., 2017). Segundo Borba et al. (2021) o aumento exponencial da aplicação desta técnica de ML sobre dados de sensoriamento remoto ocorre em função do aumento da capacidade computacional disponível para pesquisas nesta área.

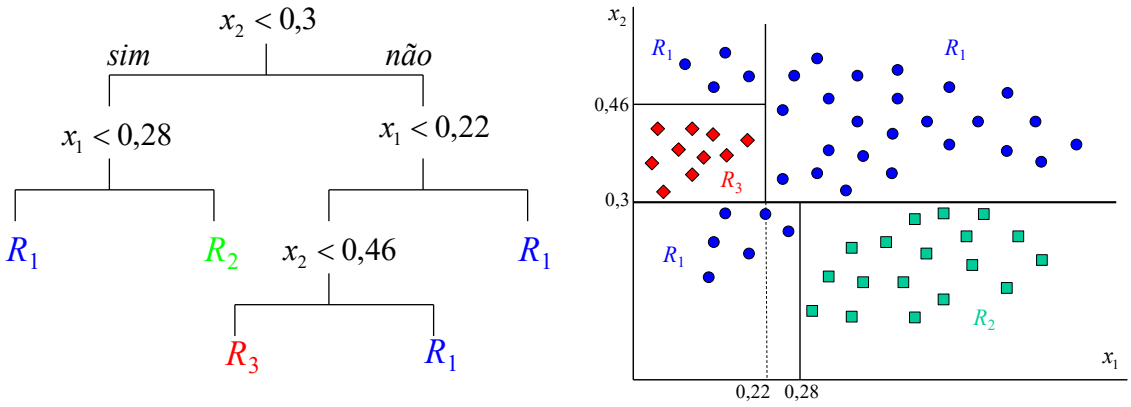
Para detalhes sobre *Deep Learning* aplicado a dados de sensoriamento remoto, sugere-se os trabalhos de Ball et al. (2017), Zhu et al. (2017) e Ma et al. (2019).

2.3.2 Árvore de Decisão

Outra abordagem utilizada para aprendizado de máquina é a de Árvore de Decisão – AD. Uma AD é constituída internamente de: (a) nós de decisão, que particionam o espaço de atributos através de superfícies de decisão e (b) nós terminais, ou "folhas", que permitem associar uma classe a cada objeto. Para atributos numéricos, utilizam-se comumente funções lineares como modelos para as superfícies de decisão. Serão retas para espaços de dois atributos, planos no caso 3-D, e, em geral (para *d* com valores maiores do que 3 atributos), serão hiperplanos.

A Figura 2.3 ilustra uma árvore de decisão univariada (ou OBCT - *Ordinary Binary Classification Tree*), suas regras de classificação e a respectiva representação da classificação dos dados em um espaço de atributos. Como exemplo, no primeiro nó de decisão, observa-se que a regra divide o espaço de atributos utilizando o eixo referente ao atributo x_2. Caso a amostra tenha um valor menor do que 0,3 no atributo x_2, ela será submetida posteriormente à regra de decisão do ramoi da esquerda. Caso contrário, irá para o ramo da direita. O mesmo acontece com a representação das amostras no espaço de atributos. Este é o caso mais simples de árvore de decisão, sendo chamada de univariada em função de suas regras estarem associadas somente a um único atributo.

Figura 2.3 – Exemplo de árvore de decisão univariada. Adaptado de Duda et al. (2001).



Em cada nó da OBCT, o hiperplano de decisão intercepta apenas um dos eixos coordenados, e seu modelo, para um dado atributo x_j é descrito pela Equação 2.12 onde **x** é o vetor de atributos do objeto, e o hiperplano intercepta o eixo j em $-w_0$.

$$f(x|j, w_0) = x_j + w_0 = 0 \tag{2.12}$$

De acordo com Quinlan (1993) e Saatchi et al. (2007b), esta abordagem tem como vantagem o fato de ser robusta por possuir natureza e propriedades não-paramétricas, podendo classificar imagens com distribuições estatísticas diferentes da gaussiana, heterogêneas e possuidoras de ruídos. Já, segundo Friedl e Brodley (1997), a principal vantagem das ADs é o fato de serem simples e flexíveis, realizando testes sequenciais e de fácil compreensão cuja semântica é praticamente intuitiva. Saatchi et al. (2007b) afirma que, para dados de sensoriamento remoto, isto permite a identificação de atributos relevantes para cada classe.

No entanto, por ocasião desta pesquisa, foi observada uma quantidade pequena de trabalhos publicados onde foram utilizadas ADs sobre dados de SAR. Na sua grande maioria, ADs são utilizadas para mineração de dados não-numéricos ou, dentro da área de sensoriamento remoto, para classificação de imagens ópticas.

Dentre os poucos trabalhos encontrados na presente pesquisa onde uma AD foi utilizada sobre um dado de SAR, destaca-se o de Simard et al. (2000). Neste trabalho, atributos de textura multiresolução foram gerados a partir de dados de SAR na banda L. Sobre estes atributos, junto às bandas de intensidade, foram construídas ADs para classificação de uso do solo, variando a quantidade de atributos de entrada, de classes e nível de poda da árvore. Os autores concluíram que, apesar dos atributos que mais se destacaram foram os de intensidade de SAR, os atributos de textura tiveram uma grande importância na diferenciação entre classes de regiões alagadas e secas. Juntamente a isto, o uso de ADs mostrou ser viável para classificação de imagens cujos dados de entrada possuem distribuições estatísticas diferentes.

Outros trabalhos onde foram construídas ADs para classificação de uso do solo utilizando dados de SAR foram os de Castro-Filho e Santos (2010) e de Castro-Filho (2010). Em ambos os trabalhos as ADs construídas utilizaram atributos extraídos de dados de PolInSAR do Projeto Radiografia da Amazônia e visavam a classificação de uso do solo. Nestes casos as classes e os dados usados como verdade de campo foram obtidos através dos mapas fitoecológicos do PROJETO RADAMBRASIL (1976 e 1977). Enquanto no primeiro trabalho foram aplicadas técnicas de SA sobre atributos extraídos somente de dados de SAR,

no segundo trabalho foram comparadas ADs construídas também com atributos extraídos de dados de sensores remotos ópticos.

Com relação à classificação para estimativa de biomassa, mais uma vez somente um único trabalho foi encontrado por este autor onde se utilizou uma AD. Saatchi et al. (2007b) utilizaram uma AD para identificar classes de biomassa acima de 150 t/ha com incrementos de 50 t/ha (totalizando sete classes). Para isto utilizaram diversos dados de sensoriamento remoto (mais especificamente de banda L de PolSAR, SRTM e dos sensores do MODIS) como entrada, os quais foram reamostrados para a resolução espacial de 1km. Como resultado obtiveram uma AD onde houve grande destaque para os atributos extraídos dos dados SRTM e cuja acurácia global chegou a 81%. Neste mesmo trabalho, Saatchi et al. (2007b) realizaram testes visando relacionar as classes de biomassa aos tipos de vegetação extraídos do PROJETO RADAMBRASIL (1976 e 1977) e aos atributos relacionados a índices climáticos, não obtendo um índice de correlação satisfatório.

Proposto por Breiman (2001), o método *Random Forest* (RF) compreende uma regra de classificação resultante de um conjunto de árvores de decisões (Dietterich et al. 2002). Formalmente, a partir de um conjunto de treinamento, são feitas n_{est} replicações, com a mesma cardinalidade, por amostragem. Para cada réplica, um subconjunto com até n_{att} atributos são considerados aleatoriamente e então usados para treinar uma única árvore de decisão. Parâmetros como a profundidade máxima a quantidade mínima de instâncias por folha precisam ser ajustados antes do processo de treinamento. Uma análise detalhada sobre estes parâmetros é apresentada por Breiman (2001).

Depois de treinar cada uma das n_{est} árvores, um dado vetor de atributo não rotulado $\boldsymbol{x}$ é classificado de acordo com uma classe cuja concordância entre as n_{est} árvores de decisão é máxima.

2.3.3 Máquina de Vetor de Suporte

Apresentado por Vladimir Vapnik (Vapnik, 1982, apud Cortes e Vapnik, 1995), o método de Máquina de Vetor de Suporte (SVM, do inglês *Support Vector Machine*) compreende um algoritmo de aprendizagem supervisionada que visa separar classes/categorias por meio de uma superfície $g(\mathbf{x}) = K(\mathbf{x}, \mathbf{w}) + b$ cuja margem é máxima, conforme descrito por Mountrakis e Ogole (2011); $\mathbf{w}$ e b são parâmetros que determinam a superfície de separação e K é uma função *kernel* adotada em função da complexidade do

problema em questão. Os tipos de função *kernel* são descritos por Shawe-Taylor e Cristianini (2004).

Conforme descrito por Bruzzone e Persello (2009), a partir de um conjunto de dados $\mathbf{D} = \{(\boldsymbol{x}_i, y_i) \in \chi \times \{-1,+1\} : i = 1,\ldots, m\}$, onde $y_i = \pm 1$ indica associação entre duas classes, o treinamento do método SVM compreende o cálculo dos parâmetros $\mathbf{w}$ e b depois de resolver o seguinte problema de otimização:

$$\max_{\alpha}\left(\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_i y_j \alpha_i \alpha_j \boldsymbol{K}(\boldsymbol{x_i}\boldsymbol{x_i})\right) \tag{2.13}$$

$$\text{considerando } \begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ 0 \le \alpha_i \le C \\ i = 1,\ldots,n \end{cases}$$

onde $\alpha_i \in \mathbb{R}$ são multiplicadores de Lagrange e $C \in \mathbb{R}_0^+$ é um fator de penalidade para erro de classificação. Sobre o kernel, a função de base radial (RBF) $\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{w}) = e^{-\gamma\|\boldsymbol{x_i}-\boldsymbol{w}\|^2}$, com $\gamma \in \mathbb{R}_0^+$, é destacada como uma opção usualmente utilizada.

Os multiplicadores de Lagrange obtidos ao resolver a Equação 2.13 permitem então definir a regra de classificação $G(\boldsymbol{x}) = \sum_{i=1}^{m} y_i \alpha_i \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x_i}) + b$ cujo valor é positivo, caso pertença à classe, ou negativo, caso não pertença.

2.3.4 Gradiente Reduzido Generalizado

Dentre os métodos de aprendizado de máquinas pouco aplicados ao sensoriamento remoto, inclui-se o do Gradiente Reduzido Generalizado (GRG) Não Linear, apresentado por Carpentier e Abadie (1966) e computacionalmente desenvolvido por Lasdon et al (1978). Ele também é simplesmente conhecido como "método de solução", ou *solver*, devido às diversas aplicações nas mais variadas áreas das ciências onde visa solucionar problemas matemáticos complexos. Atualmente, é o método de otimização mais utilizado em todo o mundo devido à disponibilização em diversos sistemas computacionais comerciais.

Apesar de ser o método de *solver* mais popular em pesquisas, seu maior problema está na incerteza de que a solução obtida seja realmente a melhor. Isto ocorre já que muitas vezes chega-se a uma solução ótima local ao invés de uma ótimo global, sendo este um fato inerente à natureza não-linear do problema. Nestes casos, na prática, termina-se o processo da otimização quando um ponto está suficientemente perto do ponto de solução (CIRILO, 1997).

Segundo Martinez e Santos (1998), a ideia presente no método GRG Não Linear é a seguinte:

- minimizar a função objetivo $z = f(\mathbf{X})$;
- tendo a matriz de parâmetros que se deseja encontrar $\mathbf{X} \equiv [x_i]^T$
- e sujeito a função de restrições $h(\mathbf{X}) = 0$, $\beta \geq x \geq \alpha$.

O problema no método GRG Não Linear deve ser iniciado com um $\mathbf{X}_k$ qualquer, preferencialmente próximo à solução. Através de um método iterativo, inicia-se a busca dos valores dos parâmetros utilizando a direção de busca identificada pelo gradiente reduzido $\nabla\varphi(\mathbf{X})$. Se o módulo do vetor gradiente reduzido $\nabla\varphi(\mathbf{X})$ for menor que a tolerância de convergência pré-definida, a variável $\mathbf{X}_k$ é tida como ponto ótimo da função.

2.4 Categorização

Em um problema computacional, os atributos a serem processados podem ser divididos em grupos de acordo com as características de seus dados. Conforme Yang e Webb (2009) atributos podem ser *qualitativos* ou *quantitativos*.

Atributos qualitativos possuem dados que não admitem operações aritméticas. Bussab e Morettin (2002) afirmam que um atributo qualitativo é nominal quando não existe nenhuma ordenação nas possíveis realizações e ordinal quando essas realizações podem ser ordenadas.

Atributos quantitativos admitem operações aritméticas e podem ser discretos, quando oriundos de uma contagem, ou contínuos, quando oriundos de uma medição (YANG e WEBB, 2009). Bussab e Morettin (2002) complementam, ainda, que os atributos discretos são aqueles que possuem os valores possíveis dentro de um conjunto finito. Por outro lado, os atributos contínuos possuem os valores possíveis pertencentes a um intervalo de números reais.

Os processos de transformação entre diferentes tipos de atributos possuem nomes específicos na literatura. A transformação de atributos quantitativos contínuos em quantitativos discretos é chamada de discretização. A discretização é feita particionando o intervalo de valores de um determinado atributo em vários subintervalos e associando um novo valor discreto específico a todas as instâncias que pertencem a cada subintervalo. Sendo A um atributo que possua valores ordenados $(x_1,...,x_i,...,x_j,...,x_m,...,x_N)$, $1 < i < j < m < N$, referentes às N instâncias e K o número de subintervalos, o processo de discretização D pode ser representado pela Equação 2.14.

$$D(A):\left\{\left[x_1;x_i\right]_1,\left[x_{i+1};x_j\right]_2,\ldots,\left[x_m;x_N\right]_K\right\} \quad (2.14)$$

Além de viabilizar a execução de alguns algoritmos referentes aos métodos de seleção de atributos e de classificação de dados, outros motivos para a discretização são o aumento na velocidade computacional e da interpretabilidade dos modelos de classificação gerados (LIU et al., 2002). Ambos os benefícios costumam ser observados, por exemplo, nos métodos de classificação por árvores de decisões que processam dados previamente discretizados.

Outro processo de transformação entre diferentes tipos de atributos é o de categorização. Para efeito do presente trabalho a categorização é definida como o processo que gera atributos qualitativos ordinais, podendo ser estes oriundos de atributos quantitativos discretos ou contínuos. O processo de categorização é semelhante ao de discretização, havendo somente a diferença de que todas as instâncias pertencentes a cada subintervalo receberão um valor categórico específico e não mais um valor numérico. Juntamente a isto, os valores categóricos recebidos deverão ser possíveis de ordenamento, pertencentes todos a um mesmo tema.

A categorização de um determinado atributo possui vasta aplicação na geração de mapas temáticos. Tais produtos cartográficos necessitam que o atributo quantitativo que será utilizado como tema seja categorizado justamente para que as categorias geradas possuam diferentes representações cartográficas. Para efeito deste trabalho, o atributo que será categorizado para ser utilizado como tema será chamado de atributo-tema que, no caso da construção da carta de estimativa de biomassa, é o atributo de biomassa.

Como exemplo de produto que passou pelo processo de categorização visando a geração de mapas temáticos, pode-se citar o Mapa Temático de Estimativa de Prevalência de Esquistossomose desenvolvido por Martins-Bedê et al. (2009). Neste caso, o atributo-tema possui valores quantitativos contínuos percentuais os quais foram associados a uma das três categorias de prevalência Baixa (0%,5%], Média (5%,15%] ou Alta (15%,100%].

A categorização pode, ainda, ser realizada em dois momentos da construção de produtos temáticos: na etapa final, onde o produto já se encontra pronto possuindo todos os valores quantitativos de cada área ou *pixel*; ou na etapa prévia à geração do modelo de clategorização do produto. Quando a categorização é realizada na etapa final, ela primeiramente captura o intervalo de valores do atributo-tema através da análise de todos os valores dos elementos existentes no produto já construído. Após isto, são construídos os subintervalos aos quais cada elemento será associado. Finalmente a cada subintervalo será designado um nome categórico. Este processo será chamado de *pós-categorização*. Por outro

lado, a categorização pode ser realizada na etapa prévia à geração do modelo de classificação do produto. Neste caso os elementos do produto temático ainda não possuem valores, sendo necessário definir um modelo de classificação a partir de amostras de treinamento. É sobre estas amostras que é realizado o processo de categorização, passando o mesmo a ser chamado de *pré-categorização*.

Conforme Dent et al. (2008), os métodos de categorização mais tradicionais são os de *intervalos iguais* e *quantil*. Em Liu et al. (2002), são descritos métodos não supervisionados para categorização já que levam em conta somente os valores do atributo-tema e não levam em consideração os valores dos demais atributos.

No método dos intervalos iguais a categorização é realizada dividindo-se os valores constantes da faixa de domínio do atributo-tema pelo número de classes de interesse. Neste caso serão obtidos subintervalos, os quais serão as categorias, de tamanhos iguais. Sendo K o número de categorias definidas pelo usuário e x_{min} e x_{max}, respectivamente, os valores mínimos e máximos observados no atributo-tema, então o método define categorias com larguras iguais a calculada pela Equação 2.15.

$$\delta=(x_{max} - x_{min}) / K \tag{2.15}$$

Portanto, a categorização pelo método dos intervalos iguais $Cat_{IntIguais}$ sobre o atributo-tema A_{tema} irá gerar o conjunto de categorias conforme a Equação 2.16.

$$Cat_{IntIguais}(A_{tema}): \{[x_{min}, x_{min}+\delta] ,]x_{min}+\delta , x_{min}+2\delta], \dots ,]x_{min}+(K\text{-}1)\delta , x_{max}]\} . \tag{2.16}$$

No método do quantil, a categorização é realizada dividindo-se o número total de instâncias N pelo número de categorias de interesse K, obtendo-se $n = \lfloor N/K \rfloor$ instâncias em cada categoria onde $\lfloor N/K \rfloor$ representa o maior inteiro menor ou igual a N/K. Portanto, a categorização pelo método do quantil $Cat_{Quantil}$ sobre o atributo-tema A_{tema} cujos valores mínimos e máximos sejam, respectivamente, x_{min} e x_{max} irá gerar o conjunto de categorias conforme a Equação 2.17.

$$Cat_{Quantil}(A_{tema}): \{[x_{min}, x_n] ,]x_n , x_{2n}], \dots ,]x_{(K\text{-}1)n} , x_{max}]\} \tag{2.17}$$

Ao analisar os *divisores de categorias* gerados por ambos os métodos, observa-se que no método dos intervalos iguais estes valores serão os que definirão os diferentes

subintervalos e são calculados por $x_{min}+(j-1)\delta$ com j variando de 1 a K. Por outro lado, no método do quantil os divisores de categorias serão calculados pela média aritmética entre os valores das instâncias adjacentes e pertencentes a diferentes subintervalos, obtendo-se o valor $x_{(j-1)n}$ onde j, mais uma vez, varia de 1 a K.

3 MATERIAL E MÉTODO

A seguir serão apresentados os materiais e o método utilizados na presente tese.

3.1 Material

Na presente tese serão utilizados dados SAR obtidos de diferentes sensores e dados de biomassa medidos *in loco* por técnicas de manejo florestal. As localidades da área de trabalho são do município de São Gabriel da Cachoeira – AM, localizado às margens do Rio Negro, e da Reserva Extrativista do Rio Unini – RESEX Unini, localizada na bacia do rio Unini, no município de Barcelos – AM.

Segundo o Portal Municipal da Prefeitura Municipal de São Gabriel da Cachoeira (2011), é o terceiro maior município brasileiro em extensão territorial (109.185,00 km^2), superando áreas de estados como Santa Catarina (95.346,18 km^2) e Pernambuco (98.311,62 km^2). Cerca de 80% de seu território é demarcado como reserva indígena, além de possuir também o Parque Nacional do Pico da Neblina, ponto culminante brasileiro com 2.993,78 metros de altitude.

Conforme MAPBIOMAS (2021) a vegetação encontrada nos blocos de processamento que serão trabalhados é de formação florestal. Já, de forma mais específica, o PROJETO RADAMBRASIL (1976) e o PROJETO RADAMBRASIL (1977) indica que são regiões fitoecológicas de contato florestal/formações edáficas (campinaranas). Estas regiões são caracterizadas de três formas:

- florestas densas, submontana e com o relevo dissecado. Neste caso, PROJETO RADAMBRASIL (1976) afirma que o volume médio de biomassa é de 107,4m^3/ha;
- florestas densas, submontana e com o relevo ondulado; e
- florestas densas, terras baixas e relevo com presença de platôs.

A RESEX Rio Unini, por sua vez, é uma unidade de conservação extrativista, com cerca de 833 hectares de extensão, e caracterizada no Projeto RadamBrasil (1977) como:
- floresta densa tropical, referente à sub-região dos baixos platôs da Amazônia; e
- áreas de tensão ecológica com densa presença aluvial.

Os dados utilizados de sensoriamento remoto são oriundos do sensor ALOS PALSAR 2 e do Projeto “Radiografia da Amazônia”. Em ambos os casos a área de trabalho está compreendida entre as latitudes 0° e 1° sul e entre as longitudes 67°w e 68°w.

Os dados do ALOS PALSAR 2 foram fornecidos pelo IBAMA e possuem as seguintes características:

- imagens no nível de processamento 1.1 – *Single Look Complex* (SLC), na banda L, quadri-polarizada (polarizações HH, HV, VH e VV), modo de imageamento Strip-map;
- 02 (duas) cenas (identificação ALOS2101587190-160410 e ALOS2101587180-160410), ambas de 10 de abril de 2016.

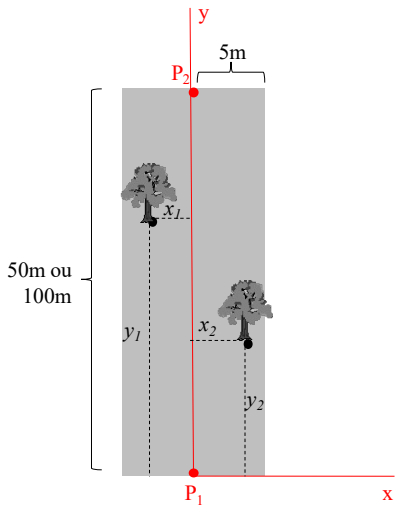
Já, os dados do Projeto "Radiografia da Amazônia" foram fornecidos pela Diretoria de Serviço Geográfico (DSG) do Exército Brasileiro, referentes ao imageamento realizado em 13 de agosto de 2010, e possuem as seguintes características:

- imagem de coerência interferométrica da banda P com resolução radiométrica de 8 bits. A imagem pode ser considerada como uma medida de qualidade dos DEM gerados;
- ortoimagens da banda X, em amplitude, com resolução radiométrica de 16 bits, polarização HH e resolução espacial de 5 metros;
- ortoimagens da banda P, em amplitude, com resolução radiométrica de 16 bits, quadri-polarizada e resolução espacial de 5 metros;
- MDS gerado a partir do processamento interferométrico dos dados oriundos da banda X, representa numericamente, em formato matricial, a altitude da superfície ao nível da copa das árvores (vegetação mais densa), com resolução radiométrica de 32 bits e espacial e 5 metros; e
- MDT gerado a partir do processamento interferométrico dos dados oriundos da banda P, representa numericamente, em formato matricial, altitude da superfície ao nível do solo, com resolução radiométrica de 32 bits e espacial e 5 metros.

Os dados de biomassa foram cedidos pelo Instituto Nacional de Pesquisas da Amazônia – INPA, e seguem os métodos desenvolvidos por Higuchi et al. (1998) e descritos por Silva (2007). Além da mesma posição geográfica dos blocos de processamento do Projeto "Radiografia da Amazônia", a proximidade com a data de imageamento da região também é importante pois visa evitar grandes mudanças na vegetação a ser analisada.

Foram inventariadas e cedidas 29 (vinte e nove) parcelas de biomassa, compostas pelo valor de biomassa aérea e total existente em cada parcela, em ton/ha, e as coordenadas em UTM dos pontos de início e fim de cada parcela. A Figura 3.1 ilustra o formato, os pontos de início e fim (P_1 e P_2) e as coordenadas arbitrárias dos indivíduos arbóreos dentro da parcela.

Figura 3.1 – Parcela de inventário florestal



Juntamente aos materiais utilizados na presente pesquisa, os programas computacionais utilizados são os seguintes:

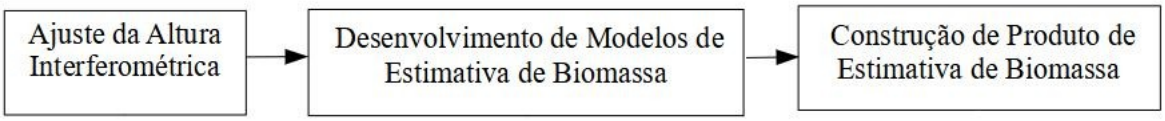
- ENVI (*Environment for Visualizing Images*), versão 4.7 (ENVI, 2009), e IDL (*Interactive Data Language*), versão 7.1 (IDL, 2009), e o SNAP (*Sentinel Application Plataform*), versão (6.0), para o tratamento dos dados de SAR;
- RAT (*Radar Tools*), versão 0.21 (RAT, 2009), e o PolSARPro (*Polarimetric SAR data Processing and Educational Tool*), versão 4.03 (POLSARPRO, 2009), para a calibração polarimétrica dos dados de radar; e
- Statistica, versão 8.0 (STATISTICA, 2007), para as análises estatísticas e de regressão; e o
- WEKA (*Waikato Environment for Knowledge*), versão 3.8 (WEKA, 2019), para o processo de mineração de dados.

3.2 Método

O método da presente tese segue as etapas constantes da Figura 3.2: ajuste da altura interferométrica (H_{int}) para desenvolvimento de modelo de estimativa de biomassa; desenvolvimento de modelos de estimativa de biomassa; e construção de produto temático de estimativa de biomassa, compondo cada parte um dos artigos apresentados.

Cada uma destas etapas possui inovações técnicas, seguindo as metodologias mais recentes, e estão apresentadas em forma de artigos técnicos. Cada artigo, por sua vez, busca aplicar técnicas de ML em diferentes etapas de construção de um produto temático, além de estar diretamente relacionado aos objetivos específicos da tese:

Figura 3.2 – Etapas / Artigos Técnicos da Tese



3.2.1 Artigo de Ajuste da Altura Interferométrica

O primeiro artigo possui como hipótese a possibilidade de ajuste da H_{int} a partir da identificação de áreas de solo exposto, isto é, onde o valor de H_{int} é, teoricamente, igual a 0 (zero). Além de inovadora, esta hipótese possibilita ajustar este atributo visando aprimorar a modelagem referente à estimativa de AGB.

Para realizar o ajuste proposto, primeiramente são identificadas áreas de solo exposto na região de estudo, isto é, em São Gabriel da Cachoeira – AM. Nesta etapa do trabalho são calculados os valores médios da H_{int} de cada uma destas áreas, excluindo pixels identificados como *outliers*.

Na sequência, cada um dos valores médios de áreas de solo exposto são aplicados em modelos matemáticos para gerar equações de ajuste da H_{int} de tal modo que estes valores, após os ajustes, se tornem próximos de 0 (zero). Para o cálculo dos parâmetros dos modelos matemáticos são utilizadas técnicas paramétricas (Método dos Mínimos Quadrados – MMQ) e de ML (Gradiente Reduzido Generalizado).

Ao término do artigo são realizados testes para analisar se os ajustes realizados sobre a H_{int} foram significativos e se possibilitaram melhoria na correlação com os valores de AGB medidos *in loco*.

No contexto da presente tese, o primeiro artigo aplica técnica de ML para ajustar valores de um atributo de SAR e buscar aprimorar a modelagem de estimativa de AGB. Neste sentido, esta aplicação encontra-se tanto relacionada ao objetivo geral da tese, como ao objetivo específico de analisar técnica de ajuste da altura interferométrica para o desenvolvimento de modelo de estimativa de biomassa.

Destaca-se o fato de que o artigo já encontra aprovados e publicado no Anuário do Instituto de Geociências da UFRJ, conforme apresentado no item 7, Apêndices, da presente tese.

3.2.2 Artigo de Desenvolvimento de Modelos de Estimativa de Biomassa

O segundo artigo dá continuidade ao primeiro e apresenta a hipótese de viabilidade na aplicação de técnicas de ML, sobre os atributos de SAR extraídos dos dados disponíveis, para construir modelos de AGB mais precisos. De forma inédita avalia e compara modelos de estimativa de AGB baseados em atributos qualitativos e quantitativos.

Para tal, inicialmente são processados os dados de AGB e de SAR e extraídos os atributos independentes para a modelagem. A planilha estruturada é então construída tendo as linhas como instâncias, referentes às parcelas medidas in loco, e as colunas como atributos de SAR.

A partir da planilha estruturada é iniciado o processo de construção de modelos de estimativa de AGB, sendo este o atributo-tema. A modelagem é realizada sobre o dado quantitativo (numérico original) ou qualitativo (categorizado por meio do método dos intervalos iguais ou do quantil); sobre atributos submetidos ao processo de seleção de atributos ou não; e aplicando técnicas de ML de *Multilayer Perceptron*, de *Suport Vector Machine*, de Árvore de Decisão Univariada e de *Random Forest*, em comparação às técnicas paramétricas de Regressão Simples, Regressão Múltipla e Regressão Logística.

Ao término do artigo são realizadas análises comparativas entre os modelos de AGB gerados por meio de matrizes de confusão e testes de hipótese.

No âmbito da presente tese, o segundo artigo aplica, de forma direta, diversas técnicas de ML para modelar a estimativa de AGB. Neste sentido, esta aplicação encontra-se tanto relacionada ao objetivo geral da tese, como aos objetivos específicos de analisar técnicas paramétricas e não-paramétricas para construção de modelos preditivos para estimativa de biomassa arbórea a partir de variáveis extraídas de dados de SAR, avaliando e apontando o mais adequado aos dados e região em trabalho.

Destaca-se o fato de que o artigo já encontra aprovados e publicado no *International Journal of Advanced Engineering Research and Science* (IJAERS), conforme apresentado no item 7, Apêndices, da presente tese.

3.2.3 Artigo de Construção de Produto de Estimativa de Biomassa

O terceiro e último artigo conclui o trabalho realizado na presente tese apresentando a hipótese de que técnicas de ML podem auxiliar na análise e construção de produtos temáticos de AGB. Nesta oportunidade foi desenvolvido um sistema computacional inovador que visa otimizar o processo de categorização, necessário à representação visual da geoinformação.

O sistema desenvolvido propõe o uso da técnica de “subida de colina” para buscar configurações de categorizações que otimizem o valor do índice de concordância Kappa do modelo construído e, consequentemente, a qualidade do produto temático final gerado. Neste sentido, são analisados diferentes intervalos das categorias, para cada qual é construído o respectivo modelo de AGB e calculando o valor de Kappa.

Ao término do artigo são apresentados os produtos temáticos de AGB que obtiveram os melhores resultados. Juntamente são feitas análises da representatividade destes produtos para as áreas em estudo.

No âmbito da presente tese, o terceiro artigo aplica técnica de ML para a construção do produto temático final de AGB. Neste sentido, esta aplicação encontra-se tanto relacionada ao objetivo geral da tese, como ao objetivo específico de implementar, analisar e propor metodologia para construção de produto temático que englobe as representações cartográficas referentes às categorias de biomassa e suas características.

O terceiro artigo foi devidamente submetido à revista técnica internacional e encontra-se em processo de revisão. Neste caso, o terceiro artigo consta do item 4 da tese, estando os documentos de submissão no item 7, Apêndices.

No caso do terceiro artigo, cujo conteúdo consta do item 4, buscou-se manter a formatação e idioma original da revista para o qual foi submetido, podendo haver variações além da língua portuguesa e das normas da ABNT.

4 ARTIGO – CATEGORIZATION OPTIMIZATION IN THE CONSTRUCTION OF THEMATIC PRODUCTS

Abstract

One of the steps in the construction of thematic products is the categorization. This process is done by partitioning the range of values of a given *feature* into several subranges and associating a new value, in this case ordinal, to all instances that are in that subrange. The objective of this work is to propose an innovative pre-categorization process that applies a computational search method to maximize the accuracy of thematic products during the classification stage. The theme *feature* used is the *A*bove *G*round *B*iomass (AGB), which estimation model is built *over* synthetic aperture radar *features*. The proposed heuristic is the hill climbing greedy, with the Kappa coefficient as the objective function. The results obtained shows that the proposed Categorization Optimization algorithm demonstrated the ability to obtain new states with subintervals of categories that increased the Kappa agreement index to 1.0 with much *lower* computational cost than the exhaustive search. The thematic *products* constructed maintained the representativeness of the study area while increasing in thematic accuracy.

Keywords: *categorization, search heuristic, thematic products, biomass, SAR.*

4.1 Introduction

Thematic products aims to represent geographic phenomena, physical or social, of the Earth's surface. Sometimes they are built from basic geoinformation, adding new layers of geoinformation to products referring to a specific theme, ie, thematic geoinformation (Raposo et al. 2020).

Originally thematic products are presented in the form of maps, which follow the language and grammatical rules defined by Bertin in publications from the 1970s (Bertin, 1977). However, currently thematic products are primarily intended to support decision-making by managers (Wu et al., 2019) as they provide useful information for spatial, urban and rural planning, change monitoring and ecosystem state assessments (Mitchell, 2018). With a focus on this objective, such products are presented in different ways, going beyond the classic representations of thematic maps, to queries in online WebGIS systems, with presentations in dashboards and the possibility of consumption via web online services.

One of the current highlights are products related to forest biomass stocks (Erb et al. 2018, Debastiani et al. 2019, Le Noe et al. 2020) that can support different types of programs,

such as Reducing Emissions from Deforestation and Forest Degradation (REDD+) and the launch of new orbital sensors such as the European Space Agency's Biomass Mission (Scipal et al. 2010).

In Brazil, the continuous monitoring program based on shallow deforestation data, called PRODES (PRODES, 2013) and the alert system for detecting deforestation in near real time, DETER (Diniz et al., 2015), both coordinated by the National Institute for Space Research (INPE), are examples of programs consolidated in the 1980s that provide online thematic products to support the fight against illegal activities in the Amazon rainforest in various formats, including raster, vector and summary tables.

Currently, more advanced Web GIS systems, which performs temporal analysis of spatial data with continental coverage, such as TerraBrasilis (Assis et al., 2019), have their own infrastructure with monitoring, filtering, cataloging and product availability services themed. The Map Biomass (2021), in turn, aims to prepare and make available thematic products specific to Brazilian biomes through the annual supply of land cover and use maps, image mosaics and statistics

One way to build thematic products is from remote sensing data, through classification methods (Mitchell, 2018, Foody, 2021). Critically, there are assumptions in the classification process that can degrade the accuracy of the product. One of these assumptions is that class labels and characterizations are rigid factors, that is, that they are defined before the start of the classification process and that they should not be changed during its execution. In these cases there are no interaction steps with the user.

However, regardless of the classification method adopted, without an accuracy assessment it is not possible for the user to decide whether it is fit for the purpose (Foody 2021). Wu (2019) highlights the importance of highly accurate and precise thematic maps, with focus on assessing the suitability and sustainability of prepared land for the planting of agricultural crops, with a strong impact on the value of the respective commodities and on the regional economy.

Furthermore, the process of generalizing the data used in building the thematic products, aiming to represent the phenomenon of interest, is one of the sources of error in the final product (Costa et al. 2017, Sluter et al. 2018). The nature and magnitude of this error, which deviates the product from reality, will vary depending on the method used (Foody 2021). For this reason, the generalization model should be selected in such a way that it seeks to degrade the accuracy of the final product as little as possible (Mitchell, 2018).

The generalization process, aiming at developing the thematic products in the form of maps, takes place on features that can be divided into groups, according to the characteristics of their data. According to Yang and Webb (2009) features can be qualitative or quantitative.

Qualitative features have data that do not support arithmetic operations. Bussab and Morettin (2017) states that a qualitative feature is nominal when there is no ordering in the possible realizations and ordinal when these realizations can be sorted.

Quantitative features supports arithmetic operations and can be discrete, when derived from a count, or continuous, when derived from a measurement (Yang and Webb 2009). Bussab and Morettin (2017) also adds that discrete features are those that have the possible values within a finite set. On the other hand, continuous features have possible values belonging to a range of real numbers.

Transformation processes between different types of features have specific names in the literature. Transforming continuous quantitative features into discrete quantitative features is called discretization. Discretization is done by partitioning the range of values of a given feature into several subranges and associating a specific new discrete value to all instances that belong to each subrange. Where A is a feature that has ordered values $(x_1,...,x_i,...,x_j,...,x_m,...,x_N)$, $1 < i < j < m < N$, referring to the N instances and K the number of subintervals, the discretization process D can be represented by Equation 4.1.

$$D(A): \{ [x_1;x_i]_1 , [x_{i+1};x_j]_2 , \dots , [x_m;x_N]_K \} \tag{4.1}$$

In addition to enabling the execution of some algorithms related to feature selection and data classification methods, other reasons for discretization are the increase in computational speed and the interpretability of the generated classification models, in addition to the decrease in the amount of data stored (Garcia et al. 2013, Bueno-Crespo et al. 2018, Rosenfeld et al. 2018). Both benefits are often seen, for example, in decision tree classification methods that process previously discretized data.

Several researches were carried out seeking to evaluate or propose discretization techniques (Garcia et al. 2013). Among the most recent reviews, Maslove et al. (2013) assess discretization techniques on clinical data. They conclude that supervised discretization techniques are more suitable for specific cases, while unsupervised techniques are more general.

Additionally, Bueno-Crespo et al. (2018) presents a fuzzy discretization proposal where each instance has a degree of belonging to different categories. In this case, the unsupervised algorithm demonstrated superiority to the traditional K-means technique.

Rosenfeld et al. (2018) highlights the fact that discretized features, when used together with the original numerical ones, contributes to the classification process. One of the explanations for this contribution lies in the fact that the feature selection process is implicit to the discretization process, which reduces the effect of the dimensionality curse in the development of predictive models.

Although the discretization process has computational and methodological advantages, improving the accuracy of the classification model in several studies, Rajbahadur et al. (2021) warn of the noise that can be generated due to this transformation. The authors emphasize that these noises are not safe for the classification process, generating models from samples that do not represent the characteristics, in general, of the region of interest.

Another transformation process between different types of features is categorization. Categorization is defined as the process that generates ordinal qualitative features, which can be derived from discrete or continuous quantitative features (Dent et al. 2008). The categorization process is similar to the discretization process, with the only difference that all instances belonging to each subinterval will receive a specific categorical value and no longer a numerical value. Along with this, the categorical values received must be possible to order, all belonging to the same theme.

The categorization of a given feature has wide application in the generation of thematic products. In this case it is necessary that the quantitative feature used as the theme is categorized so that the generated categories have different representations. The feature that will be categorized to be used as a theme will be called the theme-feature.

As an example of a product that went through the categorization process in order to generate thematic product, the Thematic Map for Estimating the Prevalence of Schistosomiasis was developed by Martins-Bedê et al. (2009). In this case, the theme feature has percentage continuous quantitative values which were associated with one of the three prevalence categories: Low (0.5%], Medium (5%,15%] or High (15%,100) %].

In a more recent research, Castro-Filho and Bias (2021) presents comparative analyzes between classifications carried out on numerical and categorized data. In this specific case, the authors sought to develop a biomass estimation model through the use of machine learning techniques on Synthetic Aperture Radar (SAR) data.

The categorization can also be carried out in two moments of the construction of the thematic product, that is, through two different methods: in the final stage, when the product is ready, having all the quantitative values of each area or pixel; or in the step prior to the generation of the classification model. When categorization is performed in the final step, it first captures the range of values of the theme feature by analyzing all the values of the existing elements in the already constructed product. After that, the subintervals to which each element will be associated are constructed. Finally, each subinterval will be assigned a categorical name. This process will be called post-categorization.

On the other hand, categorization can be performed in the step prior to the generation of the map classification model. In this case, the elements of the thematic map do not have values yet, and it is necessary to define a classification model based on training samples. It is on these samples that the categorization process is carried out, which is called pre-categorization.

The objective of this work is to propose a pre-categorization process that applies a computational search method to maximize the accuracy of thematic products during the classification stage. The theme feature used is Above Ground Biomass (AGB), which estimation model is built on SAR features described by Castro-Filho and Bias (2021). This objective is of great importance since the quality observed in the result of a classification process directly influences the thematic quality of products that use this type of data in their construction.

4.2 Method

4.2.1 Study Areas and Data

The study areas are located in different geographic regions of the Brazilian Amazon forest: São Gabriel da Cachoeira (SGC), a municipality located on the banks of the Rio Negro, in the northwest of the state of Amazonas; and the Rio Unini Extractive Reserve (Unini) located in the Rio Unini basin, in the municipality of Barcelos. The areas, in white, are highlighted in Figure 4.1.

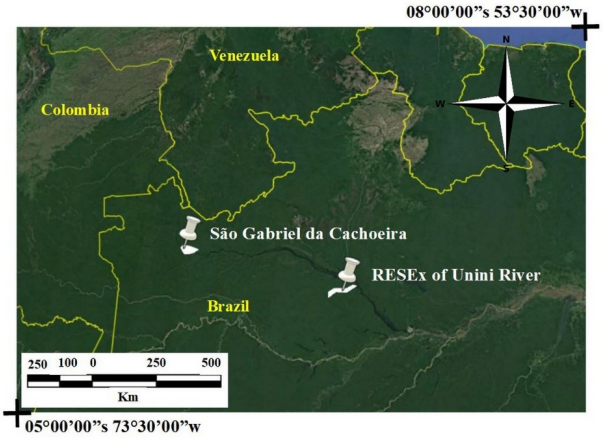


Figure 4.1 Location of the study areas, highlighted in white (Google Earth, 2021)

According to MAPBIOMAS (2021), the vegetation found in the study areas are of forest formation. More specifically, the RadamBrasil Project (1977) indicates that the vegetation found in the São Gabriel da Cachoeira area is composed of phytoecological regions of forest contact/edaphic formations (campinaranas). These regions are characterized in three ways:

- dense forests, submontane and with dissected relief. The RadamBrasil Project (1977) states that the average volume of AGB in the area is 107.4m3/ha;
- dense forests, submontane and with undulating relief; and
- dense forests, lowlands and relief with the presence of plateaus.

RESEX Rio Unini, in turn, is an extractive conservation unit, with about 833 hectares of extension, and characterized in the RadamBrasil Project (1977) as:

- tropical dense forest, referring to the sub-region of the Amazonian low plateaus; and
- areas of ecological tension with a dense alluvial presence.

Remote sensing data were obtained from the ALOS PALSAR 2 sensor and the Amazon Radiography Project. The working areas comprise between 0° and 1° south latitudes and 67° and 68° west longitudes, for the SGC region; and between 1° and 2° south latitudes and 62° and 63° west longitude, for Unini.

The AGB data were provided by the National Institute for Research in the Amazon – INPA, and follows the methods developed by Higuchi et al. (1998) and described by Silva (2007). The biomass data provided were composed of 128 inventoried plots, 58 plots from SGC and 70 from Unini, presenting the AGB values (ton / ha) and the UTM coordinates of the initial and final points of each plot.

Details about the characteristics and processing of the AGB and the SAR data, as well as the methodological approach in the construction of the structured spreadsheet used in the modeling, are described by Castro-Filho and Bias (2021). The present work will seek to focus on the proposal of the categorization optimization method and the building of the thematic map.

4.2.2 Categorization Optimization Algorithm

The proposed algorithm for categorization optimization follows the heuristic as shown in the flowchart in Figure 4.2.

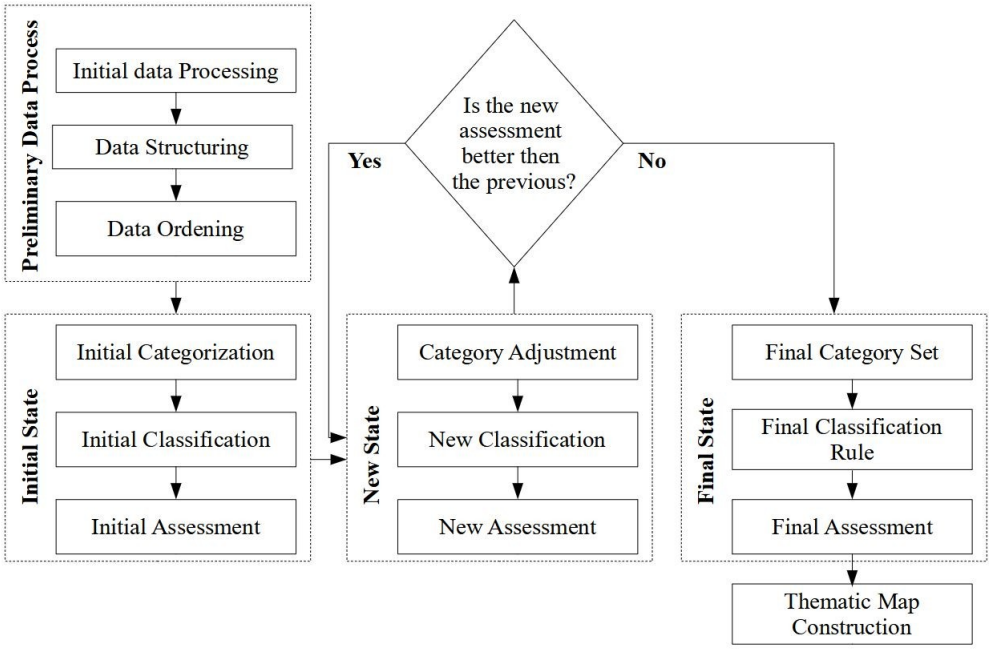


Figure 4.2 Methodological flowchart of the Categorization Optimization algorithm

The steps referring to the preliminary data process are described by Castro-Filho and Bias (2021). In general, this step aims to build a spreadsheet where the data is structured in lines, referring to instances, and columns, for the extracted features. Next, the data were sorted using the Quicksort technique applied to the AGB theme-feature. According to Theodoridis and Koutroumbas (2006), this ordering technique has computational complexity in the order of O(Nlog$N$), where N is the total number of instances of the theme feature.

The initial categorization, referring to the initial state, was obtained through the classic data categorization method of equal intervals. Liu et al. (2002) describe this method as

unsupervised for categorization since it only takes into account the values of the theme feature, despite the values of the other features.

According to Dent et al. (2008), in the equal intervals method, categorization is performed by dividing the value of the theme feature domain range by the number of classes of interest. Subranges will be obtained, which will be the categories, of equal size. With *K* being the number of user-defined categories and x_{min} and x_{max}, respectively, the minimum and maximum values observed in the theme feature, then the method defines categories with widths equal to that calculated by Equation 4.2.

$$\delta=(x_{max} - x_{min}) / K \tag{4.2}$$

Therefore, categorization by the equal intervals method Cat_{EqInt} over the theme feature A_{theme} will generate the set of categories according to Equation 4.3.

$$Cat_{EqInt}(A_{theme}):\{[x_{min}, x_{min}+\delta] ,]x_{min}+\delta , x_{min}+2\delta], \dots ,]x_{min}+(K\text{-}1)\delta , x_{max}]\} \tag{4.3}$$

Based on the initial categorization, the initial classification is performed using the method of Ordinary Binary Decision Tree (OBDT), version C4.5, due to the computational simplicity and ease of interpretation on the generated models (Hastie et al. 2009). As it is a non-parametric process, the method does not have requirements regarding the type and statistical distribution of the training data, being able to deal with continuous and discrete geoinformation simultaneously (Wu et al., 2019).

Next, the initial assessment is performed by the construction of the confusion matrix and the Kappa coefficient of agreement, using the leave-one-out cross-validation as test sample data. This procedure aimed to evaluate the ability to build the classification model from the parameters defined by the user in the process.

After the creation of the *initial state*, the proposed Categorization Optimization algorithm is started through a search process with the variation in the values of the category divisors, in such a way that different subintervals can be generated. Each set of subintervals is called a *new state* (Russel and Norvig 2020).

New states are defined in such a way that, if an instance is on the boundary between two categories, this instance is removed from the current one and inserted in the adjacent one. In this way, each of the (*K*-1) category dividers, where *K* is the number of categories, is moved to the left or to the right, varying the limits between adjacent categories. The *new*

states are then considered the *current state* of the algorithm, from which it is then possible to generate 2(*K*-1) different *new states* to be analyzed.

The proposed search method to optimize the categorization process is the *hill climbing*, using a greedy algorithm. This method is widely used in the field of Artificial Intelligence to search for optimal local values (Russel and Norvig 2020). For the applied hill climbing method, firstly an objective function is defined, which is wanted to be maximized. After this definition, the value of this function is checked in the *current state*, where the search is located, and in the *neighboring states*. If the value of the objective function in one of the *neighboring states* is higher than the *current state*, the algorithm performs a step to the *neighboring state*, making it the *current state* and restarting the entire search process.

Since the algorithm is greedy, the steps performed during the search will always be for the *neighboring state* that has the highest objective function value, ignoring the other states. In addition, the algorithm does not perform any further steps if none of the *neighboring states* has a value higher than the *current state*, ending the search process (Theodoridis and Koutroumbas, 2006).

As Russell and Norvig (2020) describe, the hill climbing method has the disadvantage of being "stuck" in local maximums and plateaus, failing to reach the global maximum. However, aiming to overcome this problem, the proposed algorithm uses a *floating step*, which will have the size defined by the user. Such floating step will only be performed by the algorithm if a more distant state have better results. In case the floating step is performed, the hill climbing algorithm restarts performing simple steps again, seeking to maximize the objective function.

The objective function defined for the proposed algorithm is the Kappa coefficient of agreement obtained by validating the result of a classification. It is observed that, as a consequence of this objective function, for each *search state* a process of categorization, classification and classification validation will be carried out in order to obtain the Kappa value. The pseudo-code of the algorithm is shown in Figure 4.3.

```
1: algorithm optimalCategorization
2:      input    structuredTable.txt
3:               K number of categories
4:               n minimum number of instances per category
5:               γ floating step
6:      sortedStructuredTable ← Quicksort (struturedTable, themeFeatureIndex)
7:      initialCategorizedState ← EqualIntervalCategorization (sortedStructuredTable)
8:      initialClassification ← Classification (initialCategorizedState)
9:      initialEvaluation ← Evaluation(initialClassification)
```

```
10:     finalCategorizedState ← initialCategorizedState
11:     finalEvaluation ← initialEvaluation
12:     for each K-1 category divider do
13:         if numberOfInstancesInPreviousCategory > n do
14:             newCategorizedState ← TransferIntanceFromPreviousCategory
15:             newClassification ← Classification (newCategorizedState)
16:             newEvaluation ← Evaluation (newClassification)
17:             if newEvaluation > finalEvaluation do
18:                 finalCategorizedState ← newCategorizedState
19:                 finalEvaluation ← newEvaluation
20:                 goto line 13
21:         FloatingStep (finalCategorizedState, γ)
22:         if numberOfInstancesInNextCategory > n do
23:             newCategorizedState ← TransferIntanceFromNextCategory
24:             newClassification ← Classification (newCategorizedState)
25:             newEvaluation ← Evaluation (newClassification)
26:             if newEvaluation > finalEvaluation do
27:                 finalCategorizedState ← newCategorizedState
28:                 finalEvaluation ← newEvaluation
29:                 goto line 20
30:         FloatingStep (finalCategorizedState, γ)
31:     return finalCategorizedState
32:     return finalEvaluation
```

Figure 4.3 Categorization Optimization algorithm pseudo-code

According to Bueno-Crespo et al. (2018), the proposed technique is classified as:

- dynamic, because the training phase of the machine learning technique is carried out together;
- local and univariate, the process being carried out only on the theme-attribute;
- incremental, as it starts with data already categorized;
- hybrid, the categories can be divided or grouped during the process;
- based on the evaluation method of the generated classification model;
- crisp for having rigid values of division between categories; and
- based on stopping criteria, as it is a greedy algorithm.

As input parameters of the proposed algorithm, the user must inform:

- the number of *K* categories you want when searching for the optimal categorization;
- the minimum number of instances per category; and
- the γ value of the floating step of the search.

It is also observed that the minimum number of instances per category limits the generation of *new states* to be analyzed. If the number of instances in a given category is lower than that informed by the user, that state will be considered invalid by the algorithm.

The Categorization Optimization algorithm was developed in the JAVA programming language in order to use the classes available in the WEKA data mining system (Waikato Environment for Knowledge Analyzes), version 3.8.4 (Witten et al. 2016).

After analyzing the complexities of all functions involved in the proposed algorithm, it is concluded that the heuristic complexity is $O(K^2N^2)$, where K is the number of categories and N is the number of instances.

4.2.3 Tests and parameters

The Categorization Optimization algorithm tests were performed with the following characteristics:

- separately for each study area, SGC and Unini;
- applied on the original values of the features and on the values transformed to logarithmics, as suggested by Gama et al (2010);
- based on two different scenarios, ie K values for 3 or 5 categories, with biomass labels as {Low, Medium, High} or {Low, Medium-Low, Medium, Medium-High, High};
- minimum number of instances per category equal to 2, aiming to seek the lowest possible restriction in the performance of steps between states; and
- floating step value equal to 4.

In order to analyze the search capability of the proposed heuristic, an exhaustive search algorithm was also developed to evaluate all the state space possibilities. Through this search, it is possible to follow the categorization status and assessment towards the maximum global value of the Kappa coefficient and compare it with that obtained by the proposed heuristic.

At the end of the process, two different ways of analyzing the results are carried out. The first analysis aims to verify the representativeness of the thematic map built in the proposed study area, that is, in specific Amazonian biomes. For this purpose, 10 (ten) selected sample areas are homogeneously distributed throughout the region, with dimensions between 40 and 50 thousand pixels. Next, the variations in the occurrence of the categories contained in the training samples are compared with the occurrences in the selected sample areas.

Additionally, a comparative analysis is made between the thematic maps generated through the Categorization Optimization, proposed heuristic, and the categorization performed by the equal intervals method, a classic process. This analysis takes place through the construction of a confusion matrix involving the pixels of both maps.

4.3 Results and discussion

The results obtained are presented in Tables 1 to 8, which differ according to the characteristics of the study area, the type of feature value and the number of categories. In each table are the results of the heuristic search, that is, the proposed categorization optimization algorithm, and the respective exhaustive search, performed with the same characteristics.

The cells of the tables referring to the heuristic search have the following information:

- number of categories;
- steps taken until the local maximum value of Kappa was obtained. In this case, the value in parentheses refers to the amount of fluctuations performed;
- initial Kappa value, obtained through categorization using the equal intervals method, and the final Kappa value, after using the proposed algorithm;
- size of the final OBDT obtained, related to the complexity of the constructed model;
- computational processing time; and
- the obtained sets of initial and final categories, both with respect to the instance index and the value of the AGB theme-feature.

On the other hand, the cells referring to the exhaustive search, in addition to the number of categories and the computational processing time, have information regarding:

- the number of analyzed states; and
- the maximum Kappa value found, including the respective number of possibilities and its percentage.

Table 4.1 Search results on logarithmic feature values of SGC for the set of 3 categories.

Test Name	**Heuristic Search Results**					
SGC-Log-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	3 (1)	0.93	1.00	15	1.74
	Initial Categorization Set Instances: {[1 , 5] ,]6 , 25] ,]26 , 58]} Values: {[1.96 , 2.21] ,]2.21 , 2.35] ,]2.35 , 2.55]}			**Final Categorization Set** Instances: {[1 , 5] ,]6 , 29] ,]30 , 58]} Values: {[1.96 , 2.21] ,]2.21 , 2.37] ,] 2.37 , 2.55]}		
	Exhaustive Search Results					

	Analysed states	Max Kappa	Processing Time (sec)
	1431	1.00 (51 possibilities – 3,5% of the analysed states)	9.95

Table 4.2 Search results on logarithmic feature values of SGC for the set of 5 categories.

Test Name	Heuristic Search Results					
SGC-Log-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	4 (0)	0.88	0.94	19	1.94
	Initial Categorization Set Instances: {[1 , 4] ,]5 , 7] ,]8 , 18] ,]19 , 52] ,]53 , 58] } Values: {[1.96 , 2.11] ,]2.22 , 2.24] ,]2.26 , 2.30] ,]2.32 , 2.43] ,]2.44 , 2.55]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 5] ,]6 , 18] ,]19 , 52] ,]53 , 58] } Values: {[1.96 , 1.99] ,]2.00 , 2.21] ,]2.22 , 2.31] ,]2.32 , 2.43] ,]2.44 , 2.55]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	270,676	1.00 (693 possibilities – 2,5% of the analysed states)				1,666.19

Table 4.3 Search results on original feature values of SGC for the set of 3 categories.

Test Name	Heuristic Search Results					
SGC-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	2 (1)	0.91	1.00	15	1,92
	Initial Categorization Set Instances index: {[1 , 7] ,]8 , 51] ,]52 , 58]} Values: {[92.21 , 171.86] ,]181.16 , 263.87] ,]272.22 , 351.73]}			**Final Categorization Set** Instances: {[1 , 6] ,]7 , 54] ,]55 , 58]} Values: {[92.21 , 167.58] ,]171.86 , 290.89] ,]301.21 , 351.73]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	1431	1.00 (66 possibilities – 4,6% of the analysed states)				9,63

Table 4.4 Search results on original feature values of SGC for the set of 5 categories.

Test Name	Heuristic Search Results					
SGC-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	3 (0)	0.95	0,98	21	1.82
	Initial Categorization Set Instances: {[1 , 4] ,]5 , 15] ,]16 , 38] ,]39 , 54] ,]55 , 58] } Values: {[92.21 , 132.50] ,]163.30 , 195.34] ,]197.37 , 247.80] ,]248.30 , 290.89] ,]301.21 , 351.73]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 15] ,]16 , 37] ,]38 , 54] ,]55 , 58] } Values: {[92.21 , 96.95] ,]101.69 , 195.34] ,]197.37 , 247.31] ,]247.80 , 290.89] ,]301.21, 351.73]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	270,676	1.00 (965 possibilities – 3,6% of the analysed states)				1,972.82

Table 4.5 Search results on logarithmic feature values of Unini for the set of 3 categories.

Test Name	Heuristic Search Results					
Unini-Log-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time**

						(sec)
	3	0	0.98	0.98	21	1.74
	Initial Categorization Set Instances: {[1 , 18] ,]19 , 40] ,]41 , 70]} Values: {[2.19 , 2.28] ,]2.29 , 2.39] ,]2.40 , 2.49]}			**Final Categorization Set** Instances: {[1 , 18] ,]19 , 40] ,]41 , 70]} Values: {[2.19 , 2.28] ,]2.29 , 2.39] ,]2.40 , 2.49]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	2145	1.00 (58 possibilities – 2,7% of the analysed states)				13.86

Table 4.6 Search results on logarithmic feature values of Unini for the set of 5 categories.

Test Name	**Heuristic Search Results**					
Unini-Log-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	0	0.96	0.96	25	1.86
	Initial Categorization Set Instances: {[1 , 2] ,]3 , 23] ,]24 , 34] ,]35 , 59] ,]60 , 70] } Values: {[2.19 , 2.22] ,]2.25 , 2.30] ,]2.31 , 2.37] ,]2.38 , 2.43] ,]2.44 , 2.49]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 23] ,]24 , 34] ,]35 , 59] ,]60 , 70] } Values: {[2.19 , 2.22] ,]2.25 , 2.30] ,]2.31 , 2.37] ,]2.38 , 2.43] ,]2.44 , 2.49]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	635,315	1.00 (700 possibilities – 0,1% of the analysed states)				5,625.84

Table 4.7 Search results on original feature values of Unini for the set of 3 categories.

Test Name	**Heuristic Search Results**					
Unini-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	3 (1)	0.94	0.96	23	1.92
	Initial Categorization Set Instances: {[1 , 23] ,]24 , 50] ,]51 , 70]} Values: {[153.32 , 202.15] ,]206.11 , 257.38] ,]258.87 , 311.57]}			**Final Categorization Set** Instances: {[1 , 24] ,]25 , 47] ,]48 , 70]} Values: {[153.32 , 206.11] ,]210.08 , 254.09] ,]254.99 , 311.57]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	2145	1.00 (104 possibilities – 4,8% of the analysed states)				20.82

Table 4.8 Search results on original feature values of Unini for the set of 5 categories.

Test Name	**Heuristic Search Results**					
Unini-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	6 (2)	0.87	0.98	29	2.73
	Initial Categorization Set Instances: {[1 , 8] ,]9 , 28] ,]29 , 41] ,]42 , 62] ,]63 , 70] } Values: {[153.32 , 184.15] ,]185.33 , 214.71] ,]218.13 , 248.24] ,]250.32 , 277.58] ,]281.85 , 311.57]}					
	Final Categorization Set Instances: {[1 , 9] ,]10 , 24] ,]25 , 37] ,]38 , 59] ,]60 , 70] } Values: {[153.32 , 185.33] ,]185.67 , 206.11] ,]210.08 , 237.05] ,]240.17 , 267.57] ,]270.44 , 311.57]}					
	Exhaustive Search Results					

	Analysed states	Max Kappa	Processing Time (sec)
	635,315	1.00 (1,614 possibilities – 0,3% of the analysed states)	9,181.66

The results obtained in the SGC-Log-3 (Table 4.1) and SGC-3 (Table 4.3) tests shows that the proposed heuristic is capable of reaching the 1.00 Kappa index value, the global maximum. In the other cases, in general, the steps between the states also provided an increase in that index.

However, it is observed that in the Unini-Log-3 (Table 4.5) and Unini-Log-5 (Table 4.6) tests, the method of equal intervals performed the initial categorization within a local maximum of Kappa. In this case, the heuristic did not identify state steps that would increase this value, maintaining the same initial and final Kappa values.

Similar situations involving local maxima are found in the SGC-Log-3 (Table 4.1), SGC-3 (Table 4.3), Unini-3 (Table 4.7) and Unini-5 (Table 4.8) tests. In these cases the hill climb method found a local maximum, however, it was surpassed due to the floating step performed.

Another result observed is related to the size of the OBDT built for the categorization process. The trees associated with the SGC study area had sizes between 15 and 21, smaller than those obtained for the Unini area, which ranged between 21 and 29. According to Frank et al. (2016), the smaller the tree, the greater its generalizability, in addition to lower computational complexity and cost.

Regarding processing time, the developed heuristic search took 1 to 3 seconds to identify a maximum Kappa value and build a categorization model. This characteristic proves the advantage of using the hill climb search method over the exhaustive search, which took, in the extreme case of the Unini-5 (Table 4.8) test, more than 2.5 hours to analyze all the possibilities.

The exhaustive search algorithm also presented good results, analyzing all possible states and identifying those referring to the maximum global Kappa index. The maximum Kappa percentages found in the exhaustive searches ranged between 0.1% and 4.8% of the analyzed states, without following a specific pattern.

From the results, 4 (four) tests that stood out for the different areas of study and number of categories were selected: SGC-3, SGC-5, Unini-Log-3 and Unini-5. On these tests, comparative analyzes were carried out between thematic maps and those of representativeness of areas.

About the tests performed, the OBDT that were built selected the most significant features. Table 4.9 presents these features that are detailed by Castro-Filho and Bias (2021).

The most prominent features are the interferometric height (Hint) and the texture and target decomposition extracted over the L-band.

Table 4.9 Characteristics of the constructed OBDT

Test	Tree Size	Selected Features
SGC-3	15	• Variance texture feature for 3x3 pixels window size over L_{hv} band (3_LHV_Var) • S1 orientation angle (ψ) over L band (TPSI1L) • Interferometric height between X and P bands (Hint) • Ratio between parallel polarizations over L band (RP_L) • Variance texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Var) • Amplitude image of the X band in the HH polarization (Xhh) • Amplitude image of the P band in the HV polarization (Phv)
SGC-5	21	• Correlation texture feature for 5x5 pixels window size over L_{hh} band (5x5_LHH_Cor) • Subtration between amplitudes of HH and HV polarizations in L band (Lhh-Lhv) • Contrast texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Con) • Homogeneity texture feature for 3x3 pixels window size over L_{hh} band (3x3_LHH_Ho) • Correlation texture feature for 7x7 pixels window size over X_{hh} band (7x7_XHH_Cor) • Mean texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Me) • Amplitude image of the L band in the HV polarization (Lhvvh) • S2 magnitude (α) over L band (TAlphaS2L) • S3 magnitude (α) over L band (TAlphaS3L)
Unini-Log-3	21	• Interferometric height between X and P bands (Hint) • Mean texture feature for 7x7 pixels window size over L_{hh} band (7x7_LHH_Me) • Entropy texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_En) • Amplitude image of the P band in the HV polarization (Phvvh) • Amplitude image of the P band in the HH polarization (Phh) • Contrast texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Con) • Variance texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Var) • Odd Scattering over L band (VanZOddL) • Variance texture feature for 3x3 pixels window size over L_{hv} band (3x3_LHV_Var)
Unini-5	27	• Interferometric height between X and P bands (Hint) • Contrast texture feature for 3x3 pixels window size over L_{hv} band (3x3_LHV_Con) • Variance texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Var) • Variance texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Var) • Mean texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Me) • Dissimilarity texture feature for 7x7 pixels window size over P_{hh} band (7x7_PHH_Di)

		• Homogeneity texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Ho) • S2 magnitude (α) over L band (TAlphaS2L) • Subtration between amplitudes of HH and HV polarizations in L band (Lhh-Lhv) • Subtration between amplitudes of HH and HV polarizations in P band (Phh-Phv) • Amplitude image of the L band in the VV polarization (Lvv) • Correlation texture feature for 7x7 pixels window size over X_{hh} band (7x7_XHH_Cor) • Entropy texture feature for 3x3 pixels window size over P_{hv} band (3x3_PHV_En)

The thematic representations were built on a delimited area of 5x5 km, referring to a sample of 1 million pixels in the 5 meters spatial resolution images. The areas were selected in the SGC and Unini regions in such a way that they were predominantly primary forests, with no influence of other natural or anthropogenic elements that could affect the representativeness of the samples, such as rivers and settlements. The representation of the biomass categories followed a choropleth representation, as shown in Figure 4.4.

From the products constructed, it is observed that those referring to the SGC-3 (Figure 4.4 A) and SGC-5 (Figure 4.4 B) tests have a speckled appearance, of "salt and pepper", with the visually balanced and homogeneous presence of all biomass categories. It is noteworthy that this obtained appearance is typical of the SAR data nature. Authors highlights that, even after using adaptive filters and increasing the equivalent number of looks in the SAR image (Woodhouse 2017, Pereira et al. 2018), steps that reduces the effects of the speckle noise, the geometric characteristics of the imaged targets remain evident (Sarker et al 2012). In fact, the imaged primary forest presents textural characteristics of granular roughness, providing homogeneous variation between biomass categories.

On the other hand, the thematic map of the Unini-Log-3 (Figure 4.4 C) test shows less categories homogeneity, with a greater presence of low biomass areas in the upper left corner. This presence also contributes to the identification of a sinuous shape of a probable river, which had not been visually identified previously.

The result of the Unini-5 (Figure 4.4D) test, in its turn, showed a large presence of the medium biomass category, with few pixels for the other categories.

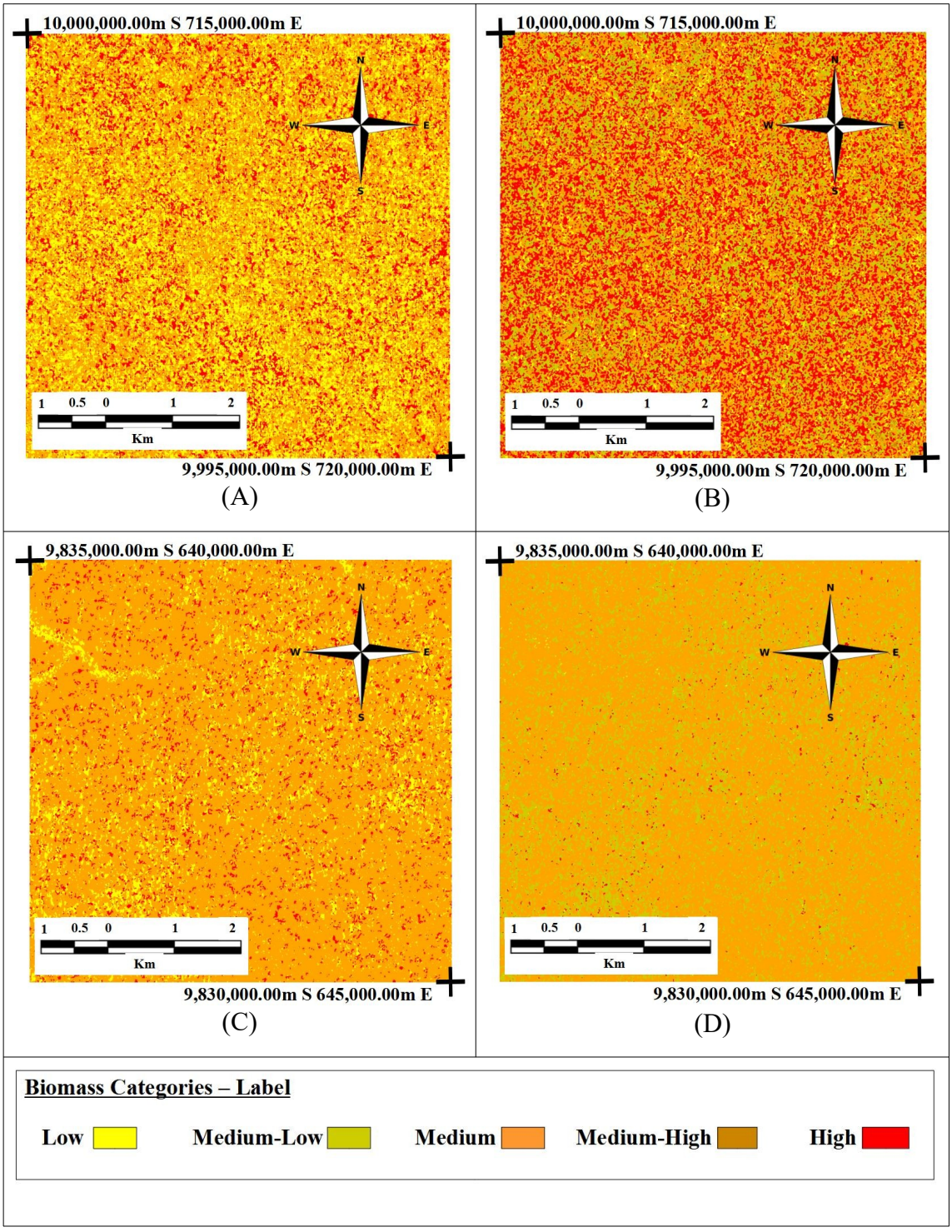


Figure 4.4 Thematic maps of biomass estimation referring to tests (A) SGC-3, (B) SGC-5, (C) Unini-Log-3 and (D) Unini-5

In order to analyze the representativeness of the biomass categorization and estimation models built for both study areas, analyzes were performed on 10 (ten) sample areas, which results are shown in Figures 4.5 to 4.8. In the graphs there are percentages of pixels, for each category, found in the sample areas randomly distributed in the respective study areas,

compared to the percentage of categories obtained by the categorization models by the method of equal intervals and by the proposed heuristic.

In this analysis, the graphs shows that the sample areas of the SGC-3 (Figure 4.5) and SGC-5 (Figure 4.6) tests presented a decrease in the average biomass category, with the respective increase in the categories with more extreme values. Despite these changes, it is observed that both equal-interval and heuristic methods were able to satisfactorily model the SGC study area, for 3 or 5 categories, with small modifications.

Conversely, the sample areas of the Unini-Log-3 test (Figure 4.7) showed a large increase in the average biomass category. However, a coherent distribution of extreme biomass values was maintained, that is, low and high biomass.

The sample areas of the Unini-5 test (Figure 4.8), in turn, did not present a distribution of categories equivalent to those built in the models. In these cases, the extreme values of categories were significantly reduced, with almost zero occurrence of the category of medium-low biomass.

It is also noteworthy that, in all graphs, the numerical value of biomass for each category are found next to the legends. The values do not present any type of anomaly that justifies a possible problem of representativeness of the study areas, that is, no cases were observed where there is a discontinuity of values or where the category intervals are insignificantly small to the point of suggesting that it does not exist. The fact that the heuristic started from a classical categorization, that is, by equal intervals, contributed for coherent numerical AGB results.

The comparative analysis between the thematic products generated through the Categorization Optimization, proposed heuristic, and the categorization performed by the method of equal intervals, a classic process generated the confusion matrices contained in Tables 4.10 to 4.12. In all of them, the values are found in percentage, with emphasis on the main diagonal in bold. The confusion matrix referring to the Unini-Log-3 test was not built, since both thematic maps are identical because there were no steps between states during the execution of the algorithm, as shown in Table 4.5.

The confusion matrix referring to the SGC-3 test (Table 4.10) shows that there is low coherence between the thematic products constructed. In this case, the proposed algorithm performed significant changes in pixel values, allocating them to other categories.

On the other hand, the SGC-5 test matrix (Table 4.11) shows strong coherence between the thematic maps. Likewise, Table 4.4 shows that there was little change in the Kappa index value, ranging only from 0.95 to 0.98.

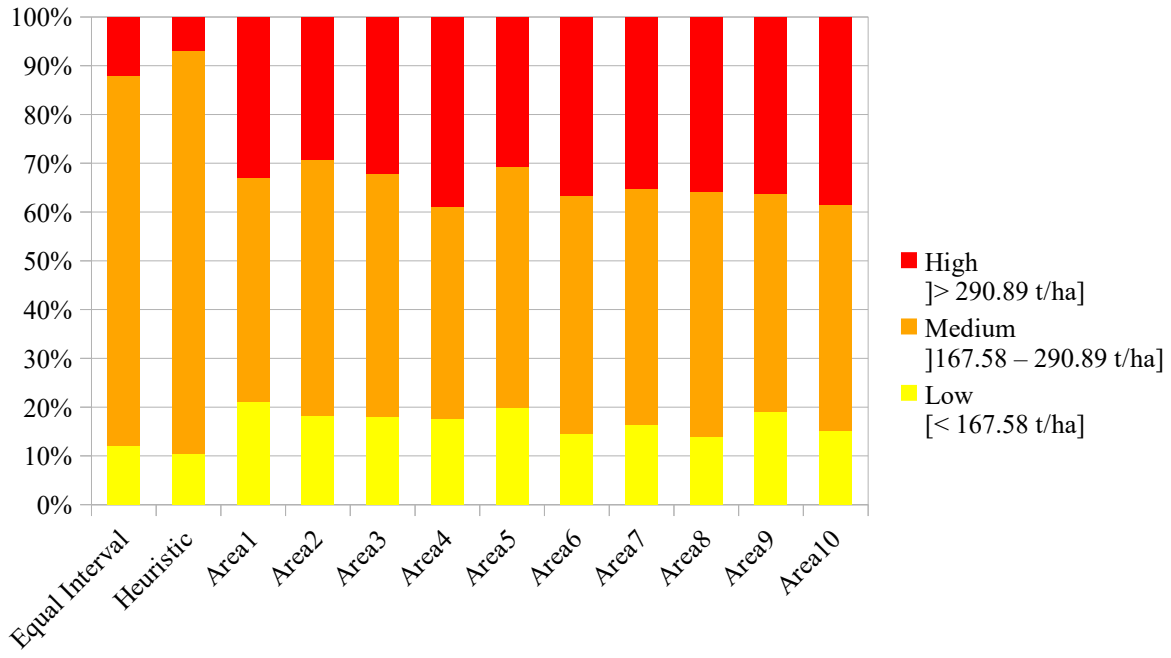


Figure 4.5 Graph showing the representativeness of the study area for the SGC-3 test.

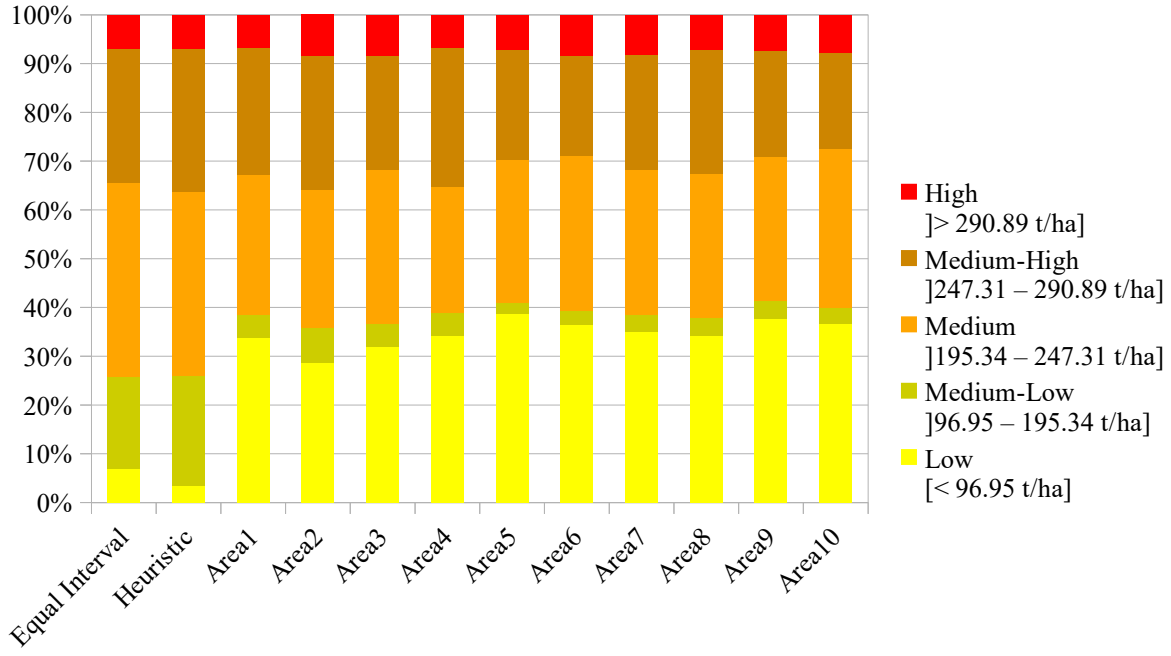


Figure 4.6 Graph showing the representativeness of the study area for the SGC-5 test.

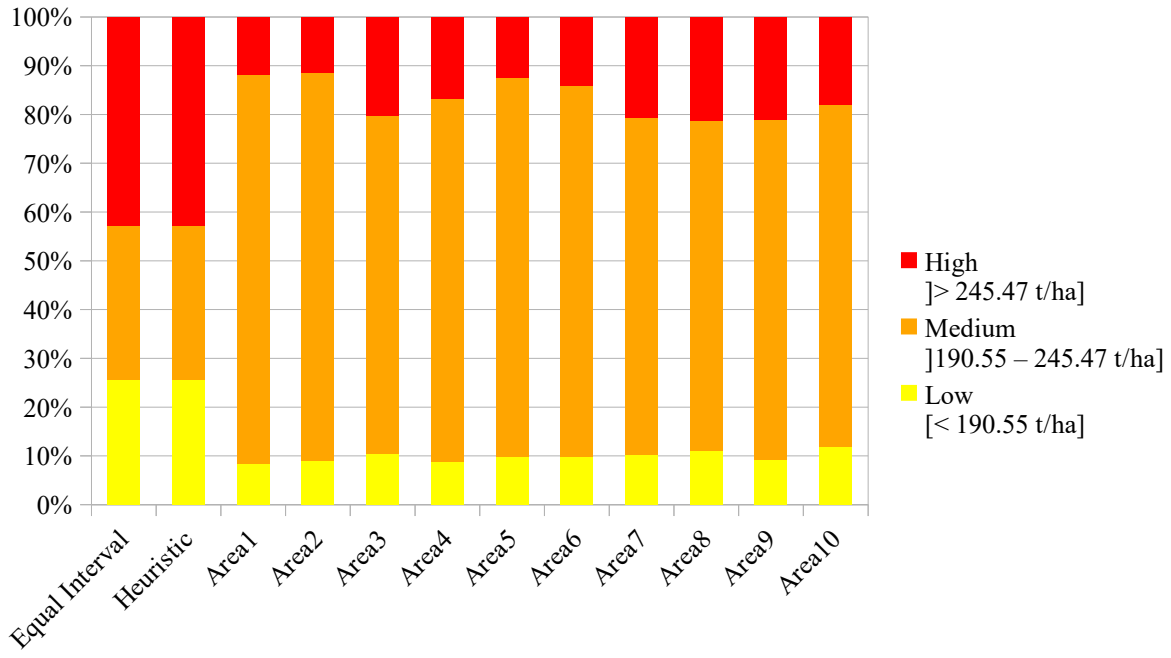


Figure 4.7 Graph showing the representativeness of the study area for the Unini-Log-3 test.

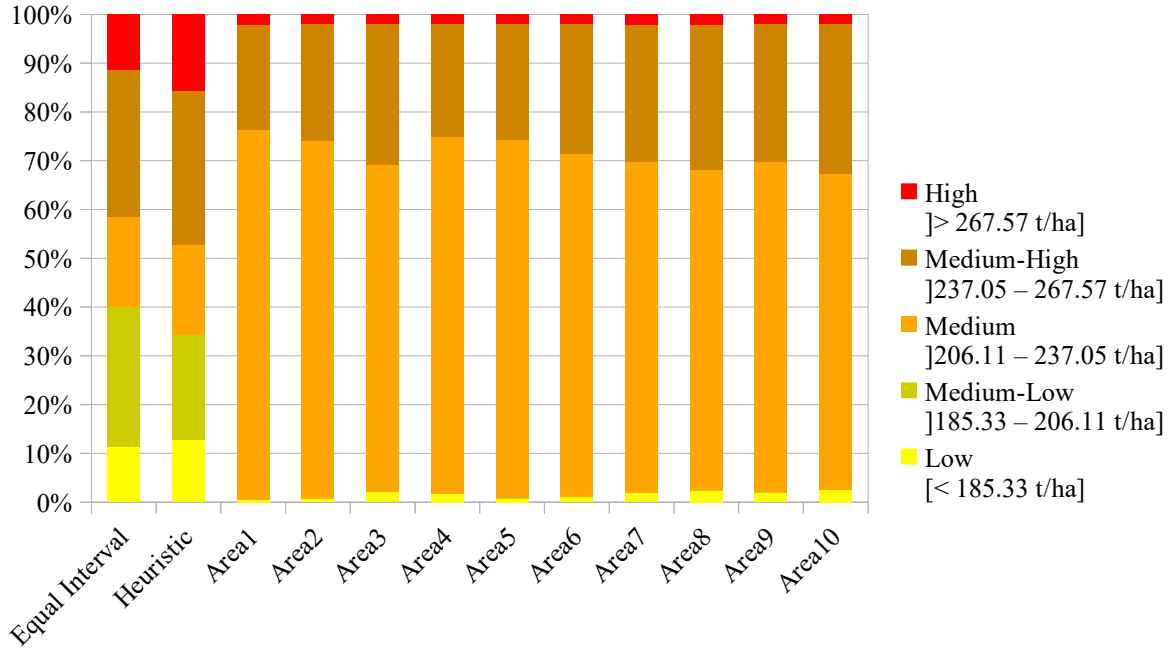


Figure 4.8 Graph showing the representativeness of the study area for the Unini-5 test.

The confusion matrix referring to the Unini-5 test (Table 4.12) showed coherence only for the medium and high-medium categories, not for the others. In this case, the area defined for the construction of the thematic product, by the method of equal intervals, does not have pixels from the high biomass category. This analysis is in accordance with the result presented in Figure 4.7 where the sample areas, in general, presented few cases for this category, in contradiction to the constructed categorization models.

Table 4.10 Confusion matrix for the SGC-3 test

	SGC-3			
Categorization Optimizer	Equal Intervals			
		Low	**Medium**	**High**
	Low	**21.84**	21.03	20.88
	Medium	44.15	**44.55**	44.47
	High	34.01	34.43	**34.66**

Table 4.11 Confusion matrix for the SGC-5 test

	SGC-5					
Categorization Optimizer	Equal Intervals					
		Low	**Medium-Low**	**Medium**	**Medium-High**	**High**
	Low	**90.89**	0	0	0	0
	Medium-Low	3.40	**48.57**	2.82	0	0
	Medium	5.66	51.43	**81.85**	0	18.75
	Medium-High	0.05	0	3.52	**100.00**	18.47
	High	0	0	11.81	0	**62.78**

Table 4.12 Confusion matrix for the Unini-5 test

	Unini-5					
Categorization Optimizer	Equal Intervals					
		Low	**Medium-Low**	**Medium**	**Medium-High**	**High**
	Low	**3.87**	0.01	0.01	0.01	0
	Medium-Low	0.04	**0.03**	0.05	0.02	0
	Medium	69.50	79.30	**78.98**	39.39	0
	Medium-High	25.14	18.90	19.07	**55.97**	0
	High	1.45	1.76	1.89	4.61	**0**

4.4 Conclusions

This article aimed to propose an innovative methodology at optimizing the categorization process in the construction of thematic products that guides elements linked to biomass existing in primary forests. The importance of this knowledge has a direct influence on the Sustainable Development Goals (SDGs) of the UN.

The theme referring to the estimation of AGB was approached for two study areas in the Amazon forest region and used remote sensing data and artificial intelligence techniques in an innovative way.

The results obtained show that the proposed Categorization Optimization algorithm demonstrated the ability to find new subintervals of categories that increased the Kappa agreement index. As a result, the constructed maps presented thematic accuracy superior to those obtained by classical categorization methods.

Together, the analyzes show that the adjustments made to the limits of the AGB categories did not impact the representativeness of the model for a study area, maintaining the characteristics of the region. According to Vogt et al. (2012), the labels of classes or categories should not be treated simply and automatically, but with human interference to improve the process.

From a computational perspective, the proposed heuristic enabled the identification of maximum values for the objective function in an efficient way, avoiding the high processing costs of the exhaustive search. In the extreme case, the search time decreased from 2.5 hours to 3 seconds.

In order to validate the proposed method for the construction of different types of thematic products, future tests with other databases are still needed.

In this sense, future works will present the development of heuristics that seek to analyze and compare the possibilities of maximum Kappa index found by the exhaustive search method, identifying possible advantages in selecting a specific categorization state and model.

4.5 References

Assis, F. G., Luiz F., Karine R. Ferreira, Lubia Vinhas, Luis Maurano, Claudio Almeida, Andre Carvalho, Jether Rodrigues, Adeline Maciel, and Claudinei Camargo. 2019. "TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale Thematic Mapping" *ISPRS International Journal of Geo-Information* 8, no. 11: 513. https://doi.org/10.3390/ijgi8110513

Bertin, J. (1977). La Graphique et le Traitment Graphique de l'Information. Flamarion, France.

Bueno-Crespo, Andrés et al. ‘An Unsupervised Technique to Discretize Numerical Values by Fuzzy Partitions’.Journal of Ambient Intelligence and Smart Environments 10 (2018) 289–300 289. DOI 10.3233/AIS-180488

Bussab, W. O.; Morettin, P. A. (2017), Estatística Básica. 9ª Edição. Editora Saraiva. 568 p. ISBN-10: 8547220224

Castro-Filho, C. A. P.; Bias, E. (2021), Comparison between Quantitative and Qualitative Theme-Feature Forest Biomass Estimation Models built over SAR Data . International Journal of Advanced Engineering Research and Science. In press.

Congalton, R. G.; Green, K. (1999), Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers. Denver. EUA. 180 p. ISBN 0873719867.

Costa, H., Foody, G.M., Boyd, D.S., 2017. Using mixed objects in the training of object-based image classifications. Remote Sens. Environ. 190, 188–197.

Debastiani, A.B., Moura, M.M., Rex, F.E., Sanquetta, C.R., Corte, A.P.D., Pinto, N. Regressões Robusta e Linear para Estimativa de Biomassa Via Imagem Sentinel em uma Floresta Tropical (2019). BIOFIX Science Journal, 4, 81–87. https://doi.org/10.5380/biofix.v4i2.62922

Dent, B.; Torquson, J.; Hodler, T. (2008), Cartography: Thematic Map Design. 6th ed. McGraw-Hill Science. 368 p. ISBN 0072943823.

Dietterich, T. G. (1997), Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Science, p. 1-24,

Diniz, C.G.; de Almeida Souza, A.A.; Santos, D.C.; Dias, M.C.; da Luz, N.C.; de Moraes, D.R.V.; Maia, J.S.; Gomes, A.R.; da Silva Narvaes, I.; Valeriano, D.M.; et al. DETER-B: The new Amazon near real-time deforestation detection system. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **2015**, 8, 3619–3628.

Erb, K.H., Kastner, T., Plutzar, C., Bais, A.L.S., Carvalhais, N., Fetzel, T., Gingrich, S., Haberl, H., Lauk, C., Niedertscheider, M., Pongratz, J., Thurner, M., Luyssaert, S (2018). Unexpectedly large impact of forest management and grazing on global vegetation biomass. Nature, 553, 73–76. https://doi.org/10.1038/nature25138

Diretoria de Serviço Geográfico (DSG). (2008), Infra-estrutura de dados espaciais para a Amazônia. São José dos Campos. Palestra realizada no Encontro de Usuários de Sensoriamento Remoto das Forças Armadas (SERFA), em 2008.

Foody, G. M. Impacts of ignorance on the accuracy of image classification and thematic mapping, Remote Sensing of Environment, Volume 259, 2021, 112367, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2021.112367.

Frank, E., Hall, M. A. and Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., and Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4):734–750, 2013.

Google Earth. 2021. Disponível em: <http://www.google.com.br/intl/pt-BR/earth/>. Acesso em: 22 fev. 2021.

Hastie T, Tibshirani R, Friedman J (2009) Elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, Berlin

Higuchi, N., Santos, J. dos, Ribeiro, R.J., Minette, L., Biot, Y. Biomassa da parte aérea da vegetação da Floresta Tropical úmida de terra-firme da Amazônia Brasileira. *Acta Amaz.* 1998, 28, 153–166. https://doi.org/10.1590/1809-43921998282166

Le Noë, J., Matej, S., Magerl, A., Bhan, M., Erb, K.H., Gingrich, S (2020). Modeling and empirical validation of long-term carbon sequestration in forests (France, 1850–2015). Global Change Biology, 26, 2421–2434. https://doi.org/10.1111/gcb.15004

Lima, A. J. N., Suwa, R., Ribeiro, G. H. P. M., Kajimoto, T., Santos, J., Silva, R. P., Souza, C. A. S., Barros, P. C., Noguchi, H., Ishizuka, M., Higuchi, N. (2012), Forest Ecology and Management Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro , Brazil. Forest Ecology and Management, v. 277, p. 163-172. doi: 10.1016/j.foreco.2012.04.028,

Liu, H.; Hussain, F.; Tan, C.; Dash, M. (2002), Discretization: an enabling technique. Data Mining and Knowledge Discovery 6(4), 393–423.

MAPBIOMAS. Disponível em: <http://mapbiomas.org/map#coverage>. Acessado em: 18 de agosto de 2021.

Martins-Bedê, F. T.; Freitas, C. D. C.; Dutra, L. V., et al. (2009), Risk Mapping of Schistosomiasis in Minas Socioeconomic Spatial Data. IEEE Transactions on Geoscience and Remote Sensing, v. 47, n. 11, p. 3899-3908.

Maslove, D.M., Podchiyska, T., and Lowe, H. J. Discretization of continuous features in clinical datasets, Journal of the American Medical Informatics Association 20(3) (2013), 544–553. doi:10.1136/amiajnl-2012-000929.

Mitchell, P. J., Downie, A. L., Diesing, M. How good is my map? A tool for semi-automated thematic mapping and spatially explicit confidence assessment. Environmental Modelling & Software, Volume 108, 2018, Pages 111-122, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2018.07.014.

Pereira, L.O., Furtado, L.F.A., Novo, E.M.L.M., Sant'Anna, S.J.S., Liesenberg, V., Silva, T.S.F (2018). Multifrequency and Full-Polarimetric SAR assessment for estimating above ground biomass and leaf area index in the Amazon Várzea Wetlands. Remote Sensing, 10, 1–23. https://doi.org/10.3390/rs10091355

Prodes, P. Monitoramento da floresta Amazônica Brasileira por satélite. Inst. Nac. De Pesqui. Espac. Proj. Prodes. 2013, 25, 2013.

Projeto RadamBrasil. 1977. Geologia, geomorfologia, pedologia, vegetação e uso potencial da terra. Rio de Janeiro, Departamento Nacional da Produção Mineral.

Quinlan, J.R**. (**1993), C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Rajbahadur, G. K., Wang, S., Kamei, Y., Hassan, A. E. "Impact of Discretization Noise of the Dependent Variable on Machine Learning Classifiers in Software Engineering," in *IEEE Transactions on Software Engineering*, vol. 47, no. 7, pp. 1414-1430, 1 July 2021, doi: 10.1109/TSE.2019.2924371.

Raposo, P., Touya G., Bereuter P. A Change of Theme: The Role of Generalization in Thematic Mapping. *ISPRS International Journal of Geo-Information*. 2020; 9(6):371. https://doi.org/10.3390/ijgi9060371

Rosenfeld, A., Illuz, R., Gottesman, D., Last, M. Using discretization for extending the set of predictive features. *EURASIP J. Adv. Signal Process.* 2018, 7 (2018). https://doi.org/10.1186/s13634-018-0528-x

Russel, S.; Norvig, P. (2020). Artificial Intelligence: A Modern Approach Edition, Pearson, 1115 páginas , ISBN-10: 0134610997

Sarker, M. L. R., Nichol, J., Iz, H. B., Ahmad, B. B., Rahman (2012), A. A. Potential of texture measurements of two-date dual polarization PALSAR data for the improvement of forest biomass estimation. ISPRS Journal of Photogrammetry and Remote Sensing, 69, 146-166. https://doi.org/10.1016/j.isprsjprs.2012.03.002.

Scipal, K., Arcioni, M., Fois, F., Lin, C-C., Chave, J., Dall, J., LeToan, T., Papathanassiou, K., Quegan, S., Rocca, F., Saatchi, S., Shugart, H., Ulander, L., & Williams, M. (2010). The BIOMASS Mission - An ESA Earth Explorer candidate to measure the BIOMASS of the Earth's forests. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 52-55) https://doi.org/10.1109/IGARSS.2010.5648979

Silva, R.P., 2007. Alometria, estoque e dinâmica da biomassa de florestas primárias e secundárias na região de Manaus (AM). National Institute for Space Research (INPE). PhD Thesis.

Sluter, C. R., Camboim, S. P., Iescheck, A. L., Pereira, L. B., Castro, M. C., Yamada, M. M., Araújo, V. S. (2018) A Proposal for Topographic Map Symbols for Large-Scale Maps of Urban Areas in Brazil, The Cartographic Journal, 55:4, 362-377, DOI: 10.1080/00087041.2018.1549307

Theodoridis, S.; Koutroumbas, K. (2008), Pattern Recognition. 4th ed. Academic Press. 961 p. ISBN 1597492728 .

Vogt, L., Grobe, P., Quast, B., Bartolomaeus, T., 2012. Fiat or bona fide boundary—a matter of granular perspective. PLoS One 7 (12), e48603.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). Data Mining: Practical Machine Learning Tools and Techniques, 2nd Editio. ed, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Massachusetts. https://doi.org/10.1016/c2009-0-19715-5

Woodhouse, I.H., 2017. Introduction to Microwave Remote Sensing, Introduction to Microwave Remote Sensing. Taylor & Francis Group CRC Press, Florida. https://doi.org/10.1201/9781315272573

Wu, T., Dong, W., Luo, J., Sun, Y., Huang, Q., Wu, W., Hu, X. Geo-parcel-based geographical thematic mapping using C5.0 decision tree: a case study of evaluating sugarcane planting suitability. Earth Sci Inform **12,** 57–70 (2019). https://doi.org/10.1007/s12145-018-0360-8

Yang, Y.; Webb, G. I. (2009), Discretization for naive-Bayes learning : managing discretization bias and variance. Machine Learning, p. 39-74. doi: 10.1007/s10994-008-5083-5.

5 CONSIDERAÇÕES E CONCLUSÕES FINAIS

5.1 Contextualização e Contribuição da Pesquisa

A presente tese buscou apresentar aplicações inovadoras de técnicas de aprendizado de máquina no desenvolvimento de metodologias que visam aprimorar os modelos de estimativa de biomassa arbórea.

No primeiro artigo houve destaque no detalhamento dos dados de AGB fornecidos pelo INPA, com apresentação das características do trabalho de manejo florestal realizado em campo. Da mesma forma, foram especificados os tipos de modelos digitais, as respectivas características e a importância na modelagem da biomassa.

As técnicas de ML foram aplicadas de forma inovadora, apresentando uma metodologia que buscou aumentar a precisão altimétrica dos dados obtidos pelo sensor de SAR e o consequente aprimoramento do modelo de estimativa de biomassa florestal.

Já, o segundo artigo, de forma inovadora, apresenta uma metodologia que envolve:

- o processo de seleção de atributos e desenvolvimento de modelos de estimativa de AGB, simultaneamente, sobre dados temáticos numéricos e categóricos; e
- as análises comparativas entre resultados numéricos e categóricos, incluindo a construção de matrizes de confusão e a respectiva comparação por meio de testes de hipóteses.

Cabe ressaltar que, para cada modelo desenvolvido no segundo artigo, a seleção de atributos e as técnicas de ML foram específicas e configuradas de forma a obter os melhores resultados.

Para finalizar, no terceiro artigo foi apresentado um sistema inovador que utiliza técnicas de ML para otimizar o processo de categorização na construção de produtos temáticos.

Os resultados obtidos no terceiro artigo mostram que o algoritmo de Otimização de Categorização proposto demonstrou capacidade de encontrar novos subintervalos de categorias que aumentaram o índice de concordância Kappa. Como resultado, foram construídos produtos temáticos que apresentaram acurácia temática superior aos obtidos pelos métodos clássicos de categorização. Juntamente, do ponto de vista computacional, a heurística proposta no algoritmo possibilitou a identificação de resultados de forma eficiente, evitando os altos custos de processamento.

De forma geral, ao analisar os artigos que compõem a tese, todos apresentaram contribuições inovadoras no sentido de desenvolver modelos de estimativa de AGB com o uso de técnicas de ML. Contribuições envolvendo técnicas de ML:

- inicialmente são apresentadas no sentido de aprimorar os dados a serem utilizados na modelagem de AGB, no primeiro artigo;
- na sequência buscam desenvolver modelos mais precisos e específicos de estimativa de AGB, no segundo artigo; e
- finalmente são apresentadas na etapa de construção de produto temático de AGB, no terceiro artigo.

5.2 Revisitando os Objetivos

Os objetivos específicos propostos foram divididos em três artigos técnicos independentes, embora sequenciais e componentes da metodologia adotada na tese.

O primeiro trabalho foca no objetivo específico da tese em analisar técnicas de ajuste da altura interferométrica para o desenvolvimento de modelo de estimativa de AGB. Neste sentido, aplica técnicas de ML sobre modelos matemáticos que visam aprimorar a H_{int} como atributo independente e diretamente relacionado à AGB.

O segundo artigo teve importância central na presente tese, com foco nas etapas de extração de atributos de SAR e na modelagem para estimativa de AGB. Abordou os seguintes objetivos específicos da tese:

- análise das técnicas não-paramétricas de aprendizado de máquinas, aplicando-as na construção de modelos preditivos para estimativa de biomassa arbórea a partir de variáveis extraídas de dados de SAR;
- implementação, análise e proposta de modelos preditivos para estimativa de biomassa arbórea utilizando métodos estatísticos; e
- avaliação dos modelos preditivos construídos visando apontar o mais adequado aos dados e região em trabalho.

O segundo artigo conclui que as diferentes regiões da Floresta Amazônica e suas respectivas características demandam modelos e técnicas específicas, não se enquadrando em um único padrão. Neste caso não foi possível identificar uma única técnica de ML que se mostrasse como a mais adequada ao objetivo, apesar dos melhores resultados apontarem para o uso de *Multilayer Perceptron* (MLP) e de árvores de decisão univariadas (OBCT).

O terceiro artigo teve por objetivo encerrar o fluxo metodológico proposto na tese e atender ao objetivo específico referente à implementação, analise e proposta de metodologia para construção de produto temático que englobe as representações cartográficas referentes às categorias de biomassa e suas características.

Ao término da presente tese todos os objetivos específicos foram atendidos por meio da pesquisa de trabalhos que envolvem o "estado da arte" e na elaboração e publicação de artigos técnicos em revistas especializadas. Desta forma, o objetivo geral da tese também foi atendido, buscando desenvolver e aplicar uma metodologia para estimar a AGB, a partir de dados de SAR, utilizando técnicas de ML.

5.3. Revisitando as Hipóteses

O primeiro artigo conclui que o ajuste da altura interferométrica não obteve o resultado esperado, não sendo possível melhorar significativamente a estimativa de biomassa florestal por meio dos modelos matemáticos aplicados. Este resultado não invalida a hipótese do respectivo artigo, isto é, de que é possível o ajuste da altura interferométrica utilizando áreas de solo exposto. Porém, sugere-se outros estudos sobre a mesma metodologia, tendo como base uma quantidade maior de áreas de solo exposto identificadas *in loco*.

No segundo artigo não foi comprovada a hipótese de que um único modelo de AGB se adequasse de forma genérica às áreas de estudo. Para cada análise realizada, baseada em dados quantitativos ou qualitativos, inclusive as submetidas ao processo de seleção de atributos, diferentes técnicas de ML se sobressaíram.

Do resultado obtido no segundo artigo, optou-se no terceiro artigo pelo uso da técnica de ML de árvore de decisão univariada (OBCT) para a construção do respectivo produto temático, em função da possibilidade intuitiva na análise dos resultados. Neste caso a hipótese é de que os parâmetros utilizados no processo de categorização possam ser ajustados, por meio de técnicas de ML, de tal modo que o produto temático final terá maior acurácia, isto é, que o processo de categorização não seja uma fonte de erro para a representação temática.

Os resultados obtidos no terceiro artigo mostram que o sistema de busca apresentado identificou intervalos de categorias que maximizaram o índice de concordância Kappa do produto temático, utilizando para isto técnicas de ML, confirmando a hipótese levantada.

5.4. Conclusões Finais

O presente trabalho apresentou três artigos sequenciais, todos com inovações técnicas, que buscaram atender aos respectivos objetivos específicos e geral da tese e aplicar técnicas de ML em diferentes etapas de construção de um produto temático de estimativa de biomassa para as regiões florestais amazônicas das áreas de estudo.

A hipótese proposta na tese, isto é, de que a aplicação de técnicas de aprendizado de máquina sobre dados de SAR permitem obter a estimativa de biomassa da região amazônica com erros abaixo de 20%, atendendo os padrões preceituados por organismos internacionais, não foi confirmada. Os resultados obtidos nos modelos elaborados são classificados somente como *moderados*. Alguns fatores podem ter contribuído para este resultado, incluindo a:

- quantidade reduzida de amostras de biomassa, além da pequena variação de valores, o que prejudicou o ajuste dos modelos gerados;
- o acesso restrito aos dados de SAR das bandas X e P, não sendo possível a obtenção dos arquivos SLC capazes de gerar novos atributos coerentes.

5.5. Aplicações e Oportunidades de Estudos Futuros

Visando dar continuidade às pequisas de estimativa de AGB na região amazônica, sugere-se o seguinte:

- promoção de novos projetos de manejo florestal na região amazônica que envolvam a medição de parcelas de biomassa em diferentes categorias, isto é, com maior variação de valores, incluindo áreas de floresta secundária;
- o desenvolvimento de modelos de estimativa de AGB a partir de dados InSAR e PolSAR oriundos de outras fontes e/ou outros comprimentos de onda. Neste sentido, novos atributos poderão ser extraídos a partir das bandas X, C, L e P, orbitais ou aerotransportadas; e
- o desenvolvimento de sistema que aplique métodos de busca na definição de parâmetros para as técnicas de ML, visando a construção de modelos regionais de estimativa de biomassa.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, A. F.; BOTELHO, M. F.; CENTENO, J. A. S. Classificação de imagens de alta resolução integrando variáveis espectrais e forma utilizando redes neurais artificiais. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 11. (SBSR), 2003, Belo Horizonte. **Anais...** São José dos Campos: INPE, 2003. p. 265-272. CD-ROM, Online. ISBN 85-17-00017-X. Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2002/11.14.14.55>. Acesso em: 13 jan. 2011.

ARAÚJO, T. M.; HIGUCHI, N.; CARVALHO JR., J. A. Comparison of formulae for biomass content determination in a tropical rain forest in the state of Pará, Brazil. **Forest Ecology and Management**, v. 117, p.43-52, 1999.

ASKNE, J.; SANTORO, M.; SMITH, G.; FRANSSON, J. Multitemporal repeat pass SAR interferometry of boreal forests. **IEEE Transactions on Geoscience and Remote Sensing**, v. 41, n. 7, p. 1540–1550, Jul. 2003.

ATKINSON P. M.; TATNALL A. R. L. Introduction Neural networks in remote sensing. **International Journal of Remote Sensing**, v. 18, n. 4, p. 699-709, 1997.

BALL, J. E.; ANDERSON, D. T.; CHAN, C. S. Comprehensive survey of deep learning in remote ensing: theories, tools, and challenges for the community. **Journal of Applied Remote Sensing,** v. 11, n. 4, p. 1, 2017.

BEAUDOIN, A.; LE TOAN, T.; GOZE, S.; NEZRY, E.; LOPES, A.; MOUGIN, E.; HSU, C. C.; HAN, H. C.; KONG, J. A.; SHIN. R. T. Retrieval of forest biomass from SAR data. **International Journal of Remote Sensing**, v. 15, n. 14, p. 2777-2796, 1994.

BELLASSEN, V.; DELBART, N.; LE MAIRE, G.; LUYSSAERT, S.; CIAIS, P.; VIOVY, N. Potential knowledge gain in large-scale simulations of the forest carbon fluxes from remotely sensed biomass and height. **Forest ecology and management**, v. 261, p. 515-530, 2011.

BISCHOF, H.; SCHNEIDER, W.; PINZ, A. J. Multispectral classification of Landsat-images using neural networks, **Transactions on Geoscience and Remote Sensing**, v. 30, n. 3, p. 482-490, May 1992.

BISHOP, C. M. **Neural networks for pattern recognition**. Clarendon. Oxford, England. 1995. 482 p. ISBN 0198538642.

BOERNER, W. M.; YAN, W. L.; XI, A. Q.; YAMAGUCHI, Y. On the basic principles of radar polarimetry: the target characteristic polarization state theory of Kennaugh, Huynen´s polarization fork concept, and its extension to the partially polarized case. **Proccedings of the IEEE**, v. 79, n.10, p. 1538-1550, 1991.

BORBA, P; BIAS, E. S.; SILVA, N. C.; ROIG, H. L. A Review of Remote Sensing Applications on Very High-Resolution Imagery Using Deep Learning-Based Semantic Segmentation Techniques, **International Journal of Advanced Engineering Research and Science**, v. 8, n. 8, 2021.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BROWN, S.; GILLESPIE, A. J. R.; LUGO, A. E. Biomass estimation methods for tropical forest with applications to forest inventory data. **Forest Science**, v.35, n.4, p.881-902, 1989.

BRUZZONE, L.; PERSELLO, C. A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples. **IEEE Transactions on Geoscience and Remote Sensing**, v. 47, n. 7, p. 2142–2154, 2009.

BUSSAB, W. O.; MORETTIN, P. A.. 2002. **Estatística Básica**. 5ª Edição. Editora Saraiva. 526 p. CDD -519.5.

CARPENTIER, J.; ABADIE, J. Généralisation de la Méthode du Gradient Réduit de Wolfe au cas des Contraintes Non Lineaires IV International Conference on Operational Research. Anais... In: **Proceedings of**... Operations Research Society of America. New York: D. B. Herts and J. Melese, 1966.

CASTRO FILHO, C. A. P. Aplicação de árvores de decisão na classificação de uso e cobertura da terra sobre imagens LANDSAT TM e PolInSAR. São José dos Campos, 2010.

CORTES, C.; VAPNIK, V. Support vector machine. **Machine learning**, v. 20, n. 3, p. 273–297, 1995.

SERFA 2010 – Encontro de Usuários de Sensoriamento Remoto das Forças Armadas, em 24 nov. 2010.

CASTRO-FILHO, C. A. P.; SANTOS, J. R. 2010. Classificação de Imagens PolInSAR Utilizando Técnicas de Mineração de Dados. In: **Proceedings of** ... IX Seminário de Atualização em Sensoriamento Remoto e Sistemas de Informações Geográficas Aplicados à Engenharia Florestal, Curitiba. IX Seminário de Atualização em Sensoriamento Remoto e Sistemas de Informações Geográficas Aplicados à Engenharia Florestal, 2010.

CHAMBERS, J. Q.; SANTOS, J.; RIBEIRO, R. J.; HIGUCHI, N. Tree damage, allometric relationships, and above-ground net primary production in central Amazon forest. **Forest Ecology and Management**, v. 152, p. 73– 84, 2001.

CHAVE, J., RÉJOU-MÉCHAIN, M., BÚRQUEZ, A., CHIDUMAYO, E., COLGAN, M. S., DELITTI, W. B., DUQUE, A., EID, T., FEARNSIDE, P. M., GOODMAN, R. C., HENRY, M., MARTÍNEZ-YRÍZAR, A., MUGASHA, W. A., MULLER-LANDAU, H. C., MENCUCCINI, M., NELSON, B. W., NGOMANDA, A., NOGUEIRA, E. M., ORTIZ-MALAVASSI, E., PÉLISSIER, R., PLOTON, P., RYAN, C. M., SALDARRIAGA, J. G. & VIEILLEDENT, G. Improved allometric models to estimate the aboveground biomass of tropical trees. **Global Change Biology**, 20, 10, p. 3177-3190, 2014. doi: 10.1111/gcb.12629.

CLOUDE, S. R.; POTTIER, E. A review of target decomposition theorem in radar polarimetry. In: International Geoscience and Remote Sensing Symposium (IGARSS), 1996. **Proceedings …** v. 34, n. 2, p. 498-518. 1996.

CIRILO, J. A., **Programação Não Linear Aplicada a Recursos Hídricos**. In: PORTO, R. L. L. et al., Técnicas Quantitativas para o Gerenciamento de Recursos Hídricos . ABRH,1ª edição, pp. 305-356, Editora da Universidade – UFRGS, 1997.

COLLINS, J. N.; HUTLEY, L. B.; WILLIAMS, R. J.; BOGGS, G.; BELL, D.; BARTOLO, R. Estimating landscape-scale vegetation carbon stocks using airborne multi-frequency polarimetric synthetic aperture radar (SAR) in the savannahs of north Australia. **International Journal of Remote Sensing** , v. 30, n. 5, p. 1141 – 1159, 2009.

CONGALTON R. G.; GREEN, K. 1999. **Assessing the Accuracy of Remotely Sensed Data: Principles and Practices**. Lewis Publishers. Denver. EUA. 180 p. ISBN 0873719867.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent Data Analysis**, v. 1, n. 3, p. 131-156, 1997.

DEL FRATE, F.; SOLIMINI, D. On neural network algorithms for retrieving forest biomass from SAR data. **IEEE Transaction on Geoscience and Remote Sensing**, v. 42, n. 1, p. 24–34, Jan. 2004.

DENT, B.; TORQUSON, J.; HODLER, T. 2008. **Cartography: Thematic Map Design**. 6th ed. McGraw-Hill Science. 368 p. ISBN 0072943823.

DIETTERICH, T. G. 1997. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. **Science**, p. 1-24.

DIETTERICH T. G. Ensemble learning. **The handbook of brain theory and neural networks**, v. 2, p.110–125, 2002.

DIRETORIA DE SERVIÇO GEOGRÁFICO (DSG). **Projeto Básico: Contratação de Serviços de Aerolevantamento na Região Amazônica e Processamento de Dados com Radares de Abertura Sintética Aerotransportados Interferométricos**. 2008.

DOBSON, C.; ULABY, F.; PIERCE, L.; SHARIK, T.; BERGEN, K.; KELLNDORFER, J.; KENDRA, J. R.; LI, E.; LIN, Y. C.; NASHASHIBI, A.; SARABANDI, K.; SIQUEIRA, Q. Estimation of forest biomass characteristics in northern Michigan with SIR-C/X-SAR data. **IEEE Transaction on Geoscience and Remote Sensing**, v. 33, n. 4, p. 877–894, July 1995.

DUBOIS-FERNANDEZ, P.; ORIOT, H.; COULOMBEIX, C., et al. TROPISAR : EXPLORING THE TEMPORAL BEHAVIOR OF P-BAND SAR DATA. **IGARSS 2010 - 2010 IEEE International Geoscience and Remote Sensing Symposium**, v. 6, n. 1, p. 1319-1322, 2010.

DUDA, R.O.; HART, P.E.; STORK, D.G. **Pattern Classification.** 2nd ed. John Wiley & Sons. NY. 2001. 513p.

DUTRA, L. V.; HUBER, R. Feature extraction and selection for ERS-1/2 InSAR classification. **International Journal of Remote Sensing**, v. 20, n. 5, pags. 993-1016, 1999.

DUTRA, L.V.; ELMIRO, M.T.; SOARES, B.S.; MURA, J.C.; SANTOS, J.R.; FREITAS, C.C.; ARAÚJO, L.S.; ALBUQUERQUE, P.C.G.; VIEIRA, P.R.; GAMA, F.F. Assessment of digital elevation models obtained in Brazilian Amazon based on P and X band airborne interferometric data. In: International Geoscience and Remote Sensing Symposium (IGARSS), 2002, Toronto. **Proceedings…** Toronto: IEEE, Jun. 2002. 1 CD-ROM.

DUTRA, L.V.; ELMIRO, M.T.; SOARES, B.S.; MURA, J.C.; SANTOS, J.R.; FREITAS, C.C.; ARAÚJO, L.S.; ALBUQUERQUE, P.C.G.; VIEIRA, P.R.; GAMA, F.F. Assessment of digital elevation models obtained in Brazilian Amazon based on P and X band airborne interferometric data. In: International Geoscience and Remote Sensing Symposium (IGARSS), 2002, Toronto. **Proceedings…** Toronto: IEEE, Jun. 2002. 1 CD-ROM.

European Space Agency (ESA). **BIOMASS**. 2021. Disponível em: <https://www.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers/Biomass >. Acesso em: 02 jun 2021.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, ALMEIDA, Tiago Agostinho de; André Carlos Ponce de Leon Ferreira de. Inteligência artificial: uma abordagem de aprendizado de máquina, 2ª Edição, 2021, 304 pag, LTC, ISBN-13 : 978-8521637349

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. American Association for Artificial Intelligence Magazine, p. 37 – 54, 1996. Disponível em: <http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf>. Acesso em 21 Ago 2017.

FOODY, G.; BOYD, D. S.; CUTLER, M. E. J. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. **Remote Sensing of Environment**, v. 85, n. 4, p. 463-474. doi: 10.1016/S0034-4257(03)00039-7, 2003.

FREEMAN, A.; DURDEN, S. L. A three-component scattering model for polarimatric SAR data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 36, n. 3, p. 963-973, 1998.

FRIEDL, M. A.; BRODLEY, C. E. Decision tree classification of land cover from remotely sensed data. **Remote Sensing of Environment**, v. 61, p.399-409, 1997.

GABOARDI, C. **Utilização de imagem de coerência SAR para classificação do uso da terra**: Floresta Nacional do Tapajós. 2002. 139 p. (INPE-9612-TDI/842). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2002. Disponível em: <http://urlib.net/sid.inpe.br/marciana/2003/04.10.08.52>. Acesso em: 10 fev. 2010.

GAMA, F. F. **Estudo da interferometria e polarimetria SAR em povoamentos florestais de eucalyptus SP**. 2007. 243 p. (INPE-14778-TDI/1231). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2007. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17@80/2007/04.04.12.36>. Acesso em: 12 jan. 2018.

GE, Y.; HEXIANG, B.; LI, S. **Geo-spatial data analysis, quality assessment and visualization**. ICCSA, Part I, LNCS 5072, p. 258-267, 2008. DOI: 10.1007/978-3-540-69839-5_20.

GHASEMI, N.; SAHEBI, M. R.; MOHAMMADZADEH, A. A review on biomass estimation methods using synthetic aperture radar. , INTERNATIONAL JOURNAL OFGEOMATICS AND GEOSCIENCES v. 1, n. 4, p. 776-788, 2011.

GOLDBERG, D. E. **Genetic algorithms in search, optimization, and machine learning**. New York: Addison-Wesley, 1989. 412 p. ISBSN 0-201-15767-5.

GONÇALVES, F. G. **Avaliação de dados SAR polarimétricos para estimativa volumétrica de florestas tropicais**. 2007. 110 p. (INPE-14777-TDI/1230). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2007. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17@80/2007/05.07.18.38>. Acesso em: 24 jan. 2018.

HAJNSEK, I.; KUGLER, F.; LEE, S. K.; PAPATHANASSIOU, K. P. Tropical forest parameter estimation by means of Pol-InSAR: The INDREX-II campaign. **IEEE Transaction on Geoscience Remote Sensing**, v. 47, n. 2, p. 481-493, 2009.

HENDERSON, F. M.; LEWIS, A. J. **Manual of remote sensing: principles and applications of imaging radars**. 3rd ed. John Wiley and Sons. New York. 1998. 896p. ISBN 0471330469.

HIGUCHI, N.; SANTOS, J.; RIBEIRO, R. J.; MINETTE, L.; BIOT, Y. Biomassa da parte aérea da vegetação da floresta tropical úmida de terra-firme da amazônia brasileira. **ACTA Amazônica**, v.28, n.2, p.153-166, 1998.

IMHOFF, M.L. Radar backscatter and biomass saturation: Ramifications for a global biomass inventory. **IEEE Transaction on Geoscience Remote Sensing**, v. 33, n. 2, p. 511–518, 1995.

Intergovernmental Panel on Climate Change (IPCC). **Good practice guidance for land use, land-use changes and forestry**. Institute for Global Environmental Strategies. Kanagawa, Japão. 2003.

Japan Aerospace Exploration Agency (JAXA). **ALOS - PALSAR**. 2021. Disponível em: <https://www.eorc.jaxa.jp/ALOS/en/about/palsar.htm>. Acesso em: 02 jun 2021.

JIN, Y.-Q.; LIU, C. Biomass retrieval from high-dimensional active/passive remote sensing data by using artificial neural networks. **International Journal of Remote Sensing**, v. 18, n. 4, p. 971-979. doi: 10.1080/014311697218863, 1997.

KASISCHKE, E.S.; MELACK, J.M.; DOBSON, M.C. The use of imaging radars for ecological applications. **Remote Sensing of Environment**, v. 59, p. 141–156, 1997.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. **Proceedings of the Ninth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, p. 249–256.

KOHAVI, R.; JOHN, G. Wrapper for feature subset selection. **Artificial Intelligence**. v. 97, p. 234–273, 1997.

KUPLICH, T.M.; CURRAN, P.J.; ATKINSON, P.M. Relating SAR image texture to the biomass of regenerating tropical forests. **International Journal of Remote Sensing**, v. 26, p. 4829–4854, 2005.

KUPLICH, T. M. Classifying regenerating forest stages in Amazônia using remotely sensed images and a neural network. **Forest Ecology and Management**, v. 234, p. 1-9, 2006.

LASDON, L. S.; WAREN, A. D.; JAIN, A.; RATNER, M. Design and testing of a generalized reduced gradient code for nolinear programming. **ACM Transactions on Mathematical Sofware**, New York, v. 4, n. 1, p. 34-50, 1978.

LIMA, A. J. N.; TEIXEIRA, L. M.; CARNEIRO, V. M. C.; PINTO, A. C. M.; PINTO, F. R.; SANTOS, J.; HIGUCHI, N. Inventário florestal contínuo em áreas manejadas e não manejadas do estado do amazonas. In: 5º Congresso Florestal Nacional, Portugal, 2005. **Anais ...** Disponível em < http://www.esac.pt/cernas/cfn5/docs/T2-43.pdf>. Acesso em: 24 jan. 2011.

LIMA, A. J. N.; SUWA, R.; HENRIQUE, G., et al. 2012. Forest Ecology and Management Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro, Brazil. **Forest Ecology and Management**, v. 277, p. 163-172. doi: 10.1016/j.foreco.2012.04.028.

LIU, H.; HUSSAIN, F.; TAN, C.; DASH, M. 2002. Discretization: an enabling technique. **Data Mining and Knowledge Discovery** 6(4), 393–423.

LOGAN, T.; RITTER, N.; BRYANT, N. Artificial neural network classification using a minimal; training set: comparison to conventional supervised classification. **Photogrammetric Engineering & Remote Sensing**, v. 56, n. 4, p. 1285-1294, 1997.

MA, L; LIU, Y.; ZHANG, X.; YE, Y.; YIN, G.; JOHNSON, B. A. Deep learning in remote sensing applications: A meta-analysis and review, **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 152, p. 166-177, 2019.

MALHI, G.; BALDOCCHI, D.D.; JARVIS, P.G. The carbon balance of tropical, temperate and boreal forests. **Plant Cell Environment**, v. 22, p. 715–740, 1999.

MAPBIOMAS. Disponível em: <http://mapbiomas.org/map#coverage>. Acessado em: 14 de junho de 2021.

MARTÍNEZ, J. M. e SANTOS, S. A. **Métodos computacionais de otimização**. 1998. Disponível em: <www.ime.unicamp.br/~martinez/mslivro.pdf>. Acessado em: 12 de julho de 2014.

MARTINEZ-ALVAREZ, F.; BUI, D. T. Advanced Machine Learning and Big Data Analytics in Remote Sensing for Natural Hazards Management. **Multidisciplinary Digital Publishing Institute**, 2020.

MARTINS-BEDÊ, F. T.; FREITAS, C. D. C.; DUTRA, L. V., et al. 2009. Risk Mapping of Schistosomiasis in Minas Socioeconomic Spatial Data. **IEEE Transactions on Geoscience and Remote Sensing,** v. 47, n. 11, p. 3899-3908.

Ministério da Ciência e Tecnologia (MCT). **Relatório de Emissão de Gases de Efeito Estufa: Crescem emissões de gases estufa**. 2010. Disponível em: <http://www.mct.gov.br/index.php/content/view/326777.html>. Acesso em: 20 jan. 2018.

MICHALSKI, R. S.; CARBONEL, J.; MITCHELL, T. 1983. **Machine Learning: An Artificial Intelligence Approach**. TIOGA Publishing Co., Palo Alto, California.

MOUNTRAKIS, G; IM, J.; OGOLE, C. Support Vector Machines in Remote Sensing: A review. **ISPRS Journal of Photogrammetry and Remote Sensing Society,** v. 66, n. 3, p. 247–259, 2011.

MUUKKONEN, P.; HEISKANEN, J. Estimating biomass for boreal forests using ASTER satellite data combined with standwise forest inventory data. **Remote Sensing of Environment**, v. 99, n. 4, p. 434-447. doi: 10.1016/j.rse.2005.09.011, 2005.

MURA, J. C. **Geocodificação automática de imagens de radar de abertura sintética interferométrico**: sistema Geo-InSAR. 2000. 159 p. (INPE-8209-TDI/764). Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2000. Disponível em: <http://urlib.net/sid.inpe.br/deise/2001/08.03.12.24>. Acesso em: 10 dez. 2010.

NARVAES, I. S. **Avaliação de dados SAR polarimétricos para estimativa de biomassa em diferentes fitofisionomias de florestas tropicais**. 2010. 164 p. (INPE-T/). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

NEEFF, T.; DUTRA, L. V.; SANTOS, J. R.; FREITAS, C. C.; ARAUJO, L. S. Tropical forest stand table modelling from SAR data. **Forest Ecology and Management**, ELSEVIER, v.186, p. 159-170, 2003.

NEEFF, T.; DUTRA, L. V.; SANTOS, J. R.; FREITAS, C. C.; ARAÚJO, L. S. Tropical forest biomass measurement by interferometric height modeling and P-band radar backscatter. **Forest Science**, v. 51, n. 6, p. 585–594, Dec. 2005.

NELSON, B.W.; MESQUITA, R.; PEREIRA, J. L. G.; SOUZA, S. G. A.; BATISTA, J. T.; COUTO, L. B. Allometric regressions for improved estimate of secondary forest biomass in the central Amazon. **Forest Ecology and Management**, v.117, p. 149 –167, 1999.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. **Applied Linear Statistical Models**. 4th ed. Boston: McGraw-Hill, 1996. 1408 p.

NI, W.; GUO, Z.; SUN, G.; CHI, H. Investigation of forest height retrieval using SRTM-DEM and ASTER-GDEM. In: International Geoscience And Remote Sensing Symposium (IGARSS), 2010. **Proceedings ...** 2010, p. 2111-2114.

NOVACK, T.; KUX, H. J. H. Classificação da cobertura do solo urbano inserindo árvores de decisão a rede hierárquica. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14. (SBSR), 2009, Natal. **Anais...** São José dos Campos: INPE, 2009. p. 7871-7876. DVD, On-line. ISBN 978-85-17-00044-7. (INPE-15960-PRE/10569). Disponível em: <http://urlib.net/dpi.inpe.br/sbsr@80/2008/11.17.17.11>. Acesso em: 26 jul. 2010.

PAOLA, J.D., SCHOWENGERDT, R.A. A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. **International Journal of Remote Sensing**, v. 16, p. 3033–3058, 1995.

POLSARPRO 4.03. European Space Agency & ESRIN, 2009. 1 CD ROM.

POPE, K. O.; BENAYAS-REY, J. M.; PARIS, J. F. Radar remote sensing of forest and wetland ecosystems in the Central American tropics. **Remote Sensing of Environment**, v. 48, n. 2, p.205-219. 1994.

PORTAL MUNICIPAL DA PREFEITURA MUNICIPAL DE SÃO GABRIEL DA CACHOEIRA. Disponível em: <http://www.saogabrieldacachoeira.am.gov.br/portal/>. Acesso em: 17 jan. 2018.

PROJETO PRODES: Monitoramento da Floresta Amazônica Brasileira por Satélite. **Ano 2017-2018**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE). Disponível em: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>. Acesso em: 1 jun. 2019.

PROJETO RADAMBRASIL. **Folha NA.19 Pico da Neblina/Amazonas**: geologia, geomorfologia, pedologia, vegetação e uso potencial da terra. Rio de Janeiro: Departamento Nacional da Produção Mineral, 1976. v. 11. 380 p. CDD 558.1.

PROJETO RADAMBRASIL. **Folha SA.19 Içá/Amazonas**: geologia, geomorfologia, pedologia, vegetação e uso potencial da terra. Rio de Janeiro: Departamento Nacional da Produção Mineral, 1977. v. 14. 452 p. CDD 558.1.

QUINLAN, J. R. **C4.5: Programs for machine learning**. Morgan Kaufmann. California. 1993. 235p.

RUSSEL, S.; NORVIG, P. 2020. **Artificial Intelligence: A Modern Approach**, Pearson, 1115 p., ISBN-10: 0134610997.

SAATCHI, S.; MOGHADDAM, M. Estimation of crown and stem water content and biomass of boreal forest using polarimetric SAR imagery. **IEEE Transactions on Geoscience and Remote Sensing**, v. 38, n. 2, p. 697–709, Mar. 2000.

SAATCHI, S.; HALLIGAN, K.; DESPAIN, D. G.; CRABTREE, R. L. Estimation of forest fuel load from radar remote sensing. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, n. 6, p. 1726 – 1740, 2007a.

SAATCHI, S. S.; HOUGHTON, R. A; DOS SANTOS ALVALÁ, R. C.; SOARES, J. V.; YU, Y. Distribution of aboveground live biomass in the Amazon basin. **Global Change Biology**, v. 13, n. 4, p. 816-837. doi: 10.1111/j.1365-2486.2007.01323.x, 2007b.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v. 23, n. 19, p. 2507-2517, 2007.

SAMUI, P.; ROY, S. S.; BALAS, V. E. 2017. **Handbook of Neural Computation 1st Edition**. Elsevier. ISBN 9780128113196.

SANTOS, J.R.; ARAÚJO, L.S.; DUTRA, L.V.; FREITAS, C.C.; MURA, J.C.; GAMA, F.F. Estimation of basal area from Amazon tropical rain forest using airborne P-band SAR data. [CD-ROM] In: IGARSS-International Geoscience and Remote Sensing Symposium, 2002, Toronto. **Proceedings…** IEEE, Jun. 2002.

SANTOS, J. R.; FREITAS, C. C.; ARAÚJO, L. S.; DUTRA, L. V.; MURA, J. C.; GAMA, F. F.; SOLER, L. S.; SANT'ANNA, S. J. S. Airborne P-band SAR applied to the aboveground biomass studies in the Brazilian tropical rainforest. **Remote Sensing of Environment**, ELSEVIER, v. 87, p. 482-493, 2003.

SCIPAL, K.; ARCIONI, M.; CHAVE, J.; DALL, J.; FOIS, F.; LETOAN, T.; LIN, C-C.; PAPATHANASSIOU, K.; QUEGAN, S.; ROCCA, F.; SAATCHI, S.; SHUGART, H.; ULANDER, L.; WILLIAMS, M. The biomass mission – an ESA earth explorer candidate to measure the biomass of the earth´s forests. In: International Geoscience and Remote Sensing Symposium (IGARSS), 2010. **Proceedings ...** 2010, p. 52-55.

SHAWE-TAYLOR, J.; CRISTIANINI, N. **Kernel methods for pattern analysis**. Cambridge University Press, 2004.

SILVA, R. P. **Alometria, estoque e dinâmica da biomassa de florestas primárias e secundárias na região de Manaus (AM)**. 2007. 135p. Tese (Doutorado em Ciências de Florestas Tropicais) – Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, 2007.

SIMARD M.; SAATCHI S. S; GRANDI G. D., The Use of Decision Tree and Multiscale Texture for Classification of JERS-1 SAR Data over Tropical Forest. **IEEE Transactions on Geoscience and Remote Sensing**, VOL. 38, NO. 5, SEPTEMBER 2000.

SNAP. **Sentinel Application Plataform.** (2019). Disponível em < https://step.esa.int/main/snap-6-0-released/>. Acesso em 11 ABR 2019.

SUN, G; RANSON K. J.; KHARUK V. I.; KOVACS K. Validation of surface height from shuttle radar topography mission using laser altimeter. **Remote Sensing of Environment**. 2002. Disponível em: <http://217.79.48.31/Articles/03/sun1.pdf >. Acesso em: 01 jun 2019.

The Economics of Ecosystems and Biodiversity (TEEB). **Mainstreaming the economics of nature**. 2008. 136p. ISBN 978-3-9813410-3-4. Disponível em: <http://www.teebweb.org>. Acesso em: 01 jun 2019.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 3rd ed. Academic Press. 2006. 856 p. ISBN 0123695317.

THIEL, C. J.; THIEL, C.; SCHMULLIUS, C. C. Operational Large-Area Forest Monitoring in Siberia Using ALOS PALSAR Summer Intensities and Winter Coherence. **IEEE Transactions on Geoscience and Remote Sensing**, v. 47, n. 12, p. 3993-4000. doi: 10.1109/TGRS.2009.2021469, 2009.

TOUZI, R. Target scattering decomposition in terms of roll-invariant target parameters. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, n. 1, p. 73 – 84, 2007.

TREUHAFT, R. N.; CHAPMAN, B. D.; SANTOS, J. R.; GONÇALVES, F. G.; DUTRA, L. V.; GRAÇA, P. M. L. A.; DRAKE, J. B. Vegetation profiles in tropical forests from multibaseline interferometric synthetic aperture radar, field, and LiDAR measuments. **Journal of Geophysical Research**, v. 114, D23110, doi:10.1029/2008JD011674, 2009.

TZENG, Y. C.; CHEN, K. S. A fuzzy neural network to SAR image classification. **IEEE Transactions on Geoscience and Remote Sensing**, v. 36, n. 1, p. 301 – 307, 1998.

UHL, C. BUSCHBACHER, R.; SERRÃO, E. A. S. Abandonned pastures in eastern Amazônia, I: patterns of plant sucession. **Journal of Ecology**, v.76, n.3, p.663 – 681, 1988.

United Nations Framework Convention on Climate Change (UNFCCC). **Kyoto Protocol Reference Manual on Accounting of Emissions and Assigned Amounts**. 2008. 130p. Disponível em: < http://unfccc.int/resource/docs/publications/08_unfccc_kp_ref_manual.pdf >. Acesso em: 22 jan. 2018.

VAN DER SANDEN, J. J. **Radar remote sensing to support tropical forest management**. 1997. 330p. Doctoral Thesis – Wageningen Agricultural University, The Netherlands, 1997.

WEKA. **Waikato Environment for Knowledge**. (2019). Disponível em < https://www.cs.waikato.ac.nz/~ml/weka/downloading.html>. Acesso em 11 Fev 2019.

WILLIAMS, R.J.; ZERIHUN, A.; MONTAGU, K.; HOFFMAN, M.; HUTLEY, L.B; CHEN, X. Allometry for estimating above-ground tree biomass in tropical and subtropical eucalypt woodlands: towards general predictive equations. **Australian Journal of Botany**, v. 53, p. 607–619, 2005.

WILLIAMS, M.L.; MILNE, T.; TAPLEY, I.; REIS, J. J.; SANFORD, M.; KOFMAN, B.; HENSLEY, S. Tropical forest biomass recovery using GeoSAR observations. In: International Geoscience and Remote Sensing Symposium (IGARSS), 2009, Cape Town, South Africa. **Proceedings …** Cape Town, 2009, p. 173-176.

WITTEN, I.H., FRANK, E., HALL, M.A., PAL, C.J., 2016. **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd Editio. ed, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Massachusetts. https://doi.org/10.1016/c2009-0-19715-5.

WOODHOUSE, I. H. **Introduction to microwave remote sensing**. Taylor & Francis Group CRC Press. 2006. 370p. ISBN 0415271231.

YANG, Y.; WEBB, G. I. 2009. Discretization for naive-Bayes learning: managing discretization bias and variance. **Machine Learning**, p. 39-74. doi: 10.1007/s10994-008-5083-5.

YUAN, Q.; SHEN, H.; LI, T.; LI, Z.; LI, S.; JIANG, Y.; XU, H.; TAN, W.; YANG, Q.; WANG, J.; GAO, J.; ZHANG, L. Deep learning in environmental remote sensing: Achievements and challenges. **Remote sensing of environment**, v. 241, p. 111716, 2020.

ZHANG, R.; MA, J.W. Feature selection for hyperspectral data based on recursive support vector machines. **International Journal of Remote Sensing**, v. 30, n. 14, p. 3669-3677, Jul. 2009.

ZHU, X. X.; TUIA, D.; MOU, L.; XIA, G. S.; ZHANG, L.; XU, F.; FRAUNDORFER, F. Deep Learning in Remote Sensing: A Comprehensive Review. **IEEE Geoscience and Remote Sensing Magazine**, v. 3, 2017.

7 APÊNDICES

Nesta seção são apresentados os seguintes documentos:

- Artigo publicado no Anuário do Instituto de Geociência – UFRJ, Vol. 43, 3 / 2020, p. 110-123, DOI: https://doi.org/10.11137/2020_3_110_123, ISSN 0101-9759 e-ISSN 1982-3908, intitulado “Proposta de Ajuste de Altura Interferométrica para Modelo de Estimativa de Biomassa / Interferometric Height Adjustment for Biomass Estimation Model”;
- Artigo publicado no International Journal of Advanced Engineering Research and Science (IJAERS), Vol-8, Issue-7, 2021, ISSN 2349-6495(P) 2456-1908(O), DOI: https://dx.doi.org/10.22161/ijaers.87.41, intitulado “Comparison between Quantitative and Qualitative Theme-Feature Forest Biomass Estimation Models built over SAR Data”; e
- Documentos de submissão à revista International Journal of Remote Sensing, intitulado “Categorization Optimization in the Construction of Thematic Products”.

Anuário do Instituto de Geociências - UFRJ
www.anuario.igeo.ufrj.br

Proposta de Ajuste de Altura Interferométrica para Modelo de Estimativa de Biomassa

Interferometric Height Adjustment for Biomass Estimation Model

Carlos Alberto Pires de Castro-Filho[1,2] & Edilson de Souza Bias[1]

[1]Universidade de Brasília (UnB), Instituto de Geociências, Programa de Pós-Graduação em Geociências Aplicadas e Geodinâmica, Campus Universitário Darcy Ribeiro, Instituto Central de Ciências – ICC, Ala Central, 70910-900, Brasília, DF, Brasil
[2]Diretoria de Serviço Geográfico, Quartel General do Exército, Bloco "F", 1º Andar, Setor Militar Urbano, 70630-901, Brasília, DF, Brasil
E-mails: carlos.pires.1976@gmail.com; edbias@unb.br
Recebido em: 06/03/2020 Aprovado em: 06/06/2020
DOI: http://doi.org/10.11137/2020_3_110_123

Resumo

A tecnologia de Radar de Abertura Sintética Interferométrica (InSAR), quando utilizada em diferentes comprimentos de onda, é capaz de gerar a altura interferométrica (H_{int}), calculada pela diferença aritmética entre o modelo digital de superfície (MDS) e o modelo digital do terreno (MDT). A H_{int} representa a altura da vegetação, visto que é o comprimento entre o terreno e o dossel da vegetação e, teoricamente, para áreas de solo exposto deve ter o valor estatisticamente igual a zero. O presente artigo tem por objetivo analisar a possibilidade de utilização de áreas identificadas como de solo exposto para ajustar dados InSAR, e consequentemente a H_{int}, visando a melhoria no modelo de estimativa de biomassa. A metodologia adotada inclui o uso de técnicas paramétricas e não paramétricas de busca de solução para definição dos parâmetros de ajuste da H_{int} sobre modelos matemáticos polinomiais e logarítmicos. Os melhores resultados foram obtidos com o modelo matemático logarítmico cujos parâmetros foram ajustados com a técnica do Gradiente Reduzido Generalizado. Entretanto, a análise dos resultados mostrou que não houve melhora significativa do coeficiente de correlação entre a biomassa florestal e a H_{int} original (r = 0,7518) e entre a biomassa florestal e a H_{int} ajustada (r = 0,7564).

Palavras-chave: *RADAR; InSAR; Biomassa*

Abstract

Interferometric Synthetic Aperture Radar (InSAR) technology, when used at different wavelengths, is capable of generating the interferometric height (H_{int}), calculated by the arithmetic difference between the digital surface model (DSM) and the digital terrain model (DTM). The H_{int} represents the height of the vegetation, since it is the length between the ground and the vegetation canopy and, theoretically, for exposed soil areas should be statistically iqual to zero. This paper aims at analyzing the possibility of using exposed soil identified areas to adjust InSAR data, and consequently the H_{int}, looking foward to improving the biomass estimation model. The adopted methodology includes the use of parametric and nonparametric solution search techniques to define the H_{int} adjustment parameters on polynomial and logarithmic mathematical models. The best results were obtained with the logarithmic mathematical model which parameters were adjusted using the Generalized Reduced Gradient technique. However, analysis of the results showed that there was no significant improvement in the correlation coefficient between forest biomass and original H_{int} (r = 0.7518) and between forest biomass and adjusted H_{int} (r = 0.7564).

Keywords: *RADAR; InSAR; Biomass*

1 Introdução

A estimativa de biomassa florestal é tratada por pesquisadores, universidades e institutos que atuam nas áreas de geociências, em todo o mundo, como um tema de grande relevância, suprindo pesquisas nas áreas de meteorologia e ecologia e tendo fortes impactos econômicos. Segundo Zhang *et al.* (2017), o assunto encontra-se entre os mais pesquisados e publicados, entre os anos de 2010 e 2015, nos periódicos de sensoriamento remoto, sendo utilizado como palavra-chave, em média, em 95 artigos técnicos publicados anualmente.

Dentre as tecnologias de sensoriamento remoto, as de Radar de Abertura Sintética (SAR) possuem destaque na modelagem de biomassa florestal devido a sua capacidade de caracterizar a geometria da região imageada (Saatchi *et al.*, 2007). Esta caracterização, quando associada a tecnologia interferométrica, chamada de InSAR, constrói modelos numéricos de elevação referentes ao comprimento de onda aplicado.

As técnicas de InSAR, quando utilizadas em diferentes comprimentos de onda, são capazes de gerar a altura interferométrica (H_{int}). A H_{int} pode ser calculada pela diferença aritmética das bandas de comprimento de onda X, referente ao modelo digital de superfície (MDS), e P, referente ao modelo digital do terreno (MDT). Neste caso ela representa a altura da vegetação, visto que é o comprimento entre o terreno e o dossel da vegetação.

Apesar de existir correlação entre a H_{int} e a altura da vegetação, são observadas regiões que apresentam inconsistências referentes aos dados de banda X e P. Teoricamente, o valor da H_{int} em áreas de solo exposto deve ser estatisticamente igual a zero. Isto ocorre porque tanto o retroespalhamento da banda X, que possui o comprimento de onda médio de 3 cm, como da banda P, com comprimento de onda médio de 0,7m, devem ocorrer junto ao solo exposto da região florestal imageada. Porém, em alguns casos são identificadas áreas de solo exposto cujo valor da H_{int} é estatisticamente diferente de zero.

No Brasil, o Projeto "Radiografia da Amazônia" desenvolvido pela Diretoria do Serviço Geográfico do Exército (DSG), encontra-se mapeando uma vasta área de região florestal (1.800.000 km^2) utilizando sensor de radar polarimétrico e interferométrico nas bandas X e P. Os dados polarimétricos, possibilitam a identificação e delimitação de regiões de solo exposto por meio de processos foto interpretativos ou por técnicas de classificação de imagens, como também, é possível observar que em algumas dessas regiões o valor de H_{int} é estatisticamente diferente de zero. Utilizando-se dessas premissas, o presente artigo tem por objetivo analisar a possibilidade de utilização de áreas identificadas como de solo exposto para ajustar dados InSAR, e consequentemente a H_{int}, visando a melhoria no modelo de estimativa de biomassa para futuros cálculos com o uso de técnicas de aprendizado de máquina.

2 Fundamentação Teórica

2.1 Modelo Digital de Elevação, Modelo Digital de Superfície e Modelo Digital do Terreno

De acordo com Burrough (1986), o termo Modelo Digital do Terreno (MDT) foi primeiramente utilizado por Miller e LaFlamme em 1958 e se refere a um conjunto de pontos que modelam uma região de superfície do terreno por meio de dados altimétricos. Os dados do MDT são representados pelos valores das coordenadas nos eixos x, y e z, onde z, valor a ser modelado, é em função de x e y, ou seja, z=f(x,y).

Atualmente, como consequência da evolução tecnológica, a definição clássica dos modelos tridimensionais que representam superfícies passou por alterações e recebeu uma série de termos distintos. Alguns órgãos brasileiros e internacionais apresentam diferentes definições, como: Modelo Digital de Elevação (MDE), Modelo Digital de Superfície (MDS) e, finalmente, Modelo Digital de Terreno (MDT). A conceituação de cada um deles pode ser observada na Tabela 1.

2.2 Biomassa Florestal

Biomassa florestal ou fitomassa é a quantidade, em unidade de massa, do material lenhoso contido em uma unidade de área de floresta (Araújo *et al.*, 1999). Além dos métodos *in loco* de estimativa de biomassa, outros métodos são realizados com o uso de sensores remotos, não havendo a necessidade de extensos trabalhos de campo, o que viabiliza financeiramente o levantamento de grandes áreas. Nestes casos busca-se relacionar as características representadas pelas imagens às características físicas medidas das árvores ou, diretamente, à biomassa existente na região imageada.

Diversos autores vêm buscando utilizar dados de SAR para modelar e, consequentemente, estimar a quantidade de biomassa em diversos tipos de florestas. Dentre esses autores, destacam-se os trabalhos de Pope *et al.* (1994), Santos *et al.* (2003), Saatchi *et al.* (2007), Debastiani *et al.* (2019) e Oliveira & Locks (2019).

2.3 Radar de Abertura Sintética Interferométrico com Aplicação na Estimativa de Biomassa

Juntamente aos dados polarimétricos, o potencial da informação interferométrica de um SAR para modelagem de biomassa em ambientes tropicais ainda possui vasta área de pesquisa. Gama (2007), ao buscar uma modelagem para estimar a biomassa em um povoamento de *Eucalyptus*, concluiu que os atributos que obtiveram melhores resultados foram provenientes dos produtos interferométricos.

Sigla	Definição	Orgão ou Instituição
MDE	Produto cartográfico obtido a partir de um modelo matemático que representa um fenômeno, de forma contínua, a partir de dados adequadamente estruturados e amostrados no mundo real.	Especificação Técnica para Produtos de Conjunto de Dados Geoespaciais (ET-PCDG), elaborada pela DSG (2016)
	Representação matemática computacional da distribuição de um fenômeno espacial que ocorre dentro de uma região da superfície terrestre.	Felgueiras (2004), do Instituto Nacional de Pesquisas Espaciais (INPE)
MDS	Produto obtido a partir de um modelo digital de elevações que representa o solo exposto e os acidentes naturais e/ou construídos pelo ser humano encontrados acima do solo (edificações, pontes, corte e aterro do terreno, vegetação, etc), de forma contínua e suavizada, a partir de dados adequadamente estruturados e amostrados do mundo real.	ET-PCDG (DSG, 2016); *European Space Agency* – ESA (Mouratidis, 2014); e *Japanese Space Agency* – Jaxa (JAXA, 2019)
	Modelo digital que representa as altitudes da superfície topográfica agregada aos elementos geográficos existentes sobre ela, como cobertura vegetal e edificações.	IBGE (2019)
MDT	O modelo representa o solo exposto, de forma contínua e suavizada, a partir de dados adequadamente estruturados e amostrados da superfície física da Terra.	ET-PCDG (DSG, 2016); e ESA (Mouratidis, 2014)
	Representam as altitudes da superfície topográfica, desconsiderando as alturas dos elementos geográficos existentes sobre ela, como cobertura vegetal e edificações	IBGE (2019)

Tabela 1 Quadro Comparativo de Definições de Modelos Digitais.

Dentre os produtos gerados pela interferometria de radar, Neeff *et al.* (2005) caracteriza a H_{int} como a diferença entre os modelos numéricos de elevação obtidos em bandas que representam modelos gerados no solo (MDT) e no dossel (MDS). Segundo Neeff *et al.* (2005), Gama (2007), Santoro & Cartus (2018) e Schlund *et al.* (2019) o valor da H_{int} tem relação com a característica física de altura total do indivíduo arbóreo, bem como o de biomassa florestal.

Na região amazônica, Neeff *et al.* (2005) trabalharam na Floresta Nacional (FLONA) do Tapajós utilizando dados polarimétricos e interferométricos de SAR das bandas X e P. Neste caso os autores analisaram a correlação entre a altura florestal e a H_{int} e entre a biomassa, a área basal, o retroespalhamento obtido pelas bandas polarimétricas e a H_{int}.

Para o cálculo da biomassa, Neeff *et al.* (2005) atingiram um coeficiente de determinação em torno de $r^2 = 0,84$ ao desenvolverem o modelo da Equação 1 onde a biomassa é dada em toneladas por hectare (t/ha), a H_{int} está em metros (m) e σ^0_{HH} é o valor do coeficiente retroespalhamento na polarização HH em unidade de decibéis (dB).

$$Biomassa = 44,965+13,87H_{int}+10,566\ \sigma^0_{HH} \quad (1)$$

Mais recentemente, Castro-Filho *et al.* (2013), após analisar diversas variáveis obtidas por dados SAR interferométricos e polarimétricos, conclui que a H_{int} foi a que obteve maior fator de correlação com a biomassa florestal aérea da região amazônica, no valor de r = 0,70.

2.4 Projeto Radiografia da Amazônia

No Brasil, entre os projetos que visam gerar imagens polarimétricas e interferométricas de SAR e que poderão ser utilizadas na estimativa de biomassa destaca-se o de Cartografia da Amazônia, mais especificamente o Subprojeto Cartografia Terrestre, também conhecido como "Radiografia da Amazônia". Este projeto, coordenado pelo CENSIPAM (Centro Gestor e Operacional do Sistema de Proteção da Amazônia), visa recobrir uma área total de cerca de 1,1 milhão de km² da região amazônica, visando a elaboração de cartas na escala 1:50.000 pela DSG (2019). Além de realizar este mapeamento, o projeto visa também gerar dados necessários ao suporte de projetos de infraestrutura e exploração sustentável de recursos naturais da região.

No projeto será utilizada a tecnologia de SAR aerotransportado (sensor OrbiSAR), gerando diversos produtos, todos com resolução espacial de 5m: orto-imagens e imagens complexas X-HH e P-HH/HV/VV; MDS; e MDT . Além do grande volume de dados gerados, novos atributos

Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 43 - 3 / 2020 p. 110-123

que melhor representam as características estruturais e de biomassa florestal poderão ser extraídos, como a H_{int}.

2.5 Técnicas de Busca de Solução

Diversas são as técnicas utilizadas para busca de solução, visando a definição dos valores de incógnitas em equações matemáticas. As técnicas de busca de solução podem ser divididas em paramétricas e não paramétricas, apresentando vantagens e limitações para cada caso.

As técnicas de busca de soluções paramétricas analisam os dados a partir de suas distribuições estatísticas. Uma das vantagens destas técnicas é a possibilidade de se utilizar as propriedades e técnicas estatísticas, como a de associar medidas de incerteza às estimativas.

No caso da solução paramétrica, existe grande destaque por parte das técnicas de regressão múltipla que buscam modelar uma variável, chamada de dependente, a partir de múltiplas outras variáveis, chamadas de independentes (Neter *et al.*, 1996). Esta abordagem é plenamente utilizada sobre dados de sensoriamento remoto, onde as variáveis dependentes comumente são bandas obtidas diretamente de sensores ou extraídas das mesmas. Dentre as técnicas paramétricas mais utilizadas, destaca-se o método dos mínimos quadrados, conforme Equação 2. Nele, é utilizado um sistema de equações a ser solucionada pela forma matricial

$$A^TAX=A^TY \quad (2)$$

onde as matrizes *A* e *Y* possuem valores conhecidos e a matriz *X* é a solução composta pelas incógnitas apresentadas nos modelos matemáticos.

Diferentemente, as técnicas de busca de solução não paramétricas não utilizam as características estatísticas dos dados de entrada, tornando-se independente de suas distribuições. No âmbito do sensoriamento remoto, estas técnicas costumam ser aplicadas sobre dados de SAR devido aos diferentes tipos de distribuição estatística que estes dados apresentam.

Um dos métodos de busca de solução não paramétricos é o do Gradiente Reduzido Generalizado (GRG) Não Linear apresentado por Carpentier & Abadie (1966), também conhecido como "método de solução", ou "*solver"*, devido às diversas aplicações nas mais variadas áreas das ciências onde visa solucionar problemas matemáticos complexos.

Segundo Martínez & Santos (1995), o método GRG Não Linear tem a seguinte função:

- minimizar a função objetivo $z = f(\mathbf{X})$
- tendo a matriz de parâmetros que se deseja encontrar $\mathbf{X} \equiv [x_i]^T$
- e sujeito a função de restrições $h(\mathbf{X}) = 0$, $\beta \geq x \geq \alpha$.

O problema deve ser iniciado com um $\mathbf{X}_k$ qualquer, preferencialmente próximo à solução. Através de um método iterativo, inicia-se a busca dos valores dos parâmetros utilizando a direção de busca identificada pelo gradiente reduzido $\nabla\varphi(\mathbf{X})$. Se o módulo do vetor gradiente reduzido $\nabla\varphi(\mathbf{X})$ for menor que a tolerância de convergência pré-definida, a variável $\mathbf{X}_k$ é tida como ponto ótimo da função.

Outro método de busca de solução vastamente utilizado é o de Algoritmo Genético. Primeiramente utilizado por John Holland, em 1975, em seu livro *Adaptation in Natural and Artificial Systems*, o método foi apresentado como uma adaptação à Teoria da Evolução de Charles Darwin (Reeves, 2010).

O algoritmo genético clássico busca uma solução sub ótima para problemas matemáticos visando a maximização (ou minimização, dependendo do problema) de uma função objetivo. A busca inicia com um conjunto de soluções aleatórias chamado de "população", onde os valores dos parâmetros matemáticos iniciais são os "cromossomos". Na sequência, são realizadas iterações computacionais onde os valores dos cromossomos são ajustados, por meio de técnicas de "*crossover*" e de "mutação", para que a população maximize a função objetivo.

No âmbito do sensoriamento remoto, algoritmos genéticos são usualmente utilizados para auxiliar na seleção de atributos, muitas vezes caracterizados por bandas espectrais, visando a classificação de tipos de uso do solo, como é o caso do trabalho de Singh & Singh (2017).

3 Metodologia

A área de estudo encontra-se inserida no município de São Gabriel da Cachoeira que é localizado às margens do Rio Negro, no noroeste do estado do Amazonas, conforme ilustrado na Figura 1. Conforme o Projeto RadamBrasil (1977) a maior parte da vegetação encontrada na área de estudo é composta por regiões fitoecológicas de contato florestal/formações edáficas (campinaranas). Estas regiões são caracterizadas de três formas:

- florestas densas, submontana e com o relevo dissecado. O Projeto RadamBrasil (1977) afirma que o volume médio de biomassa da área é de 107,4m^3/ha;
- florestas densas, submontana e com o relevo ondulado; e
- florestas densas, terras baixas e relevo com presença de platôs.

Figura 1 Imagem de alta resolução do Google Earth (2019) da região amazônica da "cabeça do cachorro" com destaque para os blocos de processamento do Projeto "Radiografia da Amazônia".

Os materiais utilizados são imagens obtidas por meio do Projeto "Radiografia da Amazônia" (DSG, 2019). As áreas de trabalho e as respectivas imagens utilizadas referem-se ao bloco compreendido entre as latitudes 0º e 1º sul e entre as longitudes 67º e 68º oeste, na localidade de São Gabriel da Cachoeira - AM. A Figura 1 apresenta imagens do sistema Google Earth (2019) referentes à região do extremo noroeste brasileiro, cuja fronteira com a Colômbia tem o formato semelhante à de uma "cabeça de cachorro" (em amarelo). Nesta figura observa-se, em cinza, o quadriculado referente aos blocos de processamento do Projeto "Radiografia da Amazônia", de 1 (um) grau em longitude por 1 (um) grau em latitude, e em vermelho a sede municipal de São Gabriel da Cachoeira – AM.

Na Figura 2 (A,B,C e D) observa-se imagens da área de trabalho nas polarizações X-HH e P-HH/VV e imagem do Google Earth (2019). Apesar de não serem o foco do presente trabalho, por meio de imagens em diferentes polarizações é possível diferenciar importantes informações do terreno, como as diferentes classes de tipo e uso do solo.

A pesquisa foi estruturada de acordo com fluxograma apresentado na Figura 3, sendo que cada etapa será descrita nos itens a seguir.

3.1 Construção de Modelo Interferométrico Preliminar de Estimativa de Biomassa

Inicialmente foi construído um modelo preliminar de estimativa de biomassa, fundamental para possibilitar a posterior comparação com os demais modelos onde foram utilizadas H_{int} ajustadas, utilizando como variável independente, ou atributo, somente a H_{int} original. Detalhes sobre a construção deste modelo encontram-se descritos por Castro-Filho *et al.* (2013).

3.2 Definição de Modelo de Ajuste de Altura Interferométrica (H_{int})

A principal condição de contorno para o problema apresentado no presente trabalho é de que o valor da H_{int} nos pontos identificados como de solo exposto tenham o valor próximo a zero. Por outro lado, a hipótese a ser

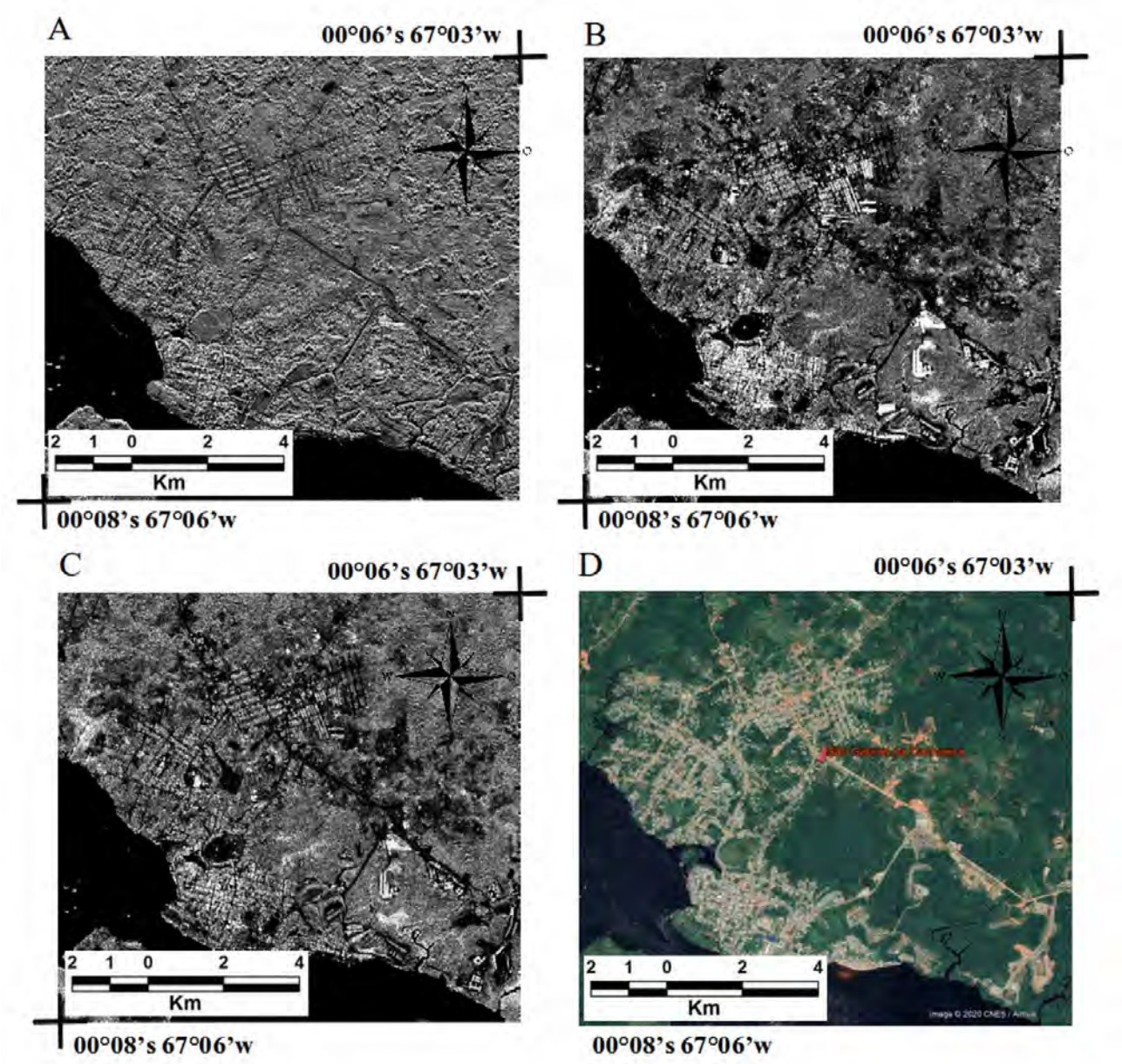


Figura 2 Imagens da região de trabalho; A. X-HH; B. P-HH; C. P-VV; D. Imagem Google Earth (2019).

testada é de que seja possível o ajuste da H_{int} por meio da utilização de modelos matemáticos, visando melhoria no modelo de estimativa de biomassa florestal.

Os modelos matemáticos de ajuste utilizados serão os modelos polinomiais

$$aX^2 + bY^2 + cXY + dX + eY + f - H_{int} = 0 \qquad (3)$$

e logarítmicos

$$a\log X + b\log Y + c - d\log H_{int} = 0 \qquad (4)$$

Nos modelos apresentados, os parâmetros de ajuste são representados pelas incógnitas *a*, *b*, *c*, *d*, *e* e *f* e as variáveis X e Y são valores das coordenadas no sistema UTM (Universal Transversa de Mercator) obtidos nas amostras de solo exposto.

3.3 Identificação e Coleta de Amostras de Solo Exposto

O processo de identificação de áreas de solo exposto foi foto-interpretativo, considerando a textura e as características do retroespalhamento nas imagens polarimétricas. Logo, as áreas de solo expostos foram

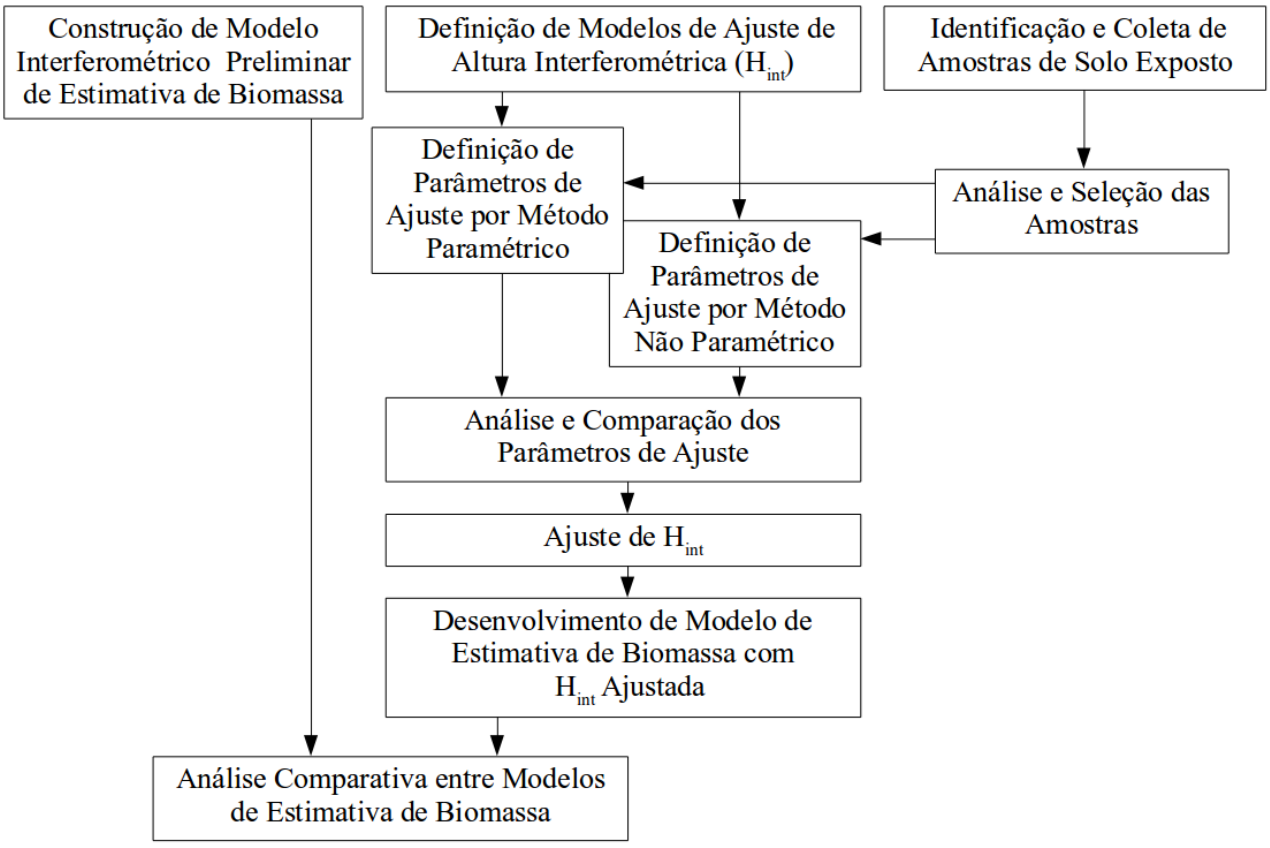


Figura 3 Fluxograma dos processos.

identificadas visualmente e os polígonos das áreas de interesse foram delimitados para compor as amostras da respectiva classe em estudo.

3.4 Análise e Seleção de Amostras

O processo de análise e seleção de amostras foi necessário pois foram observados que alguns pixels pertencentes às áreas de interesse coletadas representavam outros objetos distintos de solo exposto, como árvores isoladas, cujo valor da H_{int} deve ser diferente de zero. Logo, foi realizado um processo de identificação de *outliers* utilizando-se o parâmetro de 3 sigma (σ) e excluindo os valores destes pixels da média aritmética de cada uma das áreas de interesse.

3.5 Definição de Parâmetros de Ajuste por Método Paramétrico

O ajuste paramétrico sobre os modelos matemáticos apresentados no item 3.2 foi realizado por meio do método dos mínimos quadrados, conforme apresentado na Equação 2. Neste caso, a matriz *A* foi composta pelos valores de coordenadas médias de cada área de interesse, a matriz *Y* serão os valores da H_{int} média de cada área de interesse e a matriz *X* será a solução composta pelas incógnitas dos modelos matemáticos.

3.6 Definição de Parâmetros de Ajuste por Método Não Paramétrico

A técnica de busca de solução não paramétrica utilizada foi um método de otimização que buscou por parâmetros que minimizassem o valor do erro médio quadrático, objetivando o valor da H_{int} média para cada área de interesse igual a zero. Neste caso utilizou-se tanto o método de GRG como o de algoritmo genético, ambos disponíveis no sistema Microsoft Excel.

3.7 Análise e Comparação dos Parâmetros de Ajuste

Foram analisados e comparados os parâmetros de ajuste obtidos pelos métodos paramétrico e não paramétrico, com o intuito de:

Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 43 - 3 / 2020 p. 110-123

- compreender os valores de cada parâmetro, identificando as possíveis fontes de erro que levaram aos mesmos;
- identificar similaridades entre os parâmetros obtidos por diferentes métodos;
- analisar os erros médios quadráticos obtidos; e
- selecionar o(s) modelos mais adequado(s) a serem aplicados na próxima etapa do processo, isto é, de ajuste da H_{int}.

3.8 Ajuste de Altura Interferométrica (H_{int})

Os modelos selecionados foram aplicados matematicamente à H_{int} dos pixels cujos valores de biomassa são conhecidos. A partir deste momento os valores passaram a ser chamados de H_{int} Ajustada.

3.9 Desenvolvimento de Modelo de Estimativa de Biomassa com H_{int} Ajustada

A partir dos valores ajustados foram desenvolvidos modelos de estimativa de biomassa, utilizando-se somente a técnica paramétricas de regressão estatística, visando obter solução para os valores das incógnitas. Optou-se somente pela solução paramétrica já que somente esta é capaz de gerar o coeficiente de correlação, indicador de qualidade que será utilizado no presente trabalho.

3.10 Análise Comparativa entre Modelos de Estimativa de Biomassa

A análise comparativa será realizada entre o modelo de estimativa de biomassa apresentado por Castro-Filho *et al.* (2013), conforme citado no subitem 3.1 deste trabalho, e o modelo obtido pelo H_{int} Ajustado. Para tal, foram analisados ambos os valores dos coeficientes de correlação.

4 Resultados

4.1 Análise e Seleção de Amostras

Com base na identificação das áreas de solo exposto, foi realizada a coleta de 50 polígonos de áreas de interesse. Cada polígono foi composto, em média, por 338 pixels, totalizando uma amostra de 16.875 pixels de solo exposto.

A Tabela 2 apresenta, para cada um dos 50 polígonos, os valores médios iniciais de H_{int}, o valor do desvio padrão (σ), a quantidade total de pixels e a quantidade de pixels *outliers*, isto é, com módulos acima de 3σ. Ao analisar as amostras, observou-se que 36 dos 50 polígonos (72%) possuíam ao menos um pixel identificado como *outliers*. Os percentuais de *outliers* por polígono encontram-se destacados em vermelho na tabela.

Após a identificação dos pixels *outliers* e a respectiva exclusão, os valores médios de H_{int} para cada polígono foram recalculados. Os novos valores também se encontram na Tabela 2.

Apesar do percentual de polígonos que possuíam pixels identificados como *outliers* ter sido relativamente alto, a quantidade desses pixels totalizou 273, o que corresponde a somente 1,6% do total de pixels selecionados. Ao analisar as amostras selecionadas observou-se que alguns pixels isolados e internos a polígonos de solo exposto possuíam valores discrepantes, o que explica a ocorrências desses *outliers*. Embora não tenha havido trabalho de campo para comprovar *in loco* a existência de elementos referentes

Polígono	H_{int} Médio (m) Inicial	σ (m)	Qtd Tot Pixels	Qtd Outliers	% Outliers	H_{int} Médio (m) Recalculado
1	0,397	0,205	333	0	0%	0,397
2	0,315	0,114	474	7	1%	0,308
3	1,088	0,547	179	2	1%	1,069
4	0,422	0,274	320	10	3%	0,388
5	3,341	1,538	621	0	0%	3,341
6	0,430	0,489	434	14	3%	0,359
7	0,607	0,413	620	0	0%	0,607
8	0,766	0,353	98	1	1%	0,754
9	0,680	0,735	107	2	2%	0,627
10	0,290	0,117	259	0	0%	0,290
11	0,252	0,100	108	0	0%	0,252
12	0,454	0,336	110	3	3%	0,412
13	0,558	0,366	457	3	1%	0,547
14	0,575	0,428	304	5	2%	0,540
15	0,523	0,287	169	5	3%	0,494

Tabela 2 Análise de *Outliers.*

Polígono	H_{int} Médio (m) Inicial	σ (m)	Qtd Tot Pixels	Qtd Outliers	% Outliers	H_{int} Médio (m) Recalculado
16	0,163	0,098	163	0	0%	0,163
17	0,913	1,685	265	13	5%	0,602
18	0,417	0,450	117	3	3%	0,356
19	2,872	0,944	61	0	0%	2,872
20	2,792	1,727	306	1	0%	2,775
21	1,074	0,762	251	3	1%	1,039
22	0,378	0,243	93	3	3%	0,350
23	0,371	0,358	286	10	3%	0,317
24	0,448	0,483	881	18	2%	0,406
25	0,477	0,464	422	10	2%	0,436
26	0,562	0,616	1462	32	2%	0,503
27	0,333	0,437	370	11	3%	0,270
28	2,322	1,413	865	15	2%	2,226
29	1,185	0,557	194	0	0%	1,185
30	0,745	0,598	230	4	2%	0,702
31	2,270	1,048	278	0	0%	2,270
32	8,375	1,135	132	0	0%	8,375
33	3,045	1,229	290	3	1%	3,004
34	6,843	2,261	232	0	0%	6,843
35	2,084	0,881	337	1	0%	2,076
36	0,717	0,842	124	2	2%	0,666
37	1,379	1,763	225	6	3%	1,202
38	0,884	0,639	150	1	1%	0,869
39	0,364	0,123	363	1	0%	0,362
40	0,386	0,073	257	1	0%	0,387
41	0,252	0,075	229	0	0%	0,252
42	0,297	0,043	556	9	2%	0,296
43	0,363	0,175	695	20	3%	0,344
44	0,709	0,819	1171	31	3%	0,634
45	1,206	1,616	274	7	3%	1,053
46	0,471	0,679	358	8	2%	0,383
47	0,943	0,747	182	3	2%	0,892
48	2,766	1,759	157	0	0%	2,766
49	1,602	1,563	156	0	0%	1,602
50	0,789	1,145	150	5	3%	0,655

Tabela 2 Cont.

aos pixels isolados, em alguns casos pôde-se supor que se tratavam de árvores isoladas.

A Tabela 3 apresenta os valores de E (coordenada leste UTM média), N (coordenada norte UTM média) e da H_{int} média para cada polígono de solo exposto identificado. Observa-se que alguns destes polígonos apresentam valores médios de H_{int} bem acima de zero, como é o caso dos polígonos 32 e 34, com valores destacados em vermelho.

4.2 Definição de Parâmetros de Ajuste por Método Paramétrico

Os valores das médias de X, Y e H_{int}, para cada polígono, foram aplicados sobre os modelos polinomiais e logarítmicos apresentados, respectivamente, nas Equações 3 e 4. Iniciou-se, portanto, a etapa de definição dos parâmetros de ajuste dos modelos.

Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 43 - 3 / 2020 p. 110-123

Dentre os testes realizados envolvendo modelos paramétricos, o único modelo matemático que apresentou correlação significativa ($p<0,01$) foi o logarítmico, conforme Equação 5.

$$\log H_{int} = - 58,4753 \log X + 342,15 \quad (5)$$

Durante o processo de definição de parâmetros, utilizando modelo paramétrico, observou-se que a matriz *A*, descrita na Equação 2, não se mostrou inversível por possuir colunas, oriundas de variáveis, correlacionadas. Tal fato limitou os resultados possíveis no modelo paramétrico de definição de parâmetros.

4.3 Definição de Parâmetros de Ajuste por Método Não Paramétrico

Os resultados obtidos utilizando modelos não paramétricos para a definição de parâmetros de ajuste também foram limitados. Em diversos casos o algoritmo

N° Polígono	E (m)	N (m)	H_{int} (m)	N° Polígono	E (m)	N (m)	H_{int} (m)
1	721205,300	9983386,336	0,397	26	712316,143	9990773,685	0,503
2	721321,039	9983552,719	0,308	27	712524,095	9990789,387	0,270
3	722055,085	9983070,367	1,069	28	713433,471	9990653,224	2,226
4	721452,790	9983823,258	0,388	29	712827,010	9989834,098	1,185
5	720009,050	9984039,058	3,341	30	712693,960	9989846,327	0,702
6	719200,476	9984949,631	0,359	31	710890,162	9993316,259	2,270
7	716096,976	9986232,734	0,607	32	710981,818	9992679,735	8,375
8	716852,577	9984747,423	0,754	33	711530,732	9993308,031	3,004
9	715139,143	9985056,381	0,627	34	711364,698	9993100,151	6,843
10	715299,015	9984904,595	0,290	35	711947,366	9992965,655	2,076
11	714896,759	9985668,704	0,252	36	711596,434	9997336,598	0,666
12	714665,561	9985828,738	0,412	37	711481,438	9998210,868	1,202
13	712611,465	9985913,513	0,547	38	711646,678	9998572,718	0,869
14	712718,144	9985984,916	0,540	39	719636,395	9988305,967	0,362
15	712717,896	9985893,628	0,494	40	719783,008	9988354,727	0,387
16	712816,963	9986013,098	0,163	41	719868,581	9988564,410	0,252
17	713174,067	9985262,897	0,602	42	721165,932	9989448,876	0,296
18	712959,868	9985329,605	0,356	43	711923,400	9984351,370	0,344
19	712462,049	9985252,951	2,872	44	711307,092	9983909,504	0,634
20	711220,738	9987750,574	2,775	45	709222,247	9984210,318	1,053
21	710967,742	9986910,927	1,039	46	709461,771	9984239,714	0,383
22	710709,778	9986931,444	0,350	47	711958,603	9984199,749	0,892
23	712424,293	9988266,105	0,317	48	708350,000	9984405,955	2,766
24	712204,264	9989102,034	0,406	49	707629,103	9983892,436	1,602
25	712222,718	9989323,265	0,436	50	708219,207	9985361,345	0,655

Tabela 3 Valores de X (latitude média), Y (longitude média) e da H_{int} média para cada polígono de solo exposto identificado.

Modelo Polinomial (Equação 3): Precisão de Restrição: $1x10^{-13}$; Convergência: $1x10^{-14}$							
Incógnitas	*a*	*b*	*c*	*d*	*e*	*f*	Resíduo
Valores Iniciais	$1x10^{-13}$	$1x10^{-15}$	$1x10^{-14}$	$1x10^{-7}$	$1x10^{-8}$	0	153,3275
1ª Iteração (Final)	$-1,68x10^{29}$	$-8,16x10^{29}$	$5,12x10^{29}$	$5,94x10^{29}$	$8,60x10^{29}$	$-2,33x10^{29}$	**$3,03x10^{89}$**

Tabela 4 Solução 1 Não Paramétrica.

Modelo Logarítmico (Equação 4): Precisão de Restrição: 0,01; - Convergência: 0,001					
Incógnitas	*a*	*b*	*c*	*d*	Resíduo
Valores Iniciais	0,2	0,15	-1	-2	75,233
1ª Iteração	0,1301	0,1030	-1,2984	-1,7803	24,245
2ª Iteração (Final)	0,1190	0,0955	-1,3462	-0,1804	**0,2483**

Tabela 5 Solução 2 Não Paramétrica.

Modelo Logarítmico (Equação 4): Precisão de Restrição: 0,001; Convergência: 0,0001					
Incógnitas	*a*	*b*	*c*	*d*	Resíduo
Valores Iniciais	0,2	0,15	-1	-2	75,233
1ª Iteração	0,1301	0,1030	-1,2984	-1,7804	24,245
2ª Iteração (Final)	0,1177	0,0946	-1,3515	-0,0002	**$3{,}22 \times 10^{-6}$**

Tabela 6 Solução 3 Não Paramétrica.

Modelo Logarítmico (Equação 4): Precisão de Restrição: 0,01; Convergência: 0,001; Início múltiplo com população de 10					
Incógnitas	*a*	*b*	*c*	*d*	Resíduo
Valores Iniciais	0,2	0,15	-1	-2	75,233
1ª Iteração	0,1301	0,1030	-1,2984	-1,7803	24,245
2ª Iteração	0,1190	0,0955	-1,3462	-0,1804	0,2483
3ª Iteração (Final)	$-3{,}1540 \times 10^{29}$	$-1{,}9523 \times 10^{29}$	$-3{,}8432 \times 10^{29}$	$6{,}4128 \times 10^{29}$	**$6{,}1893 \times 10^{62}$**

Tabela 7 Solução 4 Não Paramétrica.

de busca não convergiu para uma solução com função objetivo próxima a zero, como são os casos apresentados nas Tabelas 4 e 7.

A busca de solução por algoritmo genético não obteve qualquer resultado satisfatório por não ter convergido para uma solução, logo não serão apresentados no presente trabalho. Os parâmetros utilizados foram os seguintes: população de tamanho 500 e 1000; taxa de mutação de 2% e 5%; e *crossover* por elitismo de 20% e 30%.

Já, o ajuste por GRG convergiu para algumas soluções, tendo como parâmetros de parada de busca do algoritmo o tempo máximo de 100 segundos e o número máximo de 1000 iterações.

A seguir encontram-se alguns exemplos de soluções obtidas por meio do sistema Microsoft Excel Solver utilizando modelos matemáticos polinomiais (Tabela 4) e logarítmicos (Tabelas 5, 6 e 7). Além dos modelos matemáticos, nas tabelas também se encontram os parâmetros de precisão de restrição e convergência aplicados ao algoritmo de busca do GRG.

Das soluções obtidas, somente a Solução 2 Não Paramétrica (Tabela 5) e a Solução 3 Não Paramétrica (Tabela 6) convergiram, respectivamente, para resíduos de 0,2483m e $3{,}22 \times 10^{-6}$m, ambos na segunda iteração. A Solução 1 Não Paramétrica (Tabela 4) e a Solução 4 Não Paramétrica (Tabela 7) apresentaram resultados insatisfatórios, divergindo para valores de resíduos de $3{,}03 \times 10^{89}$m e $6{,}1893 \times 10^{62}$m. Em função destes resultados obtidos nos testes para o método não paramétrico, no presente trabalho será dada continuidade somente aos modelos não paramétricos das Soluções 2 e 3.

Em função dos parâmetros obtidos na Tabela 5, a equação de ajuste para a Solução 2 Não Paramétrica é:

$$0{,}1190 \log X + 0{,}0955 \log Y - 1{,}3462 + 0{,}1804 \log H_{int} = 0$$

ou

$$\log H_{int} = -0{,}6596 \log X - 0{,}5294 \log Y + 7{,}4623 \quad (6)$$

Da mesma forma, a equação de ajuste para a Solução 3 Não Paramétrica, conforme parâmetros da Tabela 6, é a seguinte:

$$0{,}1177 \log X + 0{,}0946 \log Y - 1{,}3515 + 0{,}0002 \log H_{int} = 0$$

ou

$$\log H_{int} = -588{,}5 \log X - 473{,}0 \log Y + 6757{,}5 \quad (7)$$

4.4 Análise e Comparação dos Parâmetros de Ajuste

Ao analisar as soluções de ajuste apresentadas nas Equações 5, 6 e 7, observa-se que, tanto para o ajuste

Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 43 - 3 / 2020 p. 110-123

paramétrico como para o não paramétrico, somente o modelo matemático logarítmico apresentou resultados significativos. Da mesma forma, em todos os casos os parâmetros que acompanham as variáveis independentes X e Y possuem sinal negativo, seguido de um valor positivo de translação no eixo Z.

O modelo paramétrico da Equação 5 independe da variável independente Y, a qual corresponde à variação geográfica na latitude. Por ser dependente somente de X, entende-se que o ajuste a ser realizado no valor de H_{int} varia somente com a longitude. Neste caso, observa-se que a imagem em trabalho possui um efeito de "rampa" em somente uma direção.

Por outro lado, os modelos não paramétricos das Equações 6 e 7 apresentaram dependência com ambas as variáveis X e Y. Destaca-se o fato de que em ambos os casos os parâmetros obtidos estão proporcionais, isto é, na Equação 6 ambos são decimais, quanto na Equação 7 ambos são centenas.

Com relação aos resíduos obtidos, a Equação 7, referente ao processamento da Tabela 4, foi a que apresentou o melhor resultado, apresentando um modelo de ajuste para H_{int} que tornou nulo o valor para áreas de solo exposto.

4.5 Ajuste de H_{int} e desenvolvimento de Modelo de Estimativa de Biomassa com H_{int} Ajustada

Os modelos de ajuste das Equações 5, 6 e 7 foram aplicados na H_{int} original obtendo então três novos valores de H_{int} ajustada, respectivamente, H_{intAj1}, H_{intAj2} e H_{intAj3}. Cada uma dessas novas bandas foi então utilizada como atributo em um modelo de estimativa de biomassa, obtendo os seguintes resultados de coeficiente de correlação:

$$H_{intAj1} \rightarrow r = 0{,}5290;$$
$$H_{intAj2} \rightarrow r = 0{,}7512;$$
$$H_{intAj3} \rightarrow r = 0{,}7564.$$

Observa-se que, dos resultados obtidos, somente os coeficientes de correlação envolvendo as H_{intAj2} e H_{intAj3} foram significativamente positivos, isto é, como valores acima ou próximos de 0,75.

4.6 Análise Comparativa entre Modelos de Estimativa de Biomassa

A Figura 4 apresenta o modelo preliminar de estimativa de biomassa desenvolvido com a H_{int} original (Castro-Filho *et al.*, 2013), enquanto as Figuras 5(A) e 5(B) apresentam os modelos envolvendo os atributos ajustados de H_{intAj2} e H_{intAj3}. A diferença entre o coeficiente de correlação do modelo inicial (r = 0,7518) e dos modelos ajustados (r = 0,7512 e r = 0,7564) não foi significativa.

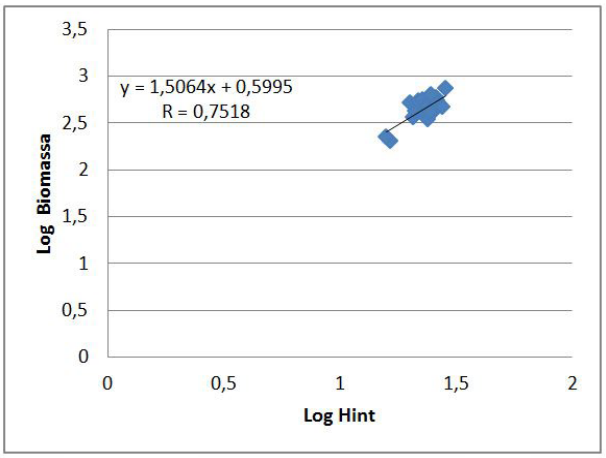
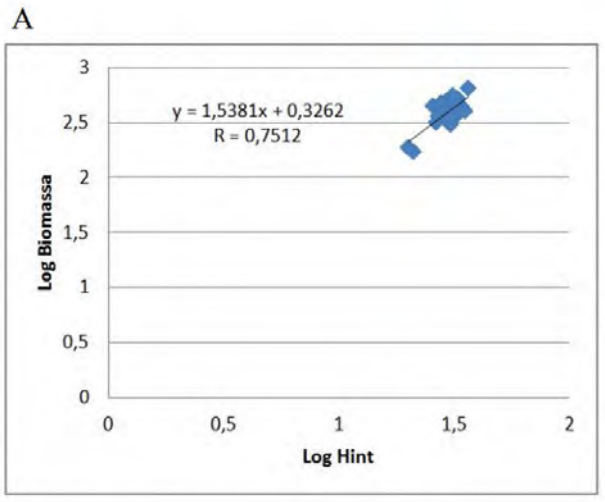


Figura 4 Modelo Preliminar.



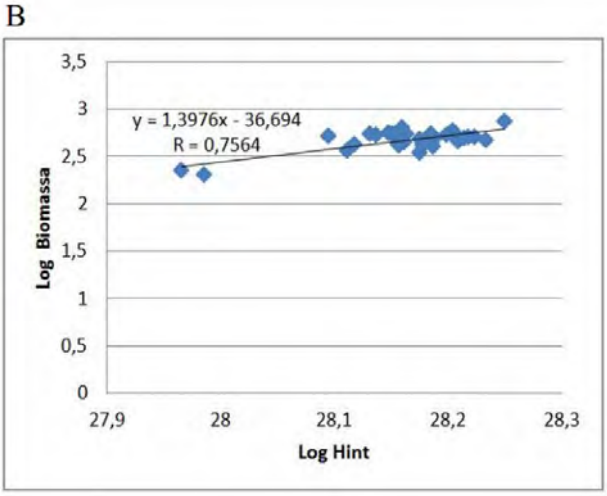


Figura 5 Modelos Ajustados; A. Modelo envolvendo a H_{intAj2}; B. Modelo envolvendo a H_{intAj3}.

Os parâmetros dos modelos de estimativa de biomassa apresentados nas Figuras 4 e 5(A) se apresentaram equivalentes, o que resultou em modelos com gráficos semelhantes. Por outro lado, a Figura 5(B) apresenta um modelo com valores mais dispersos no eixo das abscissas, o que contribuiu para um valor de coeficiente de correlação ligeiramente superior.

5 Conclusão

O presente trabalho teve por finalidade aumentar a precisão altimétrica dos dados obtidos pelo sensor de SAR e o consequente aumento na quantidade de aplicações, incluindo as na área de engenharia de infraestruturas, de monitoramento ambiental e de estimativa de biomassa florestal.

De forma inovadora, o trabalho contribuiu para esta finalidade por meio:

- da apresentação da premissa de que as regiões da classe de solo exposto devem ter altura interferométrica igual a zero; e
- do desenvolvimento de uma metodologia composta por diversos testes de modelagem de ajuste paramétrico e não paramétrico, baseados em modelos matemáticos polinomiais e logarítmicos. No âmbito da metodologia, foram aplicados os métodos de busca de Gradiente Reduzido Generalizado e de algoritmo genético, com diferentes parâmetros, utilizando o sistema computacional comercial Microsoft Excel Solver.

Os testes não obtiveram os resultados esperados, havendo poucos casos onde o algoritmo convergiu para o objetivo de anular o valor da altura interferométrica em regiões de solo exposto. Apesar das análises efetuadas sobre os resultados, não foi possível identificar o motivo da falta de convergência visto que o sistema comercial não é aberto e não permite a análise do código computacional implementado, o que o torna limitado para este tipo de aplicação.

Visando superar esta limitação, novos estudos serão realizados incluindo o uso de outros sistemas ou métodos não paramétricos de busca de solução, disponíveis comercialmente ou desenvolvidos para atender especificamente este objetivo. Dentre esses métodos, sugere-se o de "subida de colina" e o de "colônia de formigas".

Dentre os testes que obtiveram resultados convergentes, não foi possível melhorar significativamente a estimativa de biomassa florestal por meio dos modelos matemáticos aplicados. Logo, a hipótese de que é possível o ajuste da altura interferométrica utilizando áreas de solo exposto como condição de contorno também não foi comprovada.

Analisando as limitações dos dados utilizados no presente trabalho, foram identificados três fatores que poderão vir a contribuir para novas pesquisas na área em foco.

O primeiro fator se refere à quantidade de parcelas inventariadas de manejo florestal utilizadas como amostras. Uma quantidade superior de parcelas possibilitaria uma representação amostral mais fidedigna e uma maior variação de valores de biomassa na região de estudo, com o consequente aumento de aderência da reta de regressão estatística e do coeficiente de correlação.

O segundo fator está relacionado à dificuldade em observar regiões de solo exposto próximas a parcelas inventariadas de biomassa florestal. Neste caso, sugere-se o uso de dados de inventários florestais que sejam mais próximos de regiões antropizadas e fartas em áreas de solo exposto.

O terceiro fator identificado é a disponibilidade de pontos de controle sobre a área em estudo. Por meio de pontos de controle medidos *in loco* e distribuídos em regiões próximas aos inventários florestais, será possível analisar minuciosamente as variações geométricas dos modelos digitais de SAR submetidos aos modelos matemáticos de ajuste desenvolvidos.

6 Referências

Araújo, T.M.; Higuchi, N. & Junior, J.A.C. 1999. Comparison of formula for biomass content determination in a tropical rain forest in the state of Pará, Brazil. *Forest Ecology and Management, 117*: 43-52.

Burrough, P.A. 1986. *Principles of Geographical Information Systems for land resource assessment.* London, Oxford University Press, 193p.

Felgueiras, C.A. 2004. Modelagem Numérica de Terreno. *In*: CÂMARA, G.; DAVIS, C. & MONTEIRO, A.M.V. (ed.). *Introdução à Ciência da Geoinformação.* Instituto de Pesquisas Espaciais (INPE). Disponível em: <http://www.dpi.inpe.br/gilberto/livro/introd/cap7-mnt.pdf>. Acesso em: 4 dez 2019.

Carpentier, J. & Abadie, J. 1966. Généralisation de la Méthode du Gradient Réduit de Wolfe au cas des Contraintes Non Lineaires. *In*: IV INTERNATIONAL CONFERENCE ON OPERATIONAL RESEARCH. Nova York, 1966.

Castro-Filho, C.A.P.; Freitas, C.C.; Sant'anna, S.J.S.; Lima, A.J.N. & Higuchi, N. 2013. Relating Amazon Forest Biomass to PolInSAR Extracted Features. *In*: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), Melbourne, 2013. Resumos expandidos, Melbourne, p. 2111-2114.

Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 43 - 3 / 2020 p. 110-123

Debastiani, A.B.; Moura, M.M.; Rex, F.D.; Sanquetta, C.R.; Corte, A.P.D. & Pinto N. 2019. Regressões Robusta e Linear para Estimativa de Biomassa via Imagem Sentinel em uma Floresta Tropical. *BIOFIX Scientific Journal, 4:* 81-87.

DSG. 2016. Diretoria de Serviço Geográfico. Especificação Técnica para Produtos de Conjunto de Dados Geoespaciais (ET-PCDG). Disponível em: <http://www.geoportal.eb.mil.br/portal/inde2>. Acesso em: 4 dez. 2019.

DSG. 2019. Diretoria de Serviço Geográfico. Radiografia da Amazônia. Disponível em: <http://www.geoportal.eb.mil.br/portal/index.php/projetos/147-projeto-cartografia-da-amazonia>. Acesso em: 4 dez. 2019.

Gama, F.F. 2007. *Estudo da interferometria e polarimetria SAR em povoamentos florestais de eucalyptus SP*. Programa de Pós-graduação em Sensoriamento Remoto. Instituto Nacional de Pesquisas Espaciais, Tese de Doutorado, 243 p.

Google Earth. 2019. Disponível em: <http://www.google.com.br/intl/pt-BR/earth/>. Acesso em: 22 fev. 2019.

IBGE. 2019. Instituto Brasileiro de Geografia e Estatística. Manuais Técnicos em Geociências, N° 14 – Acesso e Uso de Dados Geoespaciais. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101675>. Acesso em: 4 dez. 2019.

JAXA. 2019. Japanese Space Agency. ALOS Global Digital Surface Model "ALOS World 3D - 30m (AW3D30)". Disponível em: < https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm>. Acesso em 26 de dez. 2019.

Martínez, J.M. & Santos, S.A. 1995. *Métodos computacionais de otimização*. Campinas, Departamento de Matemática Aplicada IMECC-UNICAMP, 262p.

Mouratidis A. 2014. Geographical Information Systems. *In*: ESA SAR COURSE, Malta, 2014. Disponível em: <https://earth.esa.int/documents/10174/1743079/02_Antonios_Mouratidis_GIS_overview.pdf>. Acessado em: 26 dez. 2019.

Neeff, T.; Dutra, L.V.; Santos, J.R.; Freitas, C.C. & Araújo, L.S. 2005. Tropical forest biomass measurement by interferometric height modeling and P-band radar backscatter. *Forest Science, 51*(6): 585–594.

Neter, J.; Kutner, M.H.; Nachtsheim, C.J. & Wasserman, W. 1996. *Applied Linear Statistical Models.* Boston, McGraw-Hill, 1408 p.

Oliveira, M.V.N.D & Locks C.J. 2019. Potencial de uso de SAR aerotransportado para modelagem do terreno e da biomassa acima do solo em região de floresta tropical. *In*: XIX SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, Santos, 2019. Resumo expandido, Santos, INPE, p. 1855-1858. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1113331/potencial-de-uso-de-sar-aerotransportado-para-modelagem-do-terreno-e-da-biomassa-acima-do-solo-em-regiao-de-floresta-tropical>. Acesso em: 26 dez. 2019.

Pope, K.O.; Benayas-Rey, J.M. & Paris, J.F. 1994. Radar remote sensing of forest and wetland ecosystems in the Central American tropics. *Remote Sensing of Environment, 48*(2): 205-219.

Projeto RadamBrasil. 1977. *Folha SA.19 Içá/Amazonas: geologia, geomorfologia, pedologia, vegetação e uso potencial da terra.* Rio de Janeiro, Departamento Nacional da Produção Mineral, 452 p.

Reeves, C. 2010. Genetic Algorithms. *In*: Reeves, C. (ed.). *Handbook of Metaheuristics*. Coventry University, p. 55-82. Disponível em: <https://www.researchgate.net/publication/226462334_Genetic_Algorithms>. Acesso em: 06 dez. 2019.

Saatchi, S.; Halligan, K.; Despain, D.G. & Crabtree, R.L. 2007. Estimation of forest fuel load from radar remote sensing. *IEEE Transactions on Geoscience and Remote Sensing, 45*(6): 1726 – 1740.

Santos, J.R.; Freitas, C.C.; Araújo, L.S.; Dutra, L.V.; Mura, J.C.; Gama, F.F.; Soler, L.S. & Sant'Anna, S.J.S. 2003. Airborne P-band SAR applied to the aboveground biomass studies in the Brazilian tropical rainforest. *Remote Sensing of Environment*, ELSEVIER, *87*: 482-493.

Santoro M. & Cartus O. 2018. Research Pathways of Forest Above-Ground Biomass Estimation Based on SAR Backscatter and Interferometric SAR Observations. *Remote Sensing*, *10*(4): 1-23. Disponível em: <https://www.mdpi.com/2072-4292/10/4/608>. Acesso em: 06 dez. 2019.

Schlund, M.; Erasmi, S. & Scipal, K. 2019. Comparison of Aboveground Biomass Estimation From InSAR and LiDAR Canopy Height Models in Tropical Forests. IEEE Geoscience and Remote Sensing Letters, 17(3): 367-371.

Singh, A. & Singh, K.K. 2017. Satellite image classification using Genetic Algorithm trained radial basis function neural network, application to the detection of flooded areas. *Journal of Visual Communication and Image Representation, 42*: 173–182.

Zhang, H.; Huang M.; Qing X.; Li G. & Tian C. 2017. Bibliometric Analysis of Global Remote Sensing Research during 2010–2015. *International Journal of Geo-information, 16*(3): 87-102.

International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-8, Issue-7; Jul, 2021
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.87.41

Comparison between Quantitative and Qualitative Theme-Feature Forest Biomass Estimation Models built over SAR Data

Carlos Alberto Pires de Castro-Filho[1,2], Edilson Bias[2]

[1]Brazilian Army, Brazil
[2]University of Brasília, Brasília

Received:11 Jun 2021;

Received in revised form: 12 Jul 2021;

Accepted: 21 Jul 2021;

Available online: 30 Jul 2021

***Keywords*— Amazon Forest, Biomass, Machine Learning, Remote Sensing, SAR.**

***Abstract*—** *International organizations are still in need for methodologies that accurately measures forests above ground biomass (AGB). Among the remote sensing technologies, those of Synthetic Aperture Radar (SAR) stands out in the modeling of forest biomass due to their ability to characterize the geometry of the imaged region. The semantic representation, through thematic maps, is one of the main means for the geospatial situational understanding. However, there is a gap of knowledge for models that are built by the analysis of quantitative and qualitative theme-feature in a complementary way. This article aims to develop and compare forest biomass estimation models, through an innovative methodology, over quantitative and qualitative theme-features. To this end, extracted SAR data and specific machine learning (ML) and feature selection techniques are applied for each case. The models developed are based into forest inventories with 128 plots located in two different Brazilian Amazon Forest areas and were built over 231 extracted independent variables. The methodology applied used techniques to categorize numeric data and, afterwards, comparatively evaluate numeric quantitative and categorized qualitative results. The constructions of the models were based on ML algorithms such as Multilayer Perceptron, Suport Vector Machine and Random Forest. The results showed that the different study areas had very different vegetation characteristics, significantly impacting the feature selection and ML algorithms. The different biomes of the Amazon Forest and their respective characteristics demanded specific models and techniques, not fitting into a single pattern. importance.*

I. INTRODUCTION

In 2016 more than 190 countries participated in the 21st United Nations Conference of the Parties on Climate Change (COP-21), held in Paris. This conference aimed to continue the Kyoto Protocol, expired in 2012, and, consequently, to define goals regarding the emission of polluting gases into the atmosphere. Despite the intense work, a legally binding treaty, capable of compelling the international community to cut greenhouse gas emissions, has not been signed. Among the reasons for this failure, one of the highlights was the lack of methodologies that accurately measures these cuts and establishes mechanisms for this reduction [1,2].

According to the United Nations Framework Convention on Climate Change – UNFCCC [3] the Article 3.4 of the Kyoto Protocol requires countries to report annually on changes in carbon stocks associated with forest biomass. The Intergovernmental Panel for Climate Change [4] and [5] states that reports with this information must follow a methodology based on the principles of transparency, consistency, comparability, completeness and accuracy.

However, [2,6-7] states that studies quantifying the carbon cycle between the atmosphere and forests are still needed. [2] points out that 53 to 58% of the carbon cycle comes from forests, therefore, accurate data on forest biomass are essential for many purposes, including subsidizing projects for environmental monitoring and Reducing Emissions from Deforestation and Forest Degradation (REDD +). [1,8] also states that forest biomass should be considered as a source of renewable energy and can be a source of income for national economies when used as carbon credit.

Among the remote sensing technologies, those of Synthetic Aperture Radar (SAR) stands out in the modeling of forest biomass due to their ability to characterize the geometry of the imaged region [1,2,6,8-12]. It also allows the monitoring and the verification of the type, direction, intensity and extent of the degradation in different areas, caused by human influence or by natural forest fires [6,13-16]. Due to the good results obtained by researchers, new projects that aims to use SAR data to estimate biomass are under execution or planning [6]. The Japan Aerospace Exploration Agency (JAXA) project, ALOS PALSAR 2, has been underway since 2014 and is a source of significant data for recent researches [14,17-20].

In Brazil, among the projects that aims to generate SAR images and that can be used in biomass estimation, the Amazon Radiography Project developed by the Geographic Service of the Army (DSG) stands out. By 2022, a total area of 1,800,000 km^2 of the Amazon region will be covered with airborne sensors in the X and P bands [21]. In addition to the 1:50,000 scale mapping, the project also has the potential to generate data to support infrastructure projects and sustainable exploitation of natural resources in the region [22-24].

Due to the large amount of data that can be originated from available SAR sensors, it is necessary to apply techniques that aims to organize and analyze quantitative and qualitative features in an intelligent and automated way [20,25-27]. Machine Learning – ML techniques are able to model knowledge and make associations between different types of quantitative or qualitative information [28-29]. According to [30], the main advantages of ML are accuracy, since the optimal algorithm is selected from the characteristics of the data and the problem to be solved; automation in learning, which adjusts the models according to the success or failure of the results; processing speed; customization, being suitable in any type of problem; and scalability, as they are processes that adapt to data growth.

One of the possible applications in ML is the development of models involving thematic issues and those resulting in qualitative theme-attributes [28-29]. In these cases, the theme-attribute is commonly used for the construction of thematic maps that includes different areas of human geography, from the spatial representation of health and social geography [31-33], to characteristics related to forest biomass stocks [2,12-13, 16-18].The semantic representation, through thematic maps, grows in importance, being one of the main means for the geospatial situational understanding and, consequently, the implementation of public administrations [34-35].

Recent published researches referring to biomass estimation presents ML originated models which output results are quantitative theme-attribute, that is, numerical [1,16,18-19]. However, studies that builds and analyzes simultaneously quantitative and qualitative theme-attributes models were not observed. Therefore, it is necessary researches that seeks to cover this gap of knowledge and that aims at building thematic maps models using, in a complementary way, quantitative and qualitative theme-attributes.

This article aims to develop and compare forest biomass estimation models built over quantitative and qualitative theme-feature based on extracted SAR data. To this end, machine learning and feature selection techniques are specifically selected and applied for each case.

II. METHOD

2.1 Study Area and data

The study areas are located in different geographical regions of the Brazilian Amazon rain forest: São Gabriel da Cachoeira (SGC), a municipality located on the banks of the Rio Negro, in the northwest of the state of Amazonas; and the Unini River Extractive Reserve (Unini River ExRes) located in the Unini River basin, in the municipality of Barcelos. The areas, in white, are highlighted in Figure 1, together with the location of some of the inventoried plots, in green.

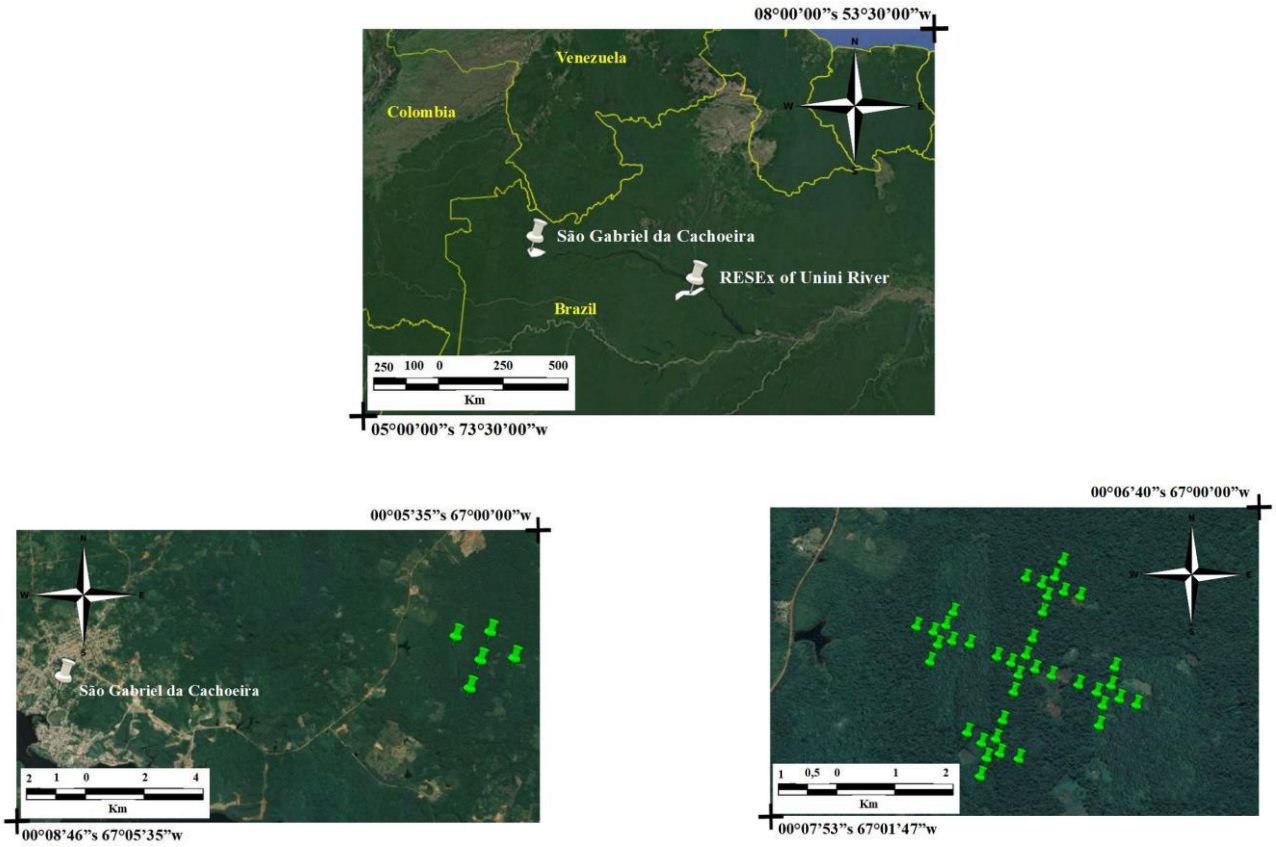


Fig.1:(a) Study areas, highlighted in white; (b) São Gabriel da Cachoeira region; (c) Location of a subset of plots inventoried and arranged in the shape of Maltese Cross.

The areas were selected for two reasons: the distinct phytoecological and land use and occupation situations and the availability of data. The SGC area has hybrid characteristics, composed of anthropized regions together with dense vegetation. In contrast, the Unini River ExRes area is composed only of primary virgin forest vegetation.

According to [31], the vegetation found in the study areas is of forest formation. More specifically, [32] indicates that the vegetation found in the São Gabriel da Cachoeira area is composed by phytoecological forest contact / edaphic formations regions (*campinaranas*). These regions are characterized in three ways:

(1) dense, submontane forests with dissected relief. [32] states that the average AGB volume in the area is 107.4 m^3/ha;

(2) dense, submontane and undulating forests; and

(3) dense forests, lowlands and relief with the presence of plateaus.

The Unini River ExRes, in its turn, is an extractive conservation unit with about 833 hectares in length and characterized in [32] as:

(1) dense tropical forest, referring to the sub-region of the low plateaus of the Amazon; and

(2) areas of ecological tension with dense alluvial presence.

The remote sensing data was obtained from the ALOS PALSAR 2 sensor and the Amazon Radiography Project. The working areas are comprised between 0° and 1° south latitudes and 67° and 68° west longitudes, for the region of São Gabriel da Cachoeira; and between 1° and 2° south latitudes and 62° and 63 ° west longitudes, for the Unini River ExRes.

The data from ALOS PALSAR 2 were provided by IBAMA and are Level 1.1 – Single Look Complex (SLC) processing images in the quadri-polarized strip-map imaging mode.

The Amazon Radiography Project data were provided by the [21] with the following characteristics:

(1) amplitude orthoimages in X band HH polarization and P band quadri-polarized, all with 16 bits radiometric resolution and 5 meters spatial resolution;

(2) digital surface models (DSM) and digital terrain models (DTM) generated, respectively, from the interferometric processing of X and P data, with 32 bits radiometric resolution and 5 meters spatial resolution.

The AGB data were provided by the National Institute of Amazon Researches – INPA, and follow the methods developed by [33] and described by [34]. In addition to the exact same geographical position as the images, the proximity to the region's imaging date was also important as it aims to avoid major changes in the analyzed vegetation.

The given biomass data provided was composed of 128 inventoried plots, 58 plots of São Gabriel da Cachoeira and 70 of Unini River ExRes, presenting the AGB values (ton/ha) and the UTM coordinates of the start and end points of each plot. As pointed out by [35-36], different allometric equations were used to calculate the inventoried plots due to the characteristics of the region. Figure 2 illustrates the format, the start (P1) and end (P2) points and the arbitrary coordinates of each arboreal individual within the plot.

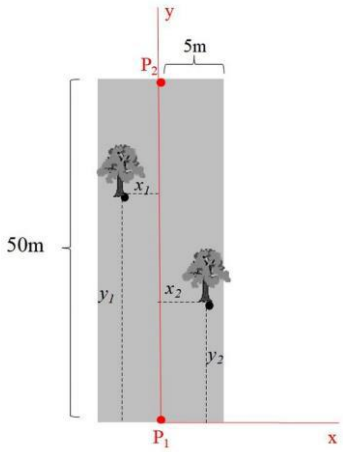


Fig.2:Plot of forest inventory.

Fig.2:Plot of forest inventory.

2.2 Methodological approach

The research was structured according to the flowchart shown in Figure 3. Each step is described in the following subitems.

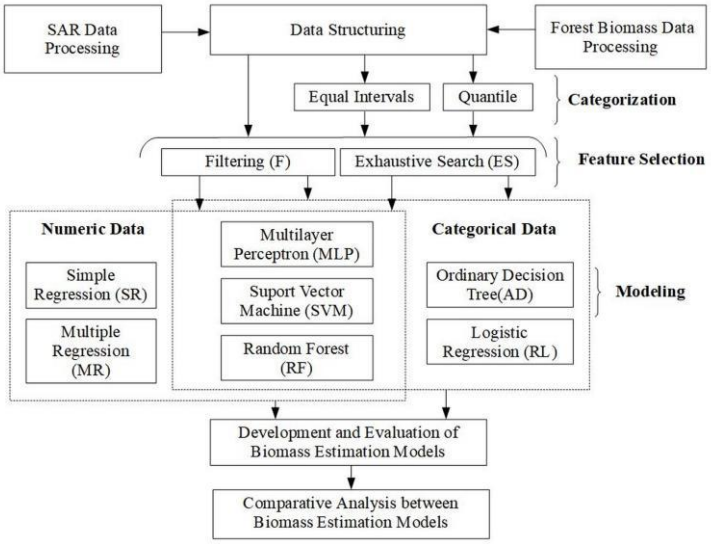


Fig.3: Methodological Flowchart.

2.2.1 Forest Biomass Data Processing

Using analytical geometry techniques, the UTM coordinates of each 4 corners of the inventoried plots were calculated and the respective vector files for each region of interest (ROI) were generated.

2.2.2 SAR Data Processing

In this stage, the ALOS PARSAR 2 images, obtained in SLC format, were processed and the features on the available X, L and P bands were extracted. All processing steps were performed using the Polarimetric SAR Data Processing and Educational Tool (PolSARpro), version 6.0

(Biomass Edition), from the European Space Agency (ESA).

The ALOS PALSAR 2 images were processed according to the flowchart shown in Figure 4. The following parameters were used:

- multilook processing with 2 looks for the rows and 1 look for the columns, as suggested by [19];
- Lee Refined speckle filter with 2 looks and 7x7 size window;
- calculation of the covariance [C] and coherence [T] matrices images, both 3x3;
- geocoding of the coherence matrix image [T], performing the correction of the Range-doppler terrain and the respective georeferencing using the digital elevation model automatically extracted from the Shuttle Radar Topography Mission (SRTM), with 90m spatial resolution;
- polarimetric calibration and conversion to sigma-nought (σ^0) using Equation 1, where the DN is the Digital Number, in amplitude, and CF is the calibration factor in dB for the channels [37]. The value applied for the CF was -83; and
- application of target decomposition techniques.

$$\sigma_0 = 10 * \log_{10}\langle DN^2 \rangle + CF \quad (1)$$

At the end of the SAR data processing, the interferometric, incoherent and coherent features were extracted according to Table 1.

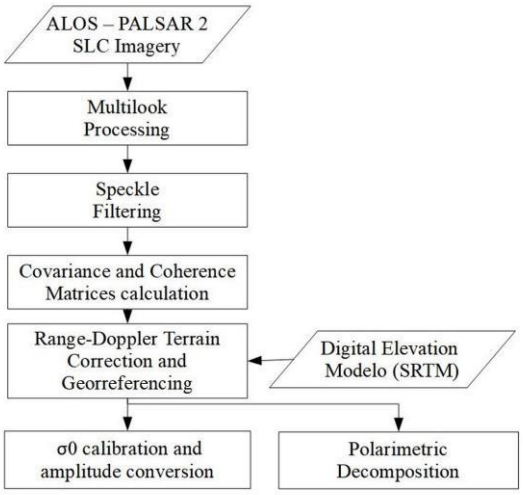


Fig.4: ALOS PALSAR 2 image processing. Adapted from [19]

Table.1: Extracted Features from SAR Data

Symbol	Description
SAR Interferometric Features	
H_{int}	Interferometric height – It is the difference in altitude between the Digital Surface Model (MDS), obtained with the X band, and the Digital Terrain Model (MDT), obtained with the P band. It represents the height of the vegetation.
Decliv	Declivity – It is the slope of the land surface in relation to the horizontal, obtained through the MDT.
Incoherent SAR Features	
Xhh	Amplitude image of the X band in the HH polarization – The backscatter of the forest canopy.
Lhh, Lhv, Lvv	Amplitude image of the L band in the polarizations HH, HV or VV – Represents the main geometric characteristics of arboreal individuals.
Phh, Phv, Pvv	Amplitude image of the P band in the polarizations HH, HV or VV – Associated with the main geometric characteristics of the terrain.
Lhh-Lhv, Lhh-Lvv, Lvv-Lhv	Subtraction between amplitude images in the L band polarizations.
Phh-Phv, Phh-Pvv, Pvv-Phv	Subtraction between amplitude images in the P band polarizations.
PC1L, PC2L, PC3L	Principal Components of the amplitude images in the L bands polarizations.
PC1P, PC2P, PC3P	Principal Components of the amplitude images in the P bands polarizations.
Henderson and Lewis Polarimetric Decomposition Features [38]	
PR_L, PR_P	Ratio between parallel polarizations (*Parallel Ratio – PR*) in the L or P bands (*PR_Band = Band_vv / Band_hh*) – Associated with the orientation and shape of the backscatter elements in the forest.

CR_L, CR_P	Ratio between crossed polarizations (*Crossed Ratio – CR*) in the L or P bands (*CR_Band = Band_hv / Band_hh*) – Referring to the volumetric backscatter of the target.
TotPow_L, TotPow_P	Total power of the L or P bands (*TotPow_Band = Band_hh + Band_vv + 2 * Band_hv*) – They represent the sum of all backscatter mechanisms occurring in the forest.
Pope Polarimetric Decomposition Features [39]	
BMI_L, BMI_P	Biomass index in bands L or P (*BMI_Band* = (*Band_hh* + *Band_vv*) / 2) – Indicator of the amount of woody structure in the forest.
CSI_L, CSI_P	Canopy structure index in the L or P bands (*CSI_Band = Band_vv* / (*Band_vv* + *Band_hh*)) – Compares the vertical structure with the horizontal vegetation.
VSI_L, VSI_P	Volumetric scattering index in the L or P bands (*VSI_Band = Band_hv* / (*Band_hv* + *BMI_Band*)) – Related to the density of the canopy, being directly proportional to the amount of elements that cause multiple type scattering.
Kim and Zyl Polarimetric Decomposition Features [40]	
RVI_L, RVI_P	Radar vegetation index (*RVI_Band* = 8 * *Band_hv* / (*Band_hh* + *Band_vv* + 2 * *Band_hv*)) – Associated with the proportion of vegetation in the soil.
Haralick Textural Features [41] The co-occurrence texture features analyzes the relationship between pixel pairs values within a window and constructs a Grey Level Co-occurence Matrix (GLCM). In the texture equations, *P (i, j)* is the co-occurrence probability of each pixel value in column *i* and row *j*; N_g is the number of distinct grey levels in the quantized image; μ is the average value of *P*; σ is the x or y deviation pattern of the image.	
J_Me_Band	$Me = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} i * P(i,j)$ Mean ( ) value within the GLCM.
J_Va_Band	$Va = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu)^2 P(i,j)$ Variance ( ) value within the GLCM.
J_Ho_Band	Homogeneity ($Ho = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\frac{1}{1+(i-j)^2}$) is the spatial correlation measurement in the GLCM.
J_Con_Band	$Con = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)(i-j)^2$ Contrast ( ) is the intensity difference between the reference pixels and its neighbors in the GLCM.
J_Di_Band	$Di = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\lvert i-j\rvert$ Dissimilarity ( ) is the amplitude difference between the reference pixels and its neighbors in the GLCM.
J_En_Band	Entropy ($En = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\log(P(i,j))$) value represents the randomness between the elements of the GLCM
J_Se_Band	$Se = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)^2$ Second Moment ( ) is the second angular moment between the elements of the GLCM.
J_Cor_Band	Correlation ($Cor = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i,j)P(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$) is the statistical difference between the reference pixels and its neighbors in the GLCM.
Coherent SAR Features	

Cloude and Pottier Polarimetric Decomposition Features [42]	
Alpha	α angle – Dominant type of scattering.
H	Entropy – Proportion in the importance of the dominant type of scattering.
A	Anisotropy – Proportion in the importance of the secondary and tertiary types of scattering.
Freeman and Durden Polarimetric Decomposition Features [43]	
FD_Vol	Volumetric – Contribution of the type of volumetric scattering, simulating the forest canopy.
FD_Dbl	Double Bounce – Result of a set of dihedral corner reflectors.
FD_Odd	Superficial – Contribution of the type of surface scattering.
Touzi Polarimetric Decomposition Features [44]	
TAlfa_S1, TAlfa_S2, TAlfa_S1, TAlfa_Sm	Magnitude (α) - Provides the type of symmetry related to the type of scattering of the target.
TPhi_S1, TPhi_S2, TPhi_S1, TPhi_Sm	Phase (φ) - Represents a more complete characterization of the target's scattering type.
TTau_S1, TTau_S2, TTau_S1, TTau_Sm	Helical angle (τ) - Allows the measurement of the target's degree of symmetry, distinguishing symmetric and asymmetric scattering.
TPsi_S1, TPsi_S2, TPsi_S1, TPsi_Sm	Orientation angle (ψ) - Associated with the target's angle of inclination.
Van Zyl Polarimetric Decomposition Features [45]	
VanZ_Vol	Volumetric Scattering – Volumetric scattering proportion.
VanZ_Dbl	Double Bounce Scattering – Double bounce scattering proportion.
VanZ_Odd	Odd Scattering – Surface (odd) scattering proportion.
Yamaguchi Polarimetric Decomposition Features [46]	
Yam_Vol	Volumetric Scattering – Volumetric scattering proportion.
Yam_Dbl	Double Bounce Scattering – Double bounce scattering proportion.
Yam_Odd	Odd Scattering – Surface (odd) scattering proportion.

2.2.3 Data Structuring

The data extracted from SAR and the AGB data were organized in a single structured spreadsheet, having the features represented in columns and the instances, referring to each inventoried forest biomass plot, as rows. The AGB feature was defined as the theme-feature (or "result" or "output" feature) of the structured spreadsheet.

For each of the extracted features, the arithmetic mean of the pixels' value corresponding to the areas of the inventoried AGB plots was calculated.

The numerical data was used in two different ways. First, using the original values of the explanatory feature set $x = (x_1, x_2, \ldots, x_p)^T$, so that the multiple regression model would be as shown in Equation 2. Second, with the logarithmic of the original value, as Equation 3. In all cases p is the number of variables, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the parameter set, y is the dependent AGB variable and ε is the random error.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_2 + \varepsilon \quad (2)$$

$$ln(y) = ln(\beta_0) + \beta_1 ln(x_1) + \ldots + \beta_p ln(x_p) + \varepsilon \quad (3)$$

2.2.4 Categorization

The numerical data of the AGB quantitative feature were categorized and associated with one of the 5 (five) categories of biomass: "Low", "Medium-Low", "Medium", "Medium-High" and "High". The categorization methods, used to transform quantitative to qualitative features, were of the equal intervals and of the quantile.

According to [47], the method of equal intervals is performed by dividing the theme-feature values in the domain range by the number of categories of interest. In Equation 4, K is the number of categories defined by the user, x_{min} and x_{max}, respectively, the minimum and

maximum values observed in the theme-feature and δ the value of the widths for each category interval.

$$\delta=(x_{max} - x_{min}) /K \quad (4)$$

In the quantile method, categorization is performed by dividing the total number of instances *N* by the number of categories of interest *K*. Therefore, at the end of this method each category will have the same number of objects.

At the end of the categorization stage, the theme-feature was classified in one of three possibilities: numeric (NumThFe), categorical by the "equal intervals" method (EqIntThFe) and categorical by the "quantile" method (QuThFe). Then, all other steps were performed for each of these cases.

2.2.5 Feature Selection

Tests were performed using the filtering type feature selection, in comparison to the exhaustive search including all features extracted from SAR data. The objective was to verify the impacts of the feature selection process on the quality of the final AGB models developed.

The feature selection technique performed was the Correlation-based Feature Subset (CFS) Selection, as described [48]. In this case, the search method used was the greedy Best First, which performs the "hill climb" heuristic in the "forward" direction.

According to [49], the CFS feature selection method is adequate to identify features that are related to the AGB by using the Pearson correlation coefficient method.

2.2.6 Modeling

In the specific cases in which the constructions of the models were based on numerical quntitative data, that is, when the theme-feature has not been categorized, the methods of simple statistical regression – SR and multiple statistical regression – MR were used. On the other hand, for the specific cases of the qualitative categorized data, the methods of logistic statistical regression – LR and ordinary decision tree – ODT were applied.

In addition to these methods, the Multilayer Perceptron – MLP, Suport Vector Machine – SVM and Random Forest – RF methods were used for all cases.

The feature selection and the model development steps were carried out entirely in the WEKA (Waikato Environment for Knowledge Analyzes) system, version 3.8.4, and followed algorithms described by [50].

2.2.7 Development and Evaluation of a Biomass Estimation Model

After the development of the models, the evaluation stage is carried out. In the case of the models based on numerical data, such as those of statistical regression, there are several parameters that can be observed and that reflects the assessment. The parameter used in this case was the correlation coefficient (r), described by [51].

In the case of the models based on categorized qualitative data, the assessment was made by building a confusion matrix and calculating the respective Kappa coefficient of agreement [52]. Due to the reduced number of instances, the process of cross-validation divided into 10 folds was used, as suggested by [53].**2.2.8 Comparative Analysis between Biomass Estimation Models**

Initially, the selected models were those that obtained the best correlation coefficient, in the case of the numerical quantitative data, and best Kappa coefficient, for the models based on categorized qualitative data.

In order to compare those different type of models, the numerical values resulting from the AGB will follow the process described in the flowchart presented in Figure 5. In this process, numerical quantitative values will be categorized using the equal intervals method, followed by the assessment obtained through the construction of the confusion matrices and calculations of the respective Kappa coefficients.

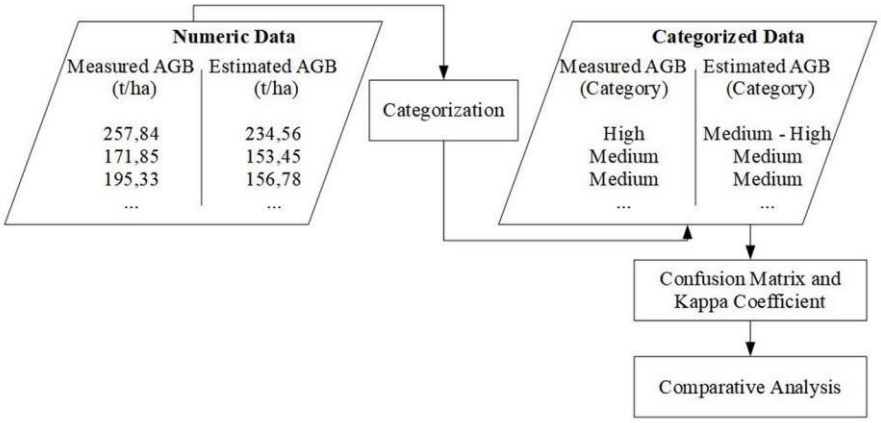


Fig. 5: Categorization process for comparative analysis.

III. RESULTS AND DISCUSSION

3.1 Forest Biomass Data Processing

From the AGB data granted by INPA, 3 sample sets were defined according to the region inventoried: São Gabriel da Cachoeira, Unini River ExRes and the joint regions. The statistics for each set, referring to the number of pixels and AGB in each plot, are shown in Table 2.

Table.2: Statistics for the number of pixels and AGB in the inventoried plots

Set	**Joint Regions**	
Statistics	Number of Pixels (un)	AGB (t/ha)
Mean	50,59	227,93
Minimum	35	92,21
Maximum	72	351,73
Standard Deviation	7,28	45,21
Number of Plots	128 plots	
Set	**São Gabriel da Cachoeira**	
Statistics	Number of Pixels (un)	AGB (t/ha)
Mean	50,17	224,95
Minimum	35	92,21
Maximum	69	351,73
Standard Deviation	8,19	52,24
Number of Plots	58 plots	
Set	**Unini River ExRes**	
Statistics	Number of Pixels (un)	AGB (t/ha)
Mean	50,93	230,40
Minimum	39	153,32
Maximum	72	311,57
Standard Deviation	6,48	38,65
Number of Plots	70 plots	

3.2 SAR Data Processing

Together with the features detailed in Table 1, the textural features were extracted for all available polarimetric bands, that is, Xhh, Phh, Phv, Pvv, Lhh, Lhv and Lvv, for 3x3, 5x5 and 7x7 window sizes.

At the end of the SAR data processing, 231 features, or independent variables, were extracted, in addition to the theme-feature.

3.3 Categorization

The categorization by the equal intervals technique obtained a δ of 52 (t / ha). Therefore, the AGB categories were defined as: Low (below 100 t/ha); Medium-Low

(between 100 and 200 t/ha); Medium (between 200 and 250 t/ha); Medium-High (between 250 and 300 t/ha); and High (above 300 t/ha). The number of categorized instances was 2 (two) for the Low class, 38 (thirty-eight) for Medium-Low, 42 (forty-two) for Medium, 40 (forty) for Medium-High and 6 (six) for High.

The categorization by the quantile method obtained 25 (twenty-five) or 26 (twenty-six) instances for each category.

3.4 Feature Selection

The process was carried out separately for numerical quantitative and categorized qualitative data. The results of the 5 (five) selected features, in decreasing order of relevance, are shown in Table 3. In the same table Pearson's correlation values between the selected feature and the respective theme-feature, quantitative or qualitative, was calculated.

In general, the selected features showed low correlation with the biomass theme-feature. The highlight was the H_{int} feature, which achieved a good correlation with the quantitative data, in addition to being selected for both cases.

Table.3: Result of the feature selection process

Quantitative Data		Qualitative Data	
Feature	**Correlation**	**Feature**	**Correlation**
H_{int}	0.449975	PC3	0.1765
Lhh	-0.188703	H_{int}	0.1592
CSI_L	-0.046255	TAlphaS3L	0.1059
FreeOddL	0.125393	7x7_Xhh_Se	0.2772
TPhiS1L	0.10413	7x7_Phh_Me	0.2851

3.5 Development of Biomass Estimation Models

The ML techniques applied in the biomass estimation modeling had the following specific configurations:

(1) SVM – the model applied to numerical quantitative data was the SMOreg, specific for statistical regression, as described by [54]. The complexity parameter c was 1.0 and the Radias Basis Function (RBF) kernel used 0.01 gamma;

(2) MLP – the models not submitted to the feature selection process were built with one (composed of 50 nodes) or two (composed of 50 and 10 nodes) hidden layers. The models submitted to the feature selection process were built with one (composed of 5 nodes) or two (composed of 5 and 5 nodes) hidden layers;

(3) RF – the parameter of 100 trees was used in the construction of the model;

(4) ODT – the minimum quantity of 2 instances per node was applied.

The correlation and kappa coefficients resulting from the tests are shown in Tables 4, 5, 6 and 7 and have the following characteristics:

(1) Tables 4 and 5 refers to models based on numerical quantitative and Tables 6 and 7 to models based on categorized qualitative theme-features;

(2) Tables 4 and 6 refer to the original values and Tables 5 and 7 refer to log values of the features ;

(3) the values before the bars (/) are those obtained by models that have not been submitted to the feature selection process, while the values after the bars are those referring to models with selected features;

(4) the results in MLP models with an asterisk (*) are those obtained with 2 (two) hidden layers and that obtained results superior to those of a single hidden layer;

(5) the results in bold are the best obtained, having been highlighted 2 (two) results for each type of region and for each type of data (quantitative or qualitative).

Table.4: Correlation coefficients of AGB estimation models for numerical quantitative theme-feature and original feature values.

ML Technique	Joint Regions	São Gabriel da Cachoeira	Unini River ExRes
SR	0.42 /0.42	0.39 /0.39	0.35 /**0.43**
MR	0.21 /0.40	0.02 /0.41	0.04 /0.38
SVM	0.12 /0.21	0.13 /0.13	0.35 /0.12
MLP	0.07 /0.32*	0.12 /**0.70**	0.13 /0.23
RF	0.16 /0.39	0.21 /0.33	0.14 /0.29

Table.5: Correlation coefficients of AGB estimation models for numerical quantitative theme-feature and logarithmic feature values.

ML Technique	Joint Regions	São Gabriel da Cachoeira	Unini River ExRes
SR	0.49 /**0.54**	0.49 /**0.58**	0.30 /0.30
MR	0.09 /0.41	0.04 /0.25	0.01 /0.31
SVM	0.20 /0.22	0.16 /0.10	0.29 /0.06
MLP	0.33 */**0.49**	0.26 */0.52*	0.06 /**0.36***
RF	0.14 /0.39	0.14 /0.47	0.19 /0.25

Table.6: Kappa index of AGB estimation models for categorized qualitative theme-features and original feature values.

ML Technique	Joint Regions		São Gabriel da Cachoeira		Unini River ExRes	
Categorization Method	Equal Intervals	Quantile	Equal Intervals	Quantile	Equal Intervals	Quantile
LR	0.10 /0.22	0.22 /0.15	0.25 /0.10	0.20 /0.10	0.18 /0.35	0.30 /0.33
MLP	0.22 /**0.38**	0.32 /0.15	0.18 /0.02	0.13 /0.07	0.31 /0.29	0.14 /0.19
SVM	0.09 /0.01	0.04 /0.01	0.01 /0.01	0.01 /0.01	0.25 /0.01	0.10 /0.01
ODT	0.09 /0.19	0.11 /0.11	0.09 /0.01	0.04 /0.01	0.22 /**0.48**	0.27 /0.21
RF	0.13 /0.28	0.19 /0.25	**0.30** /0.16	0.24 /0.01	0.19 /0.38	0.26 /0.28

Table.7: Kappa index of AGB estimation models for categorized qualitative theme-features and logarithmic feature values.

ML Technique	Joint Regions		São Gabriel da Cachoeira		Unini River EsRes	
Categorization Method	Equal Intervals	Quantile	Equal Intervals	Quantile	Equal Intervals	Quantile
LR	0.23 /0.23	0.21 /0.18	0.21 /0.24	0.26 /0.12	0.20 /0.35	0.28 /0.31
MLP	**0.36** /0.24	0.18 /0.17	**0.30** /0.12	0.22 /0.16	0.36 /**0.47**	0.28 /0.32
SVM	0.05 /0.01	0.05 /0.01	0.01 /0.01	0.02 /0.01	0.01 /0.01	0.06 /0.01
ODT	0.11 /0.22	0.18 /0.12	0.07 /0.08	0.08 /0.03	0.21 /0.39	0.18 /0.32
RF	0.24 /0.22	0.22 /0.20	0.26 /0.11	0.26 /0.06	0.24 /0.39	0.31 /0.30

3.6 Comparative Analysis between Biomass Estimation Models

As observed in Tables 4, 5, 6 and 7, in general, there was an emphasis on MLP and SR techniques, corresponding to 58% and 25% of the highlighted results, respectively. MR, RF and ODT techniques achieved results close to the best, however, with a single highlight. The SVM technique

showed results significantly lower than the other techniques.

In the case of the numerical quantitative theme-feature, presented in Tables 4 and 5, only the MLP and SR techniques showed outstanding results. The MR technique was not able to increase the r from the input of new features.

The models developed for the categorized qualitative theme-feature, Tables 6 and 7, showed an increase in results for non-parametric techniques, including MLP, RF and ODT.

The models submitted to the feature selection process showed improvement in 73% of the numerical quantitative theme-feature cases. In these cases, only 10% worsened the results, all of which refers to the SVM technique.

On the other hand, for the case of categorized qualitative theme-feature submitted to the feature selection process, the percentages of improvement, worsening and maintenance of the results were, respectively, 35%, 10% and 55%. In this case, there was no correlation to the ML technique.

Regarding the categorization method, all the best results were obtained using the method of equal intervals. Despite this, considering all cases, there was not a conclusive difference in the results between the categorization methods.

The different areas analyzed also presented different results. For the case of the numerical quantitative theme-feature, the São Gabriel da Cachoeira region obtained the best results, unlike the region of the Unini River ExRes with the worst results. The opposite result was obtained for the case of the categorized qualitative theme-feature. In both cases, the results for the joint regions, as they aggregate data from both study areas, were average.

In order to carry out the comparative analysis, the process shown in Figure 5 was applied. The comparative analysis was performed on data from the same regions (Joint Regions, SGC or Unini River ExRes), separately for quantitative or qualitative data. The results obtained are shown in Tables 8, 9, 10, 11, 12 and 13. In all cases, 3 (three) types of Z hypothesis tests were performed, with a significance level (α) of 0.05:

In order to carry out the comparative analysis, the process shown in Figure 5 was applied. The comparative analysis was performed on data from the same regions (Joint Regions, SGC or Unini River ExRes), separately for quantitative or qualitative data. The results obtained are shown in Tables 8, 9, 10, 11, 12 and 13. In all cases, 3 (three) types of Z hypothesis tests were performed, with a significance level (α) of 0.05:

- test to analyze the hypothesis of Kappa * (value referring to the first selected model) being equal to zero;
- test to analyze the hypothesis of Kappa ** (value for the second selected model) to be equal to zero;
- and test to analyze the hypothesis whether the difference between Kappa * and Kappa ** is significantly greater (or lower) than zero, that is, if both are significantly different.

Table.8: Comparative analysis between confusion matrices: numerical quantitative theme-feature of the joint region.

	SR over logarithmic values (r=0.54)*					**MLP over logarithmic values (r=0.49)****				
	Reference					**Reference**				
Categorized	**Category**	**Low**	**Medium -Low**	**Medium**	**Medium-High**	**High**	**Low**	**Medium -Low**	**Medium**	**Medium -High**	**High**
	Low	0	0	0	0	0	0	1	0	0	0
	Medium-Low	2	6	2	4	0	2	6	1	0	0
	Medium	0	9	21	9	3	0	8	18	13	3
	Medium-High	0	1	1	1	1	0	1	5	1	0

	High	0	0	0	0	0	0	0	0	0	1
	Kappa*: 0.17; Kappa Variance*: 0.0057 Global Acuracy*: 47%						Kappa**: 0.13; Kappa Variance**: 0.0073 Global Acuracy**: 43%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=2.25; p-value=0.0123; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is significantly higher than zero (z=2.25; p-value=0.0123; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is significantly higher than zero (z=2.25; p-value=0.0123; α=0.05)

Table.9: Comparative analysis between confusion matrices: numerical quantitative theme-feature, from SGC.

Categorized	**MLP over original values (r=0.70)***						**RS over logarithmic values (r=0.58)****				
		Reference						**Reference**			
	Category	**Low**	**Medium-Low**	**Medium**	**Medium-High**	**High**	**Low**	**Medium-Low**	**Medium**	**Medium-High**	**High**
	Low	0	1	0	0	0	0	0	0	0	0
	Medium-Low	2	5	1	0	0	2	5	2	4	0
	Medium	0	10	18	5	0	0	10	17	8	3
	Medium-High	0	0	3	9	1	0	1	3	2	1
	High	0	0	0	0	3	0	0	0	0	0
	Kappa*: 0.42; Kappa Variance*: 0.0082 Global Acuracy*: 60%						Kappa**: 0.11; Kappa Variance**: 0.0064 Global Acuracy**: 41%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=4.68; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is not significantly higher than zero (z=1.41; p-value=0.0798; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is significantly higher than zero (z=2.57; p-value=0.0050; α=0.05)

Table.10: Comparative analysis between confusion matrices: numerical quantitative theme-feature, from Unini River ExRes.

	RS over original values (r=0,43)*						MLP over logarithmic values (r=0,36)**				
	Reference						Reference				
Categorized	Category	Low	Medium-Low	Medium	Medium-High	High	Low	Medium-Low	Medium	Medium-High	High
	Low	0	0	0	0	0	0	0	0	0	0
	Medium-Low	0	0	1	0	0	0	0	0	0	0
	Medium	0	16	17	18	0	0	15	18	7	0
	Medium-High	0	0	0	6	2	0	1	0	13	1
	High	0	0	0	0	0	0	0	0	4	1
	Kappa*: 0.10; Kappa Variance*: 0.0029 Global Acuracy*: 38%						Kappa**: 0.33; Kappa Variance**: 0.0046 Global Acuracy**: 53%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=1.89; p-value=0.0295; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is significantly higher than zero (z=4.85; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is significantly lower than zero (z=-2.62; p-value=0.0045; α=0.05)

Table.11: Comparative analysis between confusion matrices: categorized qualitative theme-feature, from the joint region

	MLP over original values*						MLP over logarithmic values**				
	Reference						Reference				
Categorized	Category	Low	Medium-Low	Medium	Medium-High	High	Low	Medium-Low	Medium	Medium-High	High
	Low	0	0	0	0	0	2	0	0	0	0
	Medium-Low	2	25	12	4	1	0	18	10	10	0
	Medium	0	7	23	13	2	0	13	24	6	1
	Medium-High	0	6	7	23	1	0	7	8	24	2

	High	0	0	0	0	2	0	0	0	0	3
	Kappa*: 0.38; Kappa Variance*: 0.0039 Global Acuracy*: 57%						Kappa**: 0.36; Kappa Variance**: 0.0042 Global Acuracy**: 55%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=6.00; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is significantly higher than zero (z=5.60; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is not significantly different than zero (z=0.19; p-value=0.4255; α=0.05)

Table.12: Comparative analysis between confusion matrices: categorized qualitative theme-feature, from SGC.

Categorized	**RF over original values***					**MLP over logarithmic values****				
	Reference					**Reference**				
	Category	**Low**	**Medium-Low**	**Medium**	**Medium-High**	**High**	**Low**	**Medium-Low**	**Medium**	**Medium-High**	**High**
	Low	2	0	0	0	0	1	0	1	0	0
	Medium-Low	0	10	7	4	0	0	8	8	2	0
	Medium	0	5	14	6	3	1	5	10	4	0
	Medium-High	0	1	1	4	1	0	3	3	7	1
	High	0	0	0	0	0	0	0	0	1	3
	Kappa*: 0.30; Kappa Variance*: 0.0088 Global Acuracy*: 52%						Kappa**: 0.30; Kappa Variance**: 0.0091 Global Acuracy**: 50%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=3.16; p-value=0.0008; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is significantly higher than zero (z=3.20; p-value=0.0007; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is not significantly different than zero (z=-0.06; p-value=0.4762; α=0.05)

Table.13: Comparative analysis between confusion matrices: categorized qualitative theme-feature, from Unini River ExRes.

		ODT over original values*					*MLP* over logarithmic values**				
		Reference					Reference				
	Category	Low	Medium-Low	Medium	Medium-High	High	Low	Medium-Low	Medium	Medium-High	High
Categorized	Low	0	0	0	0	0	0	0	0	0	0
	Medium-Low	0	12	5	4	0	0	14	8	4	0
	Medium	0	3	14	3	0	0	4	9	2	0
	Medium-High	0	7	1	17	0	0	4	3	20	0
	High	0	0	0	2	2	0	0	0	0	2
		Kappa*: 0.48; Kappa Variance*: 0.0069 Global Acuracy*: 64%					Kappa**: 0.47; Kappa Variance**: 0.0071 Global Acuracy**: 64%				

Analysis:

Hypothesis Z-Test: Kappa* = 0

Kappa is significantly higher than zero (z=5.76; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa** =0

Kappa** is significantly higher than zero (z=5.61; p-value=0.0000; α=0.05)

Hypothesis Z-Test: Kappa*- Kappa**=0

Kappa*- Kappa** is not significantly different than zero (z=0.08; p-value=0.4697; α=0.05)

From the analysis of the results presented in the tables, it is observed that the kappa values obtained by the post-modeling categorization process (Tables 8, 9 and 10), in general, were lower than those obtained in the pre-modeling categorization process (Tables 11, 12 and 13). In both cases, the ML techniques built specific models for quantitative or qualitative data, suffering loss of accuracy in the transformation process between these types of data.

Due to the loss of accuracy in the post-modeling categorization process, the best results obtained are shown in Table 13, with insignificant difference in the kappa values for the ODT (Kappa = 0.48) and MLP (Kappa = 0.47).

The values obtained by the Kappa coefficient, in addition to serving as parameters for comparison between the categorizations, can also be evaluated, being classified in different linguistic intervals, according to their level of agreement, as shown in Figure 6. In this case, according to [55], the best results obtained in this research are classified as *moderate.*

The *moderate* results obtained may have occurred for several reasons, including: the quantity of biomass samples; the sampling distribution of biomass values; and the low correlation between the biomass theme-feature and extracted the extracted features. Regarding the latter, Table 3 shows the low correlation, including on the selected features.

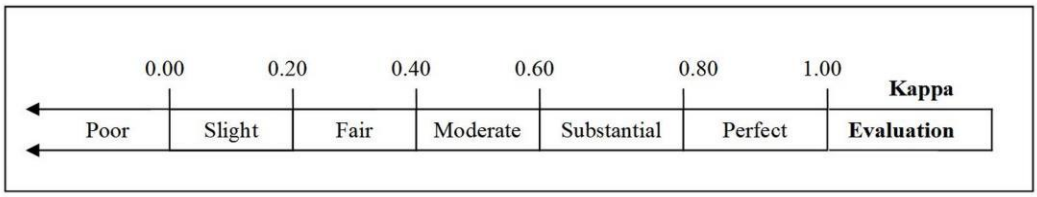


Fig.6: Linguistic evaluation of Kappa coefficient values. Adapted from [55].

IV. CONCLUSION

The present work aimed to develop and compare forest biomass estimation models, from different regions of the Amazon forest, built over numerical quantitative or categorical qualitative theme-feature. For this, ML techniques were applied on polarimetric and interferometric X, L and P bands SAR data extracted features, generating models that were analysed and compared.

In an innovative way, the work presents a methodology that involves:

- the process of feature selection and AGB estimation models development over quantitative and qualitative theme-features. It is noteworthy that, for each case, the feature selection and ML techniques were specific and configured in order to obtain the best results;
- comparative analyses between quantitative and qualitative results. In this case, the post-modeling categorization process and the respective confusion matrices construction was performed, followed by the comparison using hypothesis tests.

The results showed that the different study areas had very different characteristics, significantly impacting the feature selection and ML algorithms. The SGC area, due to the greater variation in AGB inventoried values (between 92.21 and 351.73 t/ha), obtained better results with the numeric quantitative theme-features. On the other hand, Unini's River ExRes area, that had AGB values with less variation (between 153.32 and 311.57 t/ha), was better suited to categorized qualitative data modelling.

The different biomes of the Amazon Forest and their respective characteristics demanded specific models and techniques, not fitting into a single pattern. This conclusion is in agreement with the research of [2] who affirms that the heterogeneity of tropical forests is one of the main factors for the increasing uncertainty regarding the biomass stocks measurement in the region.

The process of feature selection was unanimous in selecting the interferometric height (H_{int}) as the most relevant feature for all areas of study, both in the case of qualitative and quantitative theme-features, in agreement with the results obtained by [23-24,56-57]. Likewise, there was an emphasis on features obtained by target decomposition techniques on the L band, from the ALOS PALSAR 2 sensor. The textural features, on the other hand, did not show significant correlation with the AGB values, different from the results obtained by [58].

As a conclusion of the presented methodology, there was no significant improvement in the AGB estimation process, since the results obtained from Kappa varied between *fair* and *moderate*. Likewise, the post-modeling categorization process did not achieve the expected results, keeping the Kappa value stable and not being able to generalize the AGB values into categories. The result obtained may have occurred due to the low correlation between the biomass theme-feature and the extracted SAR features.

In order to develop more suitable AGB models for different regions of the Amazon Forest, further studies will be carried out aiming to adjust the training parameters of ML techniques. In this case, the possibility of applying search methods and deep learning, commonly used in the Artificial Intelligence area to define such parameters, will be verified.

Analysing the possible reasons that led to the limited results, two factors were identified that may contribute to new research in the area in focus.

The first factor refers to the inventoried forest management plots used as samples. In agreement with the quoted by [59-65], a large number of plots, including areas with greater variations of AGB values, allows a more reliable sample representation and more in-depth statistical analysis.

The second factor is related to the processing of SAR data and the possibility of extracting new polarimetric and interferometric features. Accessing data in SLC format of polarimetric X and P bands would enable the extraction and analysis of the respective target decomposition features. Likewise, through the construction of a digital elevation model in the L band, it would be possible to obtain new interferometric heights involving the

differences between the X-L and L-P bands and the corresponding analyzes.

REFERENCES

[1] Sinha, S., Jeganathan, C., Sharma, L.K., Nathawat, M.S (2015). A review of radar remote sensing for biomass estimation. International Journal of Environment Sciences Technologies. https://doi.org/10.1007/s13762-015-0750-0

[2] Erb, K.H., Kastner, T., Plutzar, C., Bais, A.L.S., Carvalhais, N., Fetzel, T., Gingrich, S., Haberl, H., Lauk, C., Niedertscheider, M., Pongratz, J., Thurner, M., Luyssaert, S (2018). Unexpectedly large impact of forest management and grazing on global vegetation biomass. Nature, 553, 73–76. https://doi.org/10.1038/nature25138

[3] UNFCCC – United Nation Framework Convention on Climate Change (2008). Kyoto Protocol Reference Manual on Accounting of Emissions and Assigned Amounts.

[4] IPCC – Intergovernmental Panel on Climate Change (2003). Good practice guidance for land use, land-use changes and forestry.

[5] Köhl, M., Lasco, R., Cifuentes, M., Jonsson, Ö., Korhonen, K.T., Mundhenk, P., de Jesus Navar, J., Stinson, G (2015). Changes in forest production, biomass and carbon: Results from the 2015 UN FAO. Global Forest Resource Assessment. Forest Ecology and Management, 352, 21–34. https://doi.org/10.1016/j.foreco.2015.05.036

[6] Ho Tong Minh, D., Le Toan, T., Rocca, F., Tebaldini, S., Villard, L., Réjou-Méchain, M., Phillips, O.L., Feldpausch, T.R., Dubois-Fernandez, P., Scipal, K., Chave, J (2016). SAR tomography for the retrieval of forest biomass and height: Cross-validation at two tropical forest sites in French Guiana. Remote Sensing of Environment, 175, 138–147. https://doi.org/10.1016/j.rse.2015.12.037

[7] Houghton, R.A., Nassikas, A.A (2017). Global and regional fluxes of carbon from land use and land cover change 1850–2015. Global Biogeochem. Cycles, 31, 456–472. https://doi.org/10.1002/2016GB005546

[8] Kumar, L., Sinha, P., Taylor, S., Alqurashi, A.F (2015). Review of the use of remote sensing for biomass estimation to support renewable energy generation. Journal of Applied Remote Sensing, 9, 1–29. https://doi.org/10.1117/1.jrs.9.097696

[9] Beaudoin, A., Le Toan, T., Goze, S., Nezry, E., Lopes, A., Mougin, E., Hsu, C.C., Han, H.C., Kong, J.A., Shin, R.T (1994). Retrieval of forest biomass from SAR data. International Journal of Remote Sensing, 15, 2777–2796. https://doi.org/10.1080/01431169408954284

[10] Furtado, L.F. de A., Silva, T.S.F., Novo, E.M.L. de M. Dual-season and full-polarimetric C band SAR assessment for vegetation mapping in the Amazon várzea wetlands (2016). Remote Sensing of Environment, 174, 212–222. https://doi.org/10.1016/j.rse.2015.12.013

[11] Ningthoujam, R.K., Balzter, H., Tansey, K., Feldpausch, T.R., Mitchard, E.T.A., Wani, A.A., Joshi, P.K (2017). Relationships of S-band radar backscatter and forest aboveground biomass in different forest types. Remote Sensing, 9, 1–17. https://doi.org/10.3390/rs9111116

[12] Debastiani, A.B., Moura, M.M., Rex, F.E., Sanquetta, C.R., Corte, A.P.D., Pinto, N. Regressões Robusta e Linear para Estimativa de Biomassa Via Imagem Sentinel em uma Floresta Tropical (2019). BIOFIX Science Journal, 4, 81–87. https://doi.org/10.5380/biofix.v4i2.62922

[13] Saatchi, S., Harris, N.L., Lefsky, M., Brown, S., Mitchard, E.T.A., Salas,W., Zutta, B.R., Buermann,W., Lewis, S.L., Hagen, S., Petrova, S., White, L., Silman, M. & Morel, A (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. Proceedings of the National Academy of Sciences, USA, 2011, 108: 9899–9904.

[14] Huang, W., Sun, G., Ni, W., Zhang, Z., Dubayah, R (2015). Sensitivity of multi-source SAR backscatter to changes in forest aboveground biomass, in: International Geoscience And Remote Sensing Symposium (IGARSS), Melbourne. https://doi.org/10.3390/rs70809587

[15] Treuhaft, R., Lei, Y., Gonçalves, F., Keller, M., dos Santos, J.R., Neumann, M., Almeida, A. Tropical-forest structure and biomass dynamics from TanDEM-X radar interferometry. Forests, 2017, 8, 277–294. https://doi.org/10.3390/f8080277

[16] Le Noë, J., Matej, S., Magerl, A., Bhan, M., Erb, K.H., Gingrich, S (2020). Modeling and empirical validation of long-term carbon sequestration in forests (France, 1850–2015). Global Change Biology, 26, 2421–2434. https://doi.org/10.1111/gcb.15004

[17] Avtar, R., Suzuki, R., Sawada, H (2014). Natural Forest Biomass Estimation Based on Plantation Information Using PALSAR Data. PLOS ONE, 9 (1). https://doi.org/10.1371/journal.pone.0086121

[18] Berninger, A., Lohberger, S., Stängel, M., Siegert, F (2018). SAR-based estimation of above-ground biomass and its changes in tropical forests of Kalimantan using L- and C-band. Remote Sensing, 10, 831–853. https://doi.org/10.3390/rs10060831

[19] Pereira, L.O., Furtado, L.F.A., Novo, E.M.L.M., Sant'Anna, S.J.S., Liesenberg, V., Silva, T.S.F (2018). Multifrequency and Full-Polarimetric SAR assessment for estimating above ground biomass and leaf area index in the Amazon Várzea Wetlands. Remote Sensing, 10, 1–23. https://doi.org/10.3390/rs10091355

[20] Camargo, F.F., Sano, E.E., Almeida, C.M., Mura, J.C., Almeida, T (2019). A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images. Remote Sensing, 11, 1600–1616. https://doi.org/10.3390/rs11131600

[21] DSG – Diretoria de Serviço Geográfico (2008). Contratação de Serviços de Aerolevantamento na Região Amazônica e Processamento de Dados com Radares de Abertura Sintética Aerotransportados Interferométricos. Mapping Project.

[22] Santos, J.R., Freitas, C.C., Araujo, L.S., Dutra, L. V., Mura, J.C., Gama, F.F., Soler, L.S., Sant'Anna, S.J.S (2003). Airborne P-band SAR applied to the aboveground biomass studies in the Brazilian tropical rainforest. Remote Sensing of Environment, 87, 482–493. https://doi.org/10.1016/j.rse.2002.12.001

[23] Neeff, T., Dutra, L.V., Dos Santos, J.R., Da Costa Freitas, C., Araujo, L.S (2005). Tropical forest measurement by interferometric height modeling and P-band radar backscatter. Forest Science, 51, 585–594. https://doi.org/10.1093/forestscience/51.6.585

[24] Gama, F.F., Mura, J.C., De Albuquerque, P.C.G., Dos Santos, J.R (2010). Avaliação do potencial da interferometria sar para o mapeamento altimétrico de áreas reflorestadas por eucalyptus sp. Boletim de Ciências Geodésicas. https://doi.org/10.1590/s1982-21702010000400003

[25] Del Frate, F., Solimini, D (2004). On neural network algorithms for retrieving forest biomass from SAR data. IEEE Transation Geoscience on Remote Sensing, 42, 24–34. https://doi.org/10.1109/TGRS.2003.817220

[26] Englhart, S., Keuck, V., Siegert, F (2012). Modeling aboveground biomass in tropical forests using multi-frequency SAR data – a comparison of methods. IEEE Journal Selected Topics on Applied Earth Observation Remote Sensing, 5, 298–306. https://doi.org/10.1109/JSTARS.2011.2176720

[27] Wylie, B.K., Pastick, N.J., Picotte, J.J., Deering, C.A (2019). Geospatial data mining for digital raster mapping. GIScience Remote Sensing. https://doi.org/10.1080/15481603.2018.1517445

[28] Quinlan, J.R., (1993). C4.5: Programs for Machine Learning, Machine Learning Kluwer Academic Publishers, Boston, Manufactured in The Netherlands. Morgan Kaufmann, California.

[29] Ng, A., (2018). Machine Learning Yearning: Technical Strategy for AI Engineers in the Era of Deep Learning [Draft Version], deeplearning.ai.

[30] Brink, H.B., Richard, J.W., Fetherolf, M. (2015). Real-World Machine Learning, MEAP Editi. ed, Book. Manning Publication, New York.

[31] Cavalcante, J. R. & Abreu, A. J. L (2020). COVID-19 in the city of Rio de Janeiro: spatial analysis of first confirmed cases and deaths. Epidemiologia Serviço de Saúde, Brasília, 29(3):e2020204, 2020. doi: 10.5123/S1679-49742020000300007.

[32] Pardo, I. F., Napoletano, B. M., Verges, F. R., Billa, L (2020). Spatial analysis and GIS in the study of COVID-19: a review. Science of The Total Environment, 739. https://doi.org/10.1016/j.scitotenv.2020.140033.

[33] Fatima, M., O'Keefe, K. J., Wei, W., Arshad, S., Gruebner, O (2021). Geospatial analysis of COVID-19: a scoping review. International Journal of Environment Res Public Health, 18 (5):2336. DOI: https://doi.org/10.3390/ijerph18052336.

[34] Mooney, Peter & Juhász, Levente (2020). Mapping COVID-19: How web-based maps contribute to the infodemic. Dialogues in Human Geography, 10, 265-270. 10.1177/2043820620934926.

[35] Li, R (2021). Visualizing COVID-19 information for public: Designs, effectiveness, and preference of thematic maps. Human Behavior & Emerging Technology, 3, 97– 106. https://doi.org/10.1002/hbe2.248.

[36] Mapbiomas, 2019. Mapeamento Anual da Cobertura e Uso do Solo no Brasil (MapBiomas). Mapeamento Anu. da Cober. e Uso do Solo no Bras. URL http://mapbiomas.org/map#coverage (accessed 6.14.19).

[37] RadamBrasil, 1977. Geologia, geomorfologia, pedologia, vegetação e uso potencial da terra. Mapping Project.

[38] Higuchi, N., Santos, J. dos, Ribeiro, R.J., Minette, L., Biot, Y (1998). Biomassa da parte aérea da vegetação da Floresta Tropical úmida de terra-firme da Amazônia Brasileira. Acta Amazon, 28, 153–166. https://doi.org/10.1590/1809-43921998282166

[39] Silva, R.P. (2007). Alometria, estoque e dinâmica da biomassa de florestas primárias e secundárias na região de Manaus (AM). National Institute for Space Research (INPE). PhD Thesis.

[40] Araújo, T.M., Higuchi, N., Junio, J.A. de C (1999). Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Para, Brazil. Forest Ecology and Management, 117, 43–52. https://doi.org/10.1016/S0378-1127(98)00470-8

[41] Lima, A.J.N., Suwa, R., De Mello Ribeiro, G.H.P., Kajimoto, T., Dos Santos, J., Da Silva, R.P., De Souza, C.A.S., De Barros, P.C., Noguchi, H., Ishizuka, M., Higuchi, N (2012). Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro, Brazil. Forest Ecology and Management, 277, 163–172. https://doi.org/10.1016/j.foreco.2012.04.028

[42] Woodhouse, I.H., 2017. Introduction to Microwave Remote Sensing, Introduction to Microwave Remote Sensing. Taylor & Francis Group CRC Press, Florida. https://doi.org/10.1201/9781315272573

[43] Henderson, F.M., Lewis, A.J., 1998. Manual of remote sensing: principles and applications of imaging radars, 3rd ed. ed. John Wiley and Sons, New York.

[44] Pope, K.O., Rey-Benayas, J.M., Paris, J.F (1994). Radar remote sensing of forest and wetland ecosystems in the Central American tropics. Remote Sensing of Environment, 48, 205–219. https://doi.org/10.1016/0034-4257(94)90142-2

[45] Kim, Y., Van Zyl, J.J (2009). A time-series approach to estimate soil moisture using polarimetric radar data. IEEE Transaction Geoscience Remote Sensing, 47, 2519–2527. https://doi.org/10.1109/TGRS.2009.2014944

[46] Haralick, R., Shanmugam, K., Dinstein, I (1973). Textural Features for Image Classification. IEEE Transaction System Man Cybernetics, 3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314

[47] Cloude, S.R., Pottier, E (1996). A review of target decomposition theorems in radar polarimetry. IEEE Transactions Geoscience and Remote Sensing, 34. https://doi.org/10.1109/36.485127

[48] Freeman, A., Durden, S.L (1998). A three-component scattering model for polarimetric SAR data. IEEE Transaction Geoscience Remote Sensing, 36, 963–973. https://doi.org/10.1109/36.673687

[49] Touzi, R (2007). Target scattering decomposition in terms of roll-invariant target parameters. IEEE Transaction

Geoscience Remote Sensing, 45, 73–84. https://doi.org/10.1109/TGRS.2006.886176

[50] Van Zyl, J.J., (1992). Application of Cloude's target decomposition theorem to polarimetric imaging radar data, in: Procedings Society of Photo-Optical Instrumentation Engineers. pp. 184–212. https://doi.org/10.1117/12.140615

[51] Yamaguchi, Y., Yajima, Y., Yamada, H (2006). A four-component decomposition of POLSAR images based on the coherency matrix. IEEE Geoscience Remote Sensing Letters, 3, 292–296. https://doi.org/10.1109/LGRS.2006.869986

[52] Dent, B., Torguson, J., Hodler, T. (2008). Cartography: thematic map design, 6th ed., Cartographic Perspectives. McGraw-Hill Science, New York.

[53] Hall, M. a., Smith, L. a. (1998). Practical feature subset selection for machine learning, Computer Science. Hamilton, New Zealand.

[54] Yu, X.; Ge, H.; Lu, D.; Zhang, M.; Lai, Z.; Yao, R (2019). Comparative Study on Variable Selection Approaches in Establishment of Remote Sensing Model for Forest Biomass Estimation. Remote Sensing, 11, 1437. https://doi.org/10.3390/rs11121437

[55] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). Data Mining: Practical Machine Learning Tools and Techniques, 2nd Editio. ed, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Massachusetts. https://doi.org/10.1016/c2009-0-19715-5

[56] Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (2004). Applied Linear Statistical Models, 5th Edition. ed, Journal of Education. McGraw-Hill, Boston.

[57] Congalton, R.G., Green, K. (2013). Assessing the Accuracy of Remotely Sensed Data Principles and Practices (Second Edition), CRC Press Taylor & Francis Group, Boca Raton, London, New York.

[58] Sileshi, G. W (2014). A critical review of forest biomass estimation models, common mistakes and corrective measures. Forest Ecology and Management, 329, 237-254. https://doi.org/10.1016/j.foreco.2014.06.026.

[59] Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K (1999). Improvements to the SMO algorithm for SVM regression. IEEE Transaction Neural Networks, 11, 1188–1193. https://doi.org/10.1109/72.870050

[60] Landis, J.R., Koch, G.G (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977, 33, 159–174. https://doi.org/10.2307/2529310

[61] Castro-Filho, C.A.P. de, Freitas, C.D.C., Sant'Anna, S.J.S., Lima, A.J.N., Higuchi, N. (2013). Relating Amazon forest biomass to PolInSAR extracted features, in: International Geoscience and Remote Sensing Symposium (IGARSS). Melbourne. https://doi.org/10.1109/IGARSS.2013.6721320

[62] Schlund, M., Erasmi, S., Scipal, K (2020). Comparison of Aboveground Biomass Estimation from InSAR and LiDAR Canopy Height Models in Tropical Forests. IEEE Geoscience Remote Sensing Letters, 17, 367–371. https://doi.org/10.1109/LGRS.2019.2925901

[63] Sarker, M. L. R., Nichol, J., Iz, H. B., Ahmad, B. B., Rahman (2012), A. A. Potential of texture measurements of two-date dual polarization PALSAR data for the improvement of forest biomass estimation. ISPRS Journal of Photogrammetry and Remote Sensing, 69, 146-166. https://doi.org/10.1016/j.isprsjprs.2012.03.002.

[64] Clark, D.B., Kellner, J.R (2012). Tropical forest biomass estimation and the fallacy of misplaced concreteness. Jornal Vegetation Science, 23, 1191-1196. https://doi.org/10.1111/j.1654-1103.2012.01471.x

[65] Santoro, M., Cartus, O (2018). Research pathways of forest above-ground biomass estimation based on SAR backscatter and interferometric SAR observations. Remote Sensing, 10, 1–23. https://doi.org/10.3390/rs10040608

Gmail Carlos Alberto Pires de Castro Filho <carlos.pires.1976@gmail.com>

Manuscript submission to the International Journal of Remote Sensing - Manuscript ID TRES-PAP-2021-1210

1 mensagem

International Journal of Remote Sensing <onbehalfof@manuscriptcentral.com> 28 de outubro de 2021 15:23

Responder a: IJRS-Administrator@dundee.ac.uk
Para: carlos.pires.1976@gmail.com
Cc: carlos.pires.1976@gmail.com, edbias@unb.br

28-Oct-2021

Dear Mr. Carlos Castro-Filho
(cc'd to co-authors, if any)

Your manuscript entitled "Categorization Optimization in the Construction of Thematic Products" has been successfully submitted online and is presently being given full consideration for publication in International Journal of Remote Sensing.

Your manuscript ID is TRES-PAP-2021-1210.

Please mention the above manuscript ID in all future correspondence. If there are any changes in your contact details, please log in to the International Journal of Remote Sensing - ScholarOne Manuscripts site at https://mc.manuscriptcentral.com/tres and edit your user account information as appropriate.

You can also view the status of your manuscript at any time by checking the appropriate folder in your "Corresponding Author Centre" after logging in to https://mc.manuscriptcentral.com/tres.

The journal to which you are have submitted to is participating in the PEER project. This project, which is supported by the European Union EC eContentplus programme(http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm), aims to monitor the effects of systematic self-archiving (author deposit in repositories) over time. If your submission is accepted, and you are based in the EU, you may be invited to deposit your accepted manuscript in a repository as part of this project. The project will develop models to illustrate how traditional publishing systems may coexist with self-archiving For further information please visit the PEER project website at http://www.peerproject.eu.

Thank you for submitting your manuscript to the International Journal of Remote Sensing.

Yours sincerely

Mrs Catherine Murray
Administrator, International Journal of Remote Sensing
IJRS-Administrator@Dundee.ac.uk

Gmail Carlos Alberto Pires de Castro Filho <carlos.pires.1976@gmail.com>

Editor assigned for IJRS Research Paper TRES-PAP-2021-1210

1 mensagem

International Journal of Remote Sensing <onbehalfof@manuscriptcentral.com> 8 de novembro de 2021 11:17

Responder a: IJRS-Editor-in-Chief@dundee.ac.uk
Para: carlos.pires.1976@gmail.com

08-Nov-2021

Dear Mr. Castro-Filho

Your IJRS Research Paper TRES-PAP-2021-1210, entitled "Categorization Optimization in the Construction of Thematic Products", with you as the corresponding author, was recently submitted to the International Journal of Remote Sensing.

It has now been assigned to the following Editor:-
Prof. Michael Collins
mjcollin@ucalgary.ca

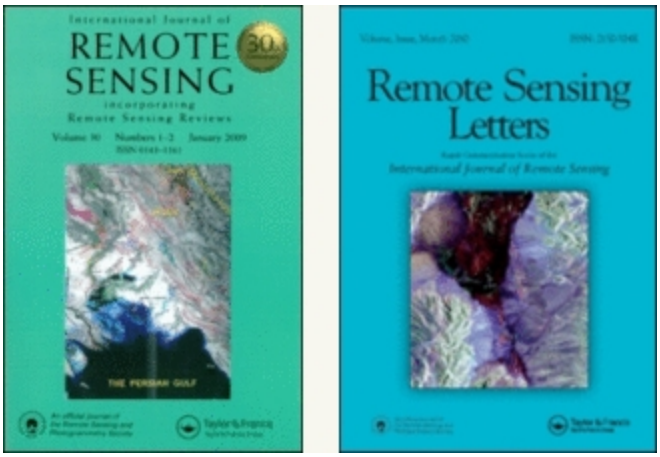
If you have any queries, please feel free to contact that Editor, or the Journal Administrator:-
Mrs Catherine Murray
IJRS-Administrator@Dundee.ac.uk

You can, of course, find out the status of your paper at any time by logging in to the International Journal of Remote Sensing - ScholarOne Manuscripts website at https://mc.manuscriptcentral.com/tres

Thank you for your interest in our journal.

Yours sincerely

Prof. Kevin Tansey
Editor-in-Chief, International Journal of Remote Sensing
IJRS-Editor-in-Chief@Dundee.ac.uk



Categorization Optimization in the Construction of Thematic Products

Journal:	*International Journal of Remote Sensing*
Manuscript ID	Draft
Manuscript Type:	IJRS Research Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Castro-Filho, Carlos; University of Brasilia, Bias, Edilson; University of Brasilia
Keywords:	SAR, BIOMASS, thematic mapping
Keywords (user defined):	categorization, search heuristic

Categorization Optimization in the Construction of Thematic Products

Carlos Alberto Pires de Castro-Filho and Edilson Bias

Institute of Geosciences, University of Brasília (UnB), Brasília, Brazil

carlos.pires.1976@gmail.com; edbias@unb.br

Categorization Optimization in the Construction of Thematic Products

Abstract: One of the steps in the construction of thematic products is the categorization. This process is done by partitioning the range of values of a given feature into several subranges and associating a new value, in this case ordinal, to all instances that are in that subrange. The objective of this work is to propose an innovative pre-categorization process that applies a computational search method to maximize the accuracy of thematic products during the classification stage. The theme feature used is the *A*bove *G*round *B*iomass (AGB), which estimation model is built over synthetic aperture radar features. A system is developed in Java script using the Weka data mining class library. The proposed heuristic is the hill climbing greedy, with the Kappa coefficient as the objective function. The results obtained shows that the proposed Categorization Optimization algorithm demonstrated the ability to obtain new states with subintervals of categories that increased the Kappa agreement index to 1.0 with much lower computational cost than the exhaustive search. The thematic products constructed maintained the representativeness of the study area while increasing in thematic accuracy.

Keywords: categorization; search heuristic; thematic products; biomass; SAR

1. Introduction

In 2016 more than 190 countries participated in the 21st United Nations Conference of the Parties on Climate Change (COP-21), held in Paris. This conference aimed to continue the Kyoto Protocol, expired in 2012, and, consequently, to define goals regarding the emission of polluting gases into the atmosphere. Despite the intense work, a legally binding treaty, capable of compelling the international community to cut greenhouse gas emissions, has not been signed. Among the reasons for this failure, one of the highlights was the lack of methodologies that accurately measures these cuts and establishes mechanisms for this reduction (Erb et al. 2018; Sinha et al. 2015).

According to the United Nations Framework Convention on Climate Change – UNFCCC (2008) the Article 3.4 of the Kyoto Protocol requires countries to report annually on changes in carbon stocks associated with forest biomass. The Intergovernmental Panel for Climate Change – IPCC (2003) and Köhl et al. (2015) states that reports with this information must follow a methodology based on the principles of transparency, consistency, comparability, completeness and accuracy.

Thematic products aims to represent geographic phenomena, physical or social, of the Earth's surface. Sometimes they are built from basic geoinformation, adding new layers of geoinformation to products referring to a specific theme, ie, thematic geoinformation (Raposo et al. 2020).

Originally thematic products are presented in the form of maps, which follow the language and grammatical rules defined by Bertin in publications from the 1970s (Bertin 1977). However, currently thematic products are primarily intended to support decision-making by managers (Wu et al. 2019) as they provide useful information for spatial, urban and rural planning, change monitoring and ecosystem state assessments (Mitchell 2018). With a focus on this objective, such products are presented in different ways, going beyond the classic representations of thematic maps, to queries in online WebGIS systems, with presentations in dashboards and the possibility of consumption via web online services.

One of the current highlights are products related to forest biomass stocks (Erb et al. 2018; Debastiani et al. 2019; Le Noe et al. 2020) that can support different types of programs, such as Reducing Emissions from Deforestation and Forest Degradation (REDD+) and the launch of new orbital sensors such as the European Space Agency's Biomass Mission (Scipal et al. 2010).

In Brazil, the continuous monitoring program based on shallow deforestation data, called PRODES (PRODES 2013) and the alert system for detecting deforestation in near real time, DETER (Diniz et al. 2015), both coordinated by the National Institute for Space Research (INPE), are examples of programs consolidated in the 1980s that provide online thematic products to support the fight against illegal activities in the Amazon rainforest in various formats, including raster, vector and summary tables.

Currently, more advanced Web GIS systems, which performs temporal analysis of spatial data with continental coverage, such as TerraBrasilis (Assis et al. 2019), have their own infrastructure with monitoring, filtering, cataloging and product availability services themed. The Map Biomass (2021), in turn, aims to prepare and make available thematic products specific to Brazilian biomes through the annual supply of land cover and use maps, image mosaics and statistics.

One way to build thematic products is from remote sensing data, through classification methods (Mitchell 2018; Foody 2021). Critically, there are assumptions in the classification process that can degrade the accuracy of the product. One of these assumptions is that class labels and characterizations are rigid factors, that is, that they are defined before the start of the classification process and that they should not be changed during its execution. In these cases there are no interaction steps with the user.

However, regardless of the classification method adopted, without an accuracy assessment it is not possible for the user to decide whether it is fit for the purpose (Foody 2021). Wu (2019) highlights the importance of highly accurate and precise thematic maps, with focus on assessing the suitability and sustainability of prepared land for the planting of agricultural crops, with a strong impact on the value of the respective commodities and on the regional economy.

Furthermore, the process of generalizing the data used in building the thematic products, aiming to represent the phenomenon of interest, is one of the sources of error in the final product (Costa et al. 2017; Sluter et al. 2018). The nature and magnitude of this error, which deviates the product from reality, will vary depending on the method used (Foody 2021). For this reason, the generalization model should be selected in such a way that it seeks to degrade the accuracy of the final product as little as possible (Mitchell 2018).

The generalization process, aiming at developing the thematic products in the form of maps, takes place on features that can be divided into groups, according to the characteristics of their data. According to Yang and Webb (2009) features can be qualitative or quantitative.

Qualitative features have data that do not support arithmetic operations. Bussab and Morettin (2017) states that a qualitative feature is nominal when there is no ordering in the possible realizations and ordinal when these realizations can be sorted.

Quantitative features supports arithmetic operations and can be discrete, when derived from a count, or continuous, when derived from a measurement (Yang and Webb 2009). Bussab and Morettin (2017) also adds that discrete features are those that have the possible values within a finite set. On the other hand, continuous features have possible values belonging to a range of real numbers.

Transformation processes between different types of features have specific names in the literature. Transforming continuous quantitative features into discrete quantitative features is called discretization. Discretization is done by partitioning the range of values of a given feature into several subranges and associating a specific new discrete value to all instances that belong to each subrange. Where A is a feature that has ordered values $(x_1,...,x_i,...,x_j,...,x_m,...,x_N)$, $1 < i < j < m < N$, referring to the N

instances and K the number of subintervals, the discretization process D can be represented by Equation 1.

$$D(A):\{ [x_1;x_i]_1 , [x_{i+1};x_j]_2 , \dots , [x_m;x_N]_K \} \quad (1)$$

In addition to enabling the execution of some algorithms related to feature selection and data classification methods, other reasons for discretization are the increase in computational speed and the interpretability of the generated classification models, in addition to the decrease in the amount of data stored (Garcia et al. 2013; Bueno-Crespo et al. 2018; Rosenfeld et al. 2018). Both benefits are often seen, for example, in decision tree classification methods that process previously discretized data.

Several researches were carried out seeking to evaluate or propose discretization techniques (Garcia et al. 2013). Among the most recent reviews, Maslove et al. (2013) assess discretization techniques on clinical data. They conclude that supervised discretization techniques are more suitable for specific cases, while unsupervised techniques are more general.

Additionally, Bueno-Crespo et al. (2018) presents a fuzzy discretization proposal where each instance has a degree of belonging to different categories. In this case, the unsupervised algorithm demonstrated superiority to the traditional K-means technique.

Rosenfeld et al. (2018) highlights the fact that discretized features, when used together with the original numerical ones, contributes to the classification process. One of the explanations for this contribution lies in the fact that the feature selection process is implicit to the discretization process, which reduces the effect of the dimensionality curse in the development of predictive models.

Although the discretization process has computational and methodological advantages, improving the accuracy of the classification model in several studies, Rajbahadur et al. (2021) warn of the noise that can be generated due to this transformation. The authors emphasize that these noises are not safe for the classification process, generating models from samples that do not represent the characteristics, in general, of the region of interest.

Another transformation process between different types of features is categorization. Categorization is defined as the process that generates ordinal qualitative features, which can be derived from discrete or continuous quantitative features (Dent et al. 2008). The categorization process is similar to the discretization process, with the only difference that all instances belonging to each subinterval will receive a specific categorical value and no longer a numerical value. Along with this, the categorical values received must be possible to order, all belonging to the same theme.

The categorization of a given feature has wide application in the generation of thematic products. In this case it is necessary that the quantitative feature used as the theme is categorized so that the generated categories have different representations. The feature that will be categorized to be used as a theme will be called the theme-feature.

As an example of a product that went through the categorization process in order to generate thematic product, the Thematic Map for Estimating the Prevalence of Schistosomiasis was developed by Martins-Bedê et al. (2009). In this case, the theme feature has percentage continuous quantitative values which were associated with one of the three prevalence categories: Low (0.5%], Medium (5%,15%] or High (15%,100) %].

In a more recent research, Castro-Filho and Bias (2021) presents comparative analyzes between classifications carried out on numerical and categorized data. In this

specific case, the authors sought to develop a biomass estimation model through the use of machine learning techniques on Synthetic Aperture Radar (SAR) data.

The categorization can also be carried out in two moments of the construction of the thematic product, that is, through two different methods: in the final stage, when the product is ready, having all the quantitative values of each area or pixel; or in the step prior to the generation of the classification model. When categorization is performed in the final step, it first captures the range of values of the theme feature by analyzing all the values of the existing elements in the already constructed product. After that, the subintervals to which each element will be associated are constructed. Finally, each subinterval will be assigned a categorical name. This process will be called post-categorization.

On the other hand, categorization can be performed in the step prior to the generation of the map classification model. In this case, the elements of the thematic map do not have values yet, and it is necessary to define a classification model based on training samples. It is on these samples that the categorization process is carried out, which is called pre-categorization.

The objective of this work is to propose a pre-categorization process that applies a computational search method to maximize the accuracy of thematic products during the classification stage. The theme feature used is Above Ground Biomass (AGB), which estimation model is built on SAR features described by Castro-Filho and Bias (2021). This objective is of great importance since the quality observed in the result of a classification process directly influences the thematic quality of products that use this type of data in their construction.

2. Materials and Methods

2.1. Study Area and data

The study areas are located in different geographical regions of the Brazilian Amazon rain forest: São Gabriel da Cachoeira (SGC), a municipality located on the banks of the Rio Negro, in the northwest of the state of Amazonas; and the Unini River Extractive Reserve (Unini) located in the Unini River basin, in the municipality of Barcelos. The areas, in white, are highlighted in Figure 1.

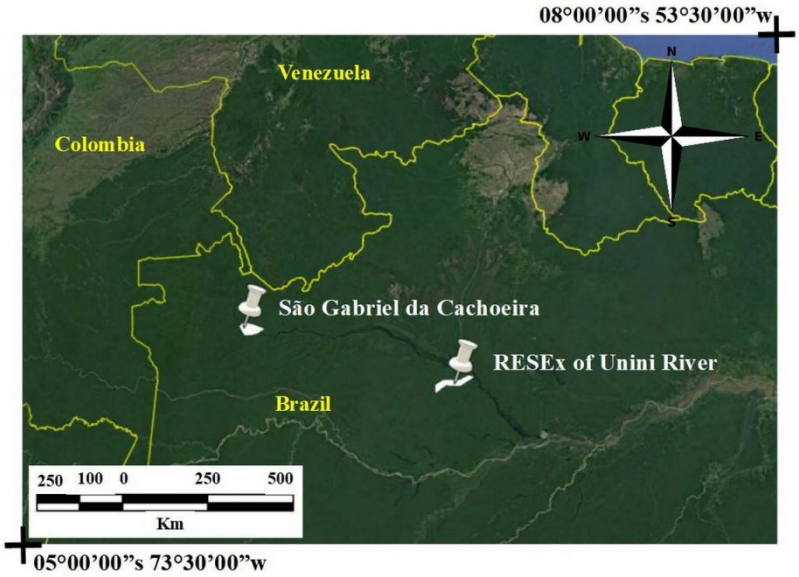


Figure 1. Location of the study areas, highlighted in white (Google Earth, 2021).

According to Mapbiomas (2021), the vegetation found in the study areas is of forest formation. More specifically, RadamBrasil (1977) indicates that the vegetation found in the São Gabriel da Cachoeira area is composed by phytoecological forest contact / edaphic formations regions (*campinaranas*). These regions are characterized in three ways:

(1) dense, submontane forests with dissected relief. RadamBrasil (1977) states that the average AGB volume in the area is 107.4 m^3/ha;

(2) dense, submontane and undulating forests; and

(3) dense forests, lowlands and relief with the presence of plateaus.

Unini, in its turn, is an extractive conservation unit with about 833 hectares in length and characterized in RadamBrasil (1977) as:

(1) dense tropical forest, referring to the sub-region of the low plateaus of the Amazon; and

(2) areas of ecological tension with dense alluvial presence.

The remote sensing data was obtained from the ALOS PALSAR 2 sensor and the Amazon Radiography Project. The working areas are comprised between 0° and 1° south latitudes and 67° and 68° west longitudes, for the region of São Gabriel da Cachoeira; and between 1° and 2° south latitudes and 62° and 63 ° west longitudes, for Unini.

The AGB data were provided by the National Institute for Research in the Amazon - INPA, and follows the methods developed by Higuchi et al. (1998) and described by Silva (2007). The biomass data provided were composed of 128 inventoried plots, 58 plots from SGC and 70 from Unini, presenting the AGB values (ton / ha) and the UTM coordinates of the initial and final points of each plot.

Details about the characteristics and processing of the AGB and the SAR data, as well as the methodological approach in the construction of the structured spreadsheet used in the modeling, are described by Castro-Filho and Bias (2021). The present work will seek to focus on the proposal of the categorization optimization method and the building of the thematic map.

2.2. Categorization Optimization Algorithm

The proposed algorithm for categorization optimization follows the heuristic as shown in the flowchart in Figure 2.

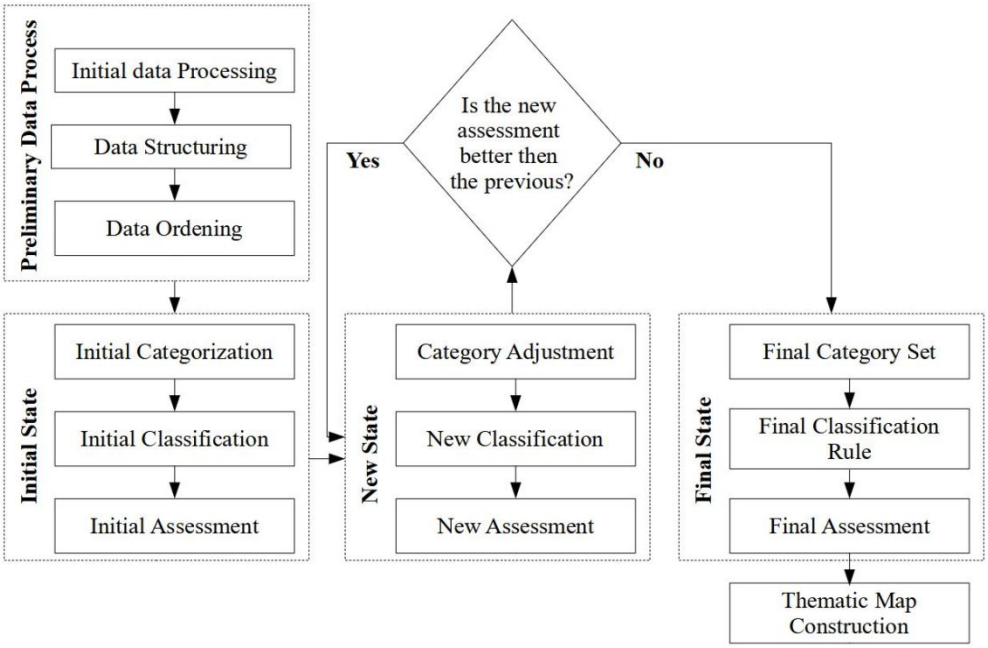


Figure 2. Methodological flowchart of the Categorization Optimization algorithm.

The steps referring to the preliminary data process are described by Castro-Filho and Bias (2021). In general, this step aims to build a spreadsheet where the data is structured in lines, referring to instances, and columns, for the extracted features. Next, the data were sorted using the Quicksort technique applied to the AGB theme-feature. According to Theodoridis and Koutroumbas (2006), this ordering technique has computational complexity in the order of O(Nlog$N$), where N is the total number of instances of the theme feature.

The initial categorization, referring to the initial state, was obtained through the classic data categorization method of equal intervals. Liu et al. (2002) describe this method as unsupervised for categorization since it only takes into account the values of the theme feature, despite the values of the other features.

According to Dent et al. (2008), in the equal intervals method, categorization is performed by dividing the value of the theme feature domain range by the number of classes of interest. Subranges will be obtained, which will be the categories, of equal size. With K being the number of user-defined categories and x_{min} and x_{max}, respectively, the minimum and maximum values observed in the theme feature, then the method defines categories with widths equal to that calculated by Equation 2.

$$\delta=(x_{max} - x_{min}) /K \quad (2)$$

Therefore, categorization by the equal intervals method Cat_{EqInt} over the theme feature A_{theme} will generate the set of categories according to Equation 3.

$$Cat_{EqInt}(A_{theme}):\{[x_{min}, x_{min}+\delta] ,]x_{min}+\delta , x_{min}+2\delta], \dots ,]x_{min}+(K-1)\delta , x_{max}]\} \quad (3)$$

Based on the initial categorization, the initial classification is performed using the method of Ordinary Binary Decision Tree (OBDT), version C4.5, due to the computational simplicity and ease of interpretation on the generated models (Hastie et al. 2009). As it is a non-parametric process, the method does not have requirements regarding the type and statistical distribution of the training data, being able to deal with continuous and discrete geoinformation simultaneously (Wu et al. 2019).

Next, the initial assessment is performed by the construction of the confusion matrix and the Kappa coefficient of agreement, using the leave-one-out cross-validation as test sample data. This procedure aimed to evaluate the ability to build the classification model from the parameters defined by the user in the process.

After the creation of the *initial state*, the proposed Categorization Optimization algorithm is started through a search process with the variation in the values of the category divisors, in such a way that different subintervals can be generated. Each set of subintervals is called a *new state* (Russel and Norvig 2004).

New states are defined in such a way that, if an instance is on the boundary between two categories, this instance is removed from the current one and inserted in the adjacent one. In this way, each of the (K-1) category dividers, where K is the number of categories, is moved to the left or to the right, varying the limits between adjacent categories. The *new states* are then considered the *current state* of the algorithm, from which it is then possible to generate 2(K-1) different *new states* to be analyzed.

The proposed search method to optimize the categorization process is the *hill climbing*, using a greedy algorithm. This method is widely used in the field of Artificial Intelligence to search for optimal local values (Russel and Norvig 2004). For the applied hill climbing method, firstly an objective function is defined, which is wanted to be maximized. After this definition, the value of this function is checked in the *current state*, where the search is located, and in the *neighboring states*. If the value of the objective function in one of the *neighboring states* is higher than the *current state*, the algorithm performs a step to the *neighboring state*, making it the *current state* and restarting the entire search process.

Since the algorithm is greedy, the steps performed during the search will always be for the *neighboring state* that has the highest objective function value, ignoring the other states. In addition, the algorithm does not perform any further steps if none of the *neighboring states* has a value higher than the *current state*, ending the search process (Theodoridis and Koutroumbas 2006).

As Russell and Norvig (2004) describe, the hill climbing method has the disadvantage of being "stuck" in local maximums and plateaus, failing to reach the global maximum. However, aiming to overcome this problem, the proposed algorithm uses a *floating step*, which will have the size defined by the user. Such floating step will

only be performed by the algorithm if a more distant state have better results. In case the floating step is performed, the hill climbing algorithm restarts performing simple steps again, seeking to maximize the objective function.

The objective function defined for the proposed algorithm is the Kappa coefficient of agreement obtained by validating the result of a classification. It is observed that, as a consequence of this objective function, for each *search state* a process of categorization, classification and classification validation will be carried out in order to obtain the Kappa value. The pseudo-code of the algorithm is shown in Figure 3.

```
1: algorithm optimalCategorization
2:     input    structuredTable.txt
3:              K number of categories
4:              n minimum number of instances per category
5:              γ floating step
6:     sortedStructuredTable ← Quicksort (struturedTable, themeFeatureIndex)
7:     initialCategorizedState ← EqualIntervalCategorization (sortedStructuredTable)
8:     initialClassification ← Classification (initialCategorizedState)
9:     initialEvaluation ← Evaluation(initialClassification)
10:    finalCategorizedState ← initialCategorizedState
11:    finalEvaluation ← initialEvaluation
12:    for each K-1 category divider do
13:        if numberOfInstancesInPreviousCategory > n do
14:            newCategorizedState ← TransferIntanceFromPreviousCategory
15:            newClassification ← Classification (newCategorizedState)
16:            newEvaluation ← Evaluation (newClassification)
17:            if newEvaluation > finalEvaluation do
18:                finalCategorizedState ← newCategorizedState
19:                finalEvaluation ← newEvaluation
20:                goto line 13
```

```
21:             FloatingStep (finalCategorizedState, γ)
22:             if numberOfInstancesInNextCategory > n do
23:                     newCategorizedState ← TransferIntanceFromNextCategory
24:                     newClassification ← Classification (newCategorizedState)
25:                     newEvaluation ← Evaluation (newClassification)
26:                     if newEvaluation > finalEvaluation do
27:                             finalCategorizedState ← newCategorizedState
28:                             finalEvaluation ← newEvaluation
29:                             goto line 20
30:             FloatingStep (finalCategorizedState, γ)
31:     return finalCategorizedState
32:     return finalEvaluation
```

Figure 3. Categorization Optimization algorithm pseudo-code.

According to Bueno-Crespo et al. (2018), the proposed technique is classified as:

- dynamic, because the training phase of the machine learning technique is carried out together;
- local and univariate, the process being carried out only on the theme-attribute;
- incremental, as it starts with data already categorized;
- hybrid, the categories can be divided or grouped during the process;
- based on the evaluation method of the generated classification model;
- crisp for having rigid values of division between categories; and
- based on stopping criteria, as it is a greedy algorithm.

As input parameters of the proposed algorithm, the user must inform:

(1) the number of K categories you want when searching for the optimal categorization;

(2) the minimum number of instances per category;

(3) the γ value of the floating step of the search.

It is also observed that the minimum number of instances per category limits the generation of *new states* to be analyzed. If the number of instances in a given category is lower than that informed by the user, that state will be considered invalid by the algorithm.

The Categorization Optimization algorithm was developed in the JAVA programming language in order to use the classes available in the WEKA data mining system (Waikato Environment for Knowledge Analyzes), version 3.8.4 (Witten et al. 2016).

After analyzing the complexities of all functions involved in the proposed algorithm, it is concluded that the heuristic complexity is $O(K^2N^2)$, where K is the number of categories and N is the number of instances.

2.3 Tests and parameters

The Categorization Optimization algorithm tests were performed with the following characteristics:

(1) separately for each study area, SGC and Unini;

(2) applied on the original values of the features and on the values transformed to logarithmics, as suggested by Gama et al (2010);

(3) based on two different scenarios, ie K values for 3 or 5 categories, with biomass labels as {Low, Medium, High} or {Low, Medium-Low, Medium, Medium-High, High};

(4) minimum number of instances per category equal to 2, aiming to seek the lowest possible restriction in the performance of steps between states; and

(5) floating step value equal to 4.

In order to analyze the search capability of the proposed heuristic, an exhaustive search algorithm was also developed to evaluate all the state space possibilities. Through this search, it is possible to follow the categorization status and assessment towards the maximum global value of the Kappa coefficient and compare it with that obtained by the proposed heuristic.

At the end of the process, two different ways of analyzing the results are carried out. The first analysis aims to verify the representativeness of the thematic map built in the proposed study area, that is, in specific Amazonian biomes. For this purpose, 10 (ten) selected sample areas are homogeneously distributed throughout the region, with dimensions between 40 and 50 thousand pixels. Next, the variations in the occurrence of the categories contained in the training samples are compared with the occurrences in the selected sample areas.

Additionally, a comparative analysis is made between the thematic maps generated through the Categorization Optimization, proposed heuristic, and the categorization performed by the equal intervals method, a classic process. This analysis takes place through the construction of a confusion matrix involving the pixels of both maps.

3. Results and discussion

The results obtained are presented in Tables 1 to 8, which differ according to the characteristics of the study area, the type of feature value and the number of categories. In each table are the results of the heuristic search, that is, the proposed categorization optimization algorithm, and the respective exhaustive search, performed with the same characteristics.

The cells of the tables referring to the heuristic search have the following information:

(1) number of categories;

(2) steps taken until the local maximum value of Kappa was obtained. In this case, the value in parentheses refers to the amount of fluctuations performed;

(3) initial Kappa value, obtained through categorization using the equal intervals method, and the final Kappa value, after using the proposed algorithm;

(4) size of the final OBDT obtained, related to the complexity of the constructed model;

(5) computational processing time; and

(6) the obtained sets of initial and final categories, both with respect to the instance index and the value of the AGB theme-feature.

On the other hand, the cells referring to the exhaustive search, in addition to the number of categories and the computational processing time, have information regarding:

(1) the number of analyzed states;

(2) the maximum Kappa value found, including the respective number of possibilities and its percentage.

Table 1. Search results on logarithmic feature values of SGC for the set of 3 categories.

Test Name	**Heuristic Search Results**					
SGC-Log-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	3 (1)	0.93	1.00	15	1.74
	Initial Categorization Set Instances: {[1 , 5] ,]6 , 25] ,]26 , 58]} Values: {[1.96 , 2.21] ,]2.21 , 2.35] ,]2.35 , 2.55]}			**Final Categorization Set** Instances: {[1 , 5] ,]6 , 29] ,]30 , 58]} Values: {[1.96 , 2.21] ,]2.21 , 2.37] ,] 2.37 , 2.55]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**

	1431	1.00 (51 possibilities – 3,5% of the analysed states)	9.95

Table 2. Search results on logarithmic feature values of SGC for the set of 5 categories.

Test Name	Heuristic Search Results					
SGC-Log-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	4 (0)	0.88	0.94	19	1.94
	Initial Categorization Set Instances: {[1 , 4] ,]5 , 7] ,]8 , 18] ,]19 , 52] ,]53 , 58] } Values: {[1.96 , 2.11] ,]2.22 , 2.24] ,]2.26 , 2.30] ,]2.32 , 2.43] ,]2.44 , 2.55]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 5] ,]6 , 18] ,]19 , 52] ,]53 , 58] } Values: {[1.96 , 1.99] ,]2.00 , 2.21] ,]2.22 , 2.31] ,]2.32 , 2.43] ,]2.44 , 2.55]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	270,676	1.00 (693 possibilities – 2,5% of the analysed states)				1,666.19

Table 3. Search results on original feature values of SGC for the set of 3 categories.

Test Name	Heuristic Search Results					
SGC-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	2 (1)	0.91	1.00	15	1,92
	Initial Categorization Set Instances index: {[1 , 7] ,]8 , 51] ,]52 , 58]} Values: {[92.21 , 171.86] ,]181.16 , 263.87] ,]272.22 , 351.73]}			**Final Categorization Set** Instances: {[1 , 6] ,]7 , 54] ,]55 , 58]} Values: {[92.21 , 167.58] ,]171.86 , 290.89] ,]301.21 , 351.73]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	1431	1.00 (66 possibilities – 4,6% of the analysed states)				9,63

Table 4. Search results on original feature values of SGC for the set of 5 categories.

Test Name	Heuristic Search Results					
SGC-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	3 (0)	0.95	0,98	21	1.82
	Initial Categorization Set Instances: {[1 , 4] ,]5 , 15] ,]16 , 38] ,]39 , 54] ,]55 , 58] } Values: {[92.21 , 132.50] ,]163.30 , 195.34] ,]197.37 , 247.80] ,]248.30 , 290.89] ,]301.21 , 351.73]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 15] ,]16 , 37] ,]38 , 54] ,]55 , 58] } Values: {[92.21 , 96.95] ,]101.69 , 195.34] ,]197.37 , 247.31] ,]247.80 , 290.89] ,]301.21, 351.73]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	270,676	1.00 (965 possibilities – 3,6% of the analysed states)				1,972.82

Table 5. Search results on logarithmic feature values of Unini for the set of 3 categories.

Test Name	Heuristic Search Results					
Unini-Log-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time**

						(sec)
	3	0	0.98	0.98	21	1.74
	Initial Categorization Set Instances: {[1 , 18] ,]19 , 40] ,]41 , 70]} Values: {[2.19 , 2.28] ,]2.29 , 2.39] ,]2.40 , 2.49]}			**Final Categorization Set** Instances: {[1 , 18] ,]19 , 40] ,]41 , 70]} Values: {[2.19 , 2.28] ,]2.29 , 2.39] ,]2.40 , 2.49]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	2145	1.00 (58 possibilities – 2,7% of the analysed states)				13.86

Table 6. Search results on logarithmic feature values of Unini for the set of 5 categories.

Test Name	**Heuristic Search Results**					
Unini-Log-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	0	0.96	0.96	25	1.86
	Initial Categorization Set Instances: {[1 , 2] ,]3 , 23] ,]24 , 34] ,]35 , 59] ,]60 , 70] } Values: {[2.19 , 2.22] ,]2.25 , 2.30] ,]2.31 , 2.37] ,]2.38 , 2.43] ,]2.44 , 2.49]}					
	Final Categorization Set Instances: {[1 , 2] ,]3 , 23] ,]24 , 34] ,]35 , 59] ,]60 , 70] } Values: {[2.19 , 2.22] ,]2.25 , 2.30] ,]2.31 , 2.37] ,]2.38 , 2.43] ,]2.44 , 2.49]}					
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	635,315	1.00 (700 possibilities – 0,1% of the analysed states)				5,625.84

Table 7. Search results on original feature values of Unini for the set of 3 categories.

Test Name	**Heuristic Search Results**					
Unini-3	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	3	3 (1)	0.94	0.96	23	1.92
	Initial Categorization Set Instances: {[1 , 23] ,]24 , 50] ,]51 , 70]} Values: {[153.32 , 202.15] ,]206.11 , 257.38] ,]258.87 , 311.57]}			**Final Categorization Set** Instances: {[1 , 24] ,]25 , 47] ,]48 , 70]} Values: {[153.32 , 206.11] ,]210.08 , 254.09] ,]254.99 , 311.57]}		
	Exhaustive Search Results					
	Analysed states	**Max Kappa**				**Processing Time (sec)**
	2145	1.00 (104 possibilities – 4,8% of the analysed states)				20.82

Table 8. Search results on original feature values of Unini for the set of 5 categories.

Test Name	**Heuristic Search Results**					
Unini-5	**Categories**	**Steps Taken**	**Initial Kappa**	**Final Kappa**	**Tree size**	**Processing Time (sec)**
	5	6 (2)	0.87	0.98	29	2.73
	Initial Categorization Set Instances: {[1 , 8] ,]9 , 28] ,]29 , 41] ,]42 , 62] ,]63 , 70] } Values: {[153.32 , 184.15] ,]185.33 , 214.71] ,]218.13 , 248.24] ,]250.32 , 277.58] ,]281.85 , 311.57]}					
	Final Categorization Set Instances: {[1 , 9] ,]10 , 24] ,]25 , 37] ,]38 , 59] ,]60 , 70] } Values: {[153.32 , 185.33] ,]185.67 , 206.11] ,]210.08 , 237.05] ,]240.17 , 267.57] ,]270.44 , 311.57]}					

	Exhaustive Search Results		
	Analysed states	**Max Kappa**	**Processing Time (sec)**
	635,315	1.00 (1,614 possibilities – 0,3% of the analysed states)	9,181.66

The results obtained in the SGC-Log-3 (Table 1) and SGC-3 (Table 3) tests shows that the proposed heuristic is capable of reaching the 1.00 Kappa index value, the global maximum. In the other cases, in general, the steps between the states also provided an increase in that index.

However, it is observed that in the Unini-Log-3 (Table 5) and Unini-Log-5 (Table 6) tests, the method of equal intervals performed the initial categorization within a local maximum of Kappa. In this case, the heuristic did not identify state steps that would increase this value, maintaining the same initial and final Kappa values.

Similar situations involving local maxima are found in the SGC-Log-3 (Table 1), SGC-3 (Table 3), Unini-3 (Table 7) and Unini-5 (Table 8) tests. In these cases the hill climb method found a local maximum, however, it was surpassed due to the floating step performed.

Another result observed is related to the size of the OBDT built for the categorization process. The trees associated with the SGC study area had sizes between 15 and 21, smaller than those obtained for the Unini area, which ranged between 21 and 29. According to Frank et al. (2016), the smaller the tree, the greater its generalizability, in addition to lower computational complexity and cost.

Regarding processing time, the developed heuristic search took 1 to 3 seconds to identify a maximum Kappa value and build a categorization model. This characteristic proves the advantage of using the hill climb search method over the exhaustive search, which took, in the extreme case of the Unini-5 (Table 8) test, more than 2.5 hours to analyze all the possibilities.

The exhaustive search algorithm also presented good results, analyzing all possible states and identifying those referring to the maximum global Kappa index. The maximum Kappa percentages found in the exhaustive searches ranged between 0.1% and 4.8% of the analyzed states, without following a specific pattern.

From the results, 4 (four) tests that stood out for the different areas of study and number of categories were selected: SGC-3, SGC-5, Unini-Log-3 and Unini-5. On these tests, comparative analyzes were carried out between thematic maps and those of representativeness of areas.

About the tests performed, the OBDT that were built selected the most significant features. Table 9 presents these features that are detailed by Castro-Filho and Bias (2021). The most prominent features are the interferometric height (Hint) and the texture and target decomposition extracted over the L-band.

Table 9. Characteristics of the constructed OBDT

Test	Tree Size	Selected Features
SGC-3	15	• Variance texture feature for 3x3 pixels window size over L_{hv} band (3_LHV_Var) • S1 orientation angle (ψ) over L band (TPSI1L) • Interferometric height between X and P bands (Hint) • Ratio between parallel polarizations over L band (RP_L) • Variance texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Var) • Amplitude image of the X band in the HH polarization (Xhh) • Amplitude image of the P band in the HV polarization (Phv)
SGC-5	21	• Correlation texture feature for 5x5 pixels window size over L_{hh} band (5x5_LHH_Cor) • Subtration between amplitudes of HH and HV polarizations in L band (Lhh-Lhv) • Contrast texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Con) • Homogeneity texture feature for 3x3 pixels window size over L_{hh} band (3x3_LHH_Ho) • Correlation texture feature for 7x7 pixels window size over X_{hh} band (7x7_XHH_Cor) • Mean texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Me)

		• Amplitude image of the L band in the HV polarization (Lhvvh) • S2 magnitude (α) over L band (TAlphaS2L) • S3 magnitude (α) over L band (TAlphaS3L)
Unini-Log-3	21	• Interferometric height between X and P bands (Hint) • Mean texture feature for 7x7 pixels window size over L_{hh} band (7x7_LHH_Me) • Entropy texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_En) • Amplitude image of the P band in the HV polarization (Phvvh) • Amplitude image of the P band in the HH polarization (Phh) • Contrast texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Con) • Variance texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Var) • Odd Scattering over L band (VanZOddL) • Variance texture feature for 3x3 pixels window size over L_{hv} band (3x3_LHV_Var)
Unini-5	27	• Interferometric height between X and P bands (Hint) • Contrast texture feature for 3x3 pixels window size over L_{hv} band (3x3_LHV_Con) • Variance texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Var) • Variance texture feature for 7x7 pixels window size over L_{vv} band (7x7_LVV_Var) • Mean texture feature for 5x5 pixels window size over L_{vv} band (5x5_LVV_Me) • Dissimilarity texture feature for 7x7 pixels window size over P_{hh} band (7x7_PHH_Di) • Homogeneity texture feature for 3x3 pixels window size over X_{hh} band (3x3_XHH_Ho) • S2 magnitude (α) over L band (TAlphaS2L) • Subtration between amplitudes of HH and HV polarizations in L band (Lhh-Lhv) • Subtration between amplitudes of HH and HV polarizations in P band (Phh-Phv) • Amplitude image of the L band in the VV polarization (Lvv) • Correlation texture feature for 7x7 pixels window size over X_{hh} band (7x7_XHH_Cor) • Entropy texture feature for 3x3 pixels window size over P_{hv} band (3x3_PHV_En)

The thematic representations were built on a delimited area of 5x5 km, referring to a sample of 1 million pixels in the 5 meters spatial resolution images. The areas were selected in the SGC and Unini regions in such a way that they were predominantly primary forests, with no influence of other natural or anthropogenic elements that could

affect the representativeness of the samples, such as rivers and settlements. The representation of the biomass categories followed a choropleth representation, as shown in Figure 4.

From the products constructed, it is observed that those referring to the SGC-3 (Figure 4 A) and SGC-5 (Figure 4 B) tests have a speckled appearance, of "salt and pepper", with the visually balanced and homogeneous presence of all biomass categories. It is noteworthy that this obtained appearance is typical of the SAR data nature. Authors highlights that, even after using adaptive filters and increasing the equivalent number of looks in the SAR image (Woodhouse 2017; Pereira et al. 2018), steps that reduces the effects of the speckle noise, the geometric characteristics of the imaged targets remain evident (Sarker et al 2012). In fact, the imaged primary forest presents textural characteristics of granular roughness, providing homogeneous variation between biomass categories.

On the other hand, the thematic map of the Unini-Log-3 (Figure 4 C) test shows less categories homogeneity, with a greater presence of low biomass areas in the upper left corner. This presence also contributes to the identification of a sinuous shape of a probable river, which had not been visually identified previously.

The result of the Unini-5 (Figure 4D) test, in its turn, showed a large presence of the medium biomass category, with few pixels for the other categories.

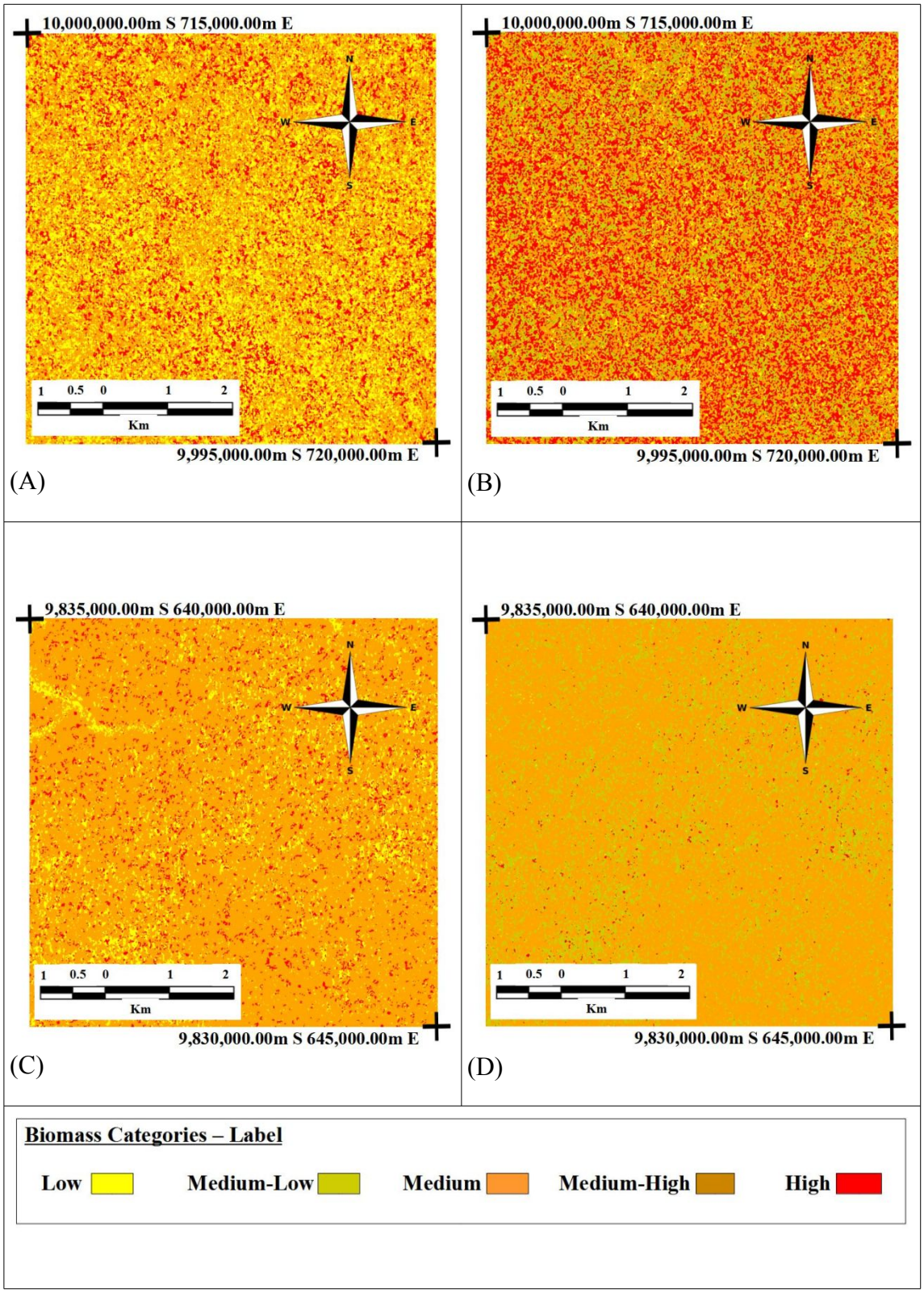


Figure 4. Thematic maps of biomass estimation referring to tests (A) SGC-3, (B) SGC-5, (C) Unini-Log-3 and (D) Unini-5

In order to analyze the representativeness of the biomass categorization and estimation models built for both study areas, analyzes were performed on 10 (ten)

sample areas, which results are shown in Figures 5 to 8. In the graphs there are percentages of pixels, for each category, found in the sample areas randomly distributed in the respective study areas, compared to the percentage of categories obtained by the categorization models by the method of equal intervals and by the proposed heuristic.

In this analysis, the graphs shows that the sample areas of the SGC-3 (Figure 5) and SGC-5 (Figure 6) tests presented a decrease in the average biomass category, with the respective increase in the categories with more extreme values. Despite these changes, it is observed that both equal-interval and heuristic methods were able to satisfactorily model the SGC study area, for 3 or 5 categories, with small modifications.

Conversely, the sample areas of the Unini-Log-3 test (Figure 7) showed a large increase in the average biomass category. However, a coherent distribution of extreme biomass values was maintained, that is, low and high biomass.

The sample areas of the Unini-5 test (Figure 8), in turn, did not present a distribution of categories equivalent to those built in the models. In these cases, the extreme values of categories were significantly reduced, with almost zero occurrence of the category of medium-low biomass.

It is also noteworthy that, in all graphs, the numerical value of biomass for each category are found next to the legends. The values do not present any type of anomaly that justifies a possible problem of representativeness of the study areas, that is, no cases were observed where there is a discontinuity of values or where the category intervals are insignificantly small to the point of suggesting that it does not exist. The fact that the heuristic started from a classical categorization, that is, by equal intervals, contributed for coherent numerical AGB results.

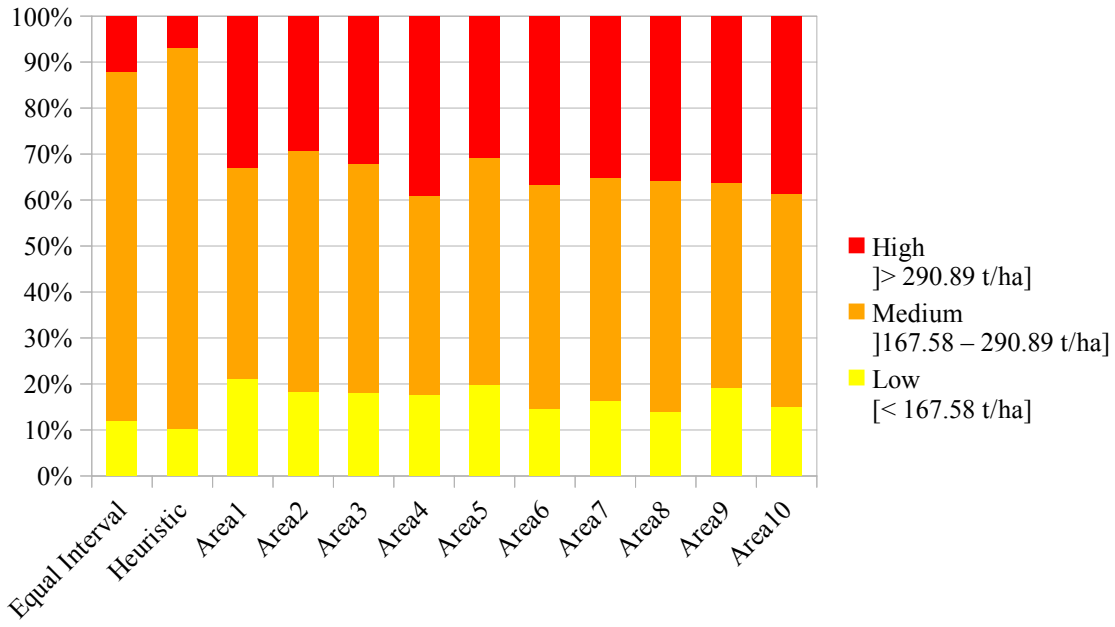


Figure 5. Graph showing the representativeness of the study area for the SGC-3 test.

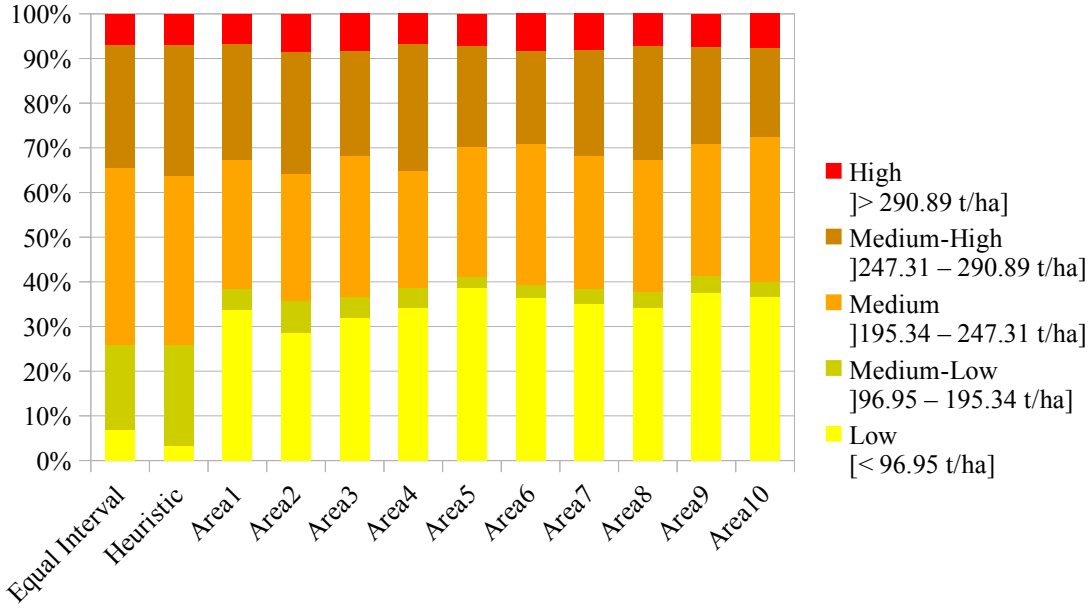


Figure 6. Graph showing the representativeness of the study area for the SGC-5 test.

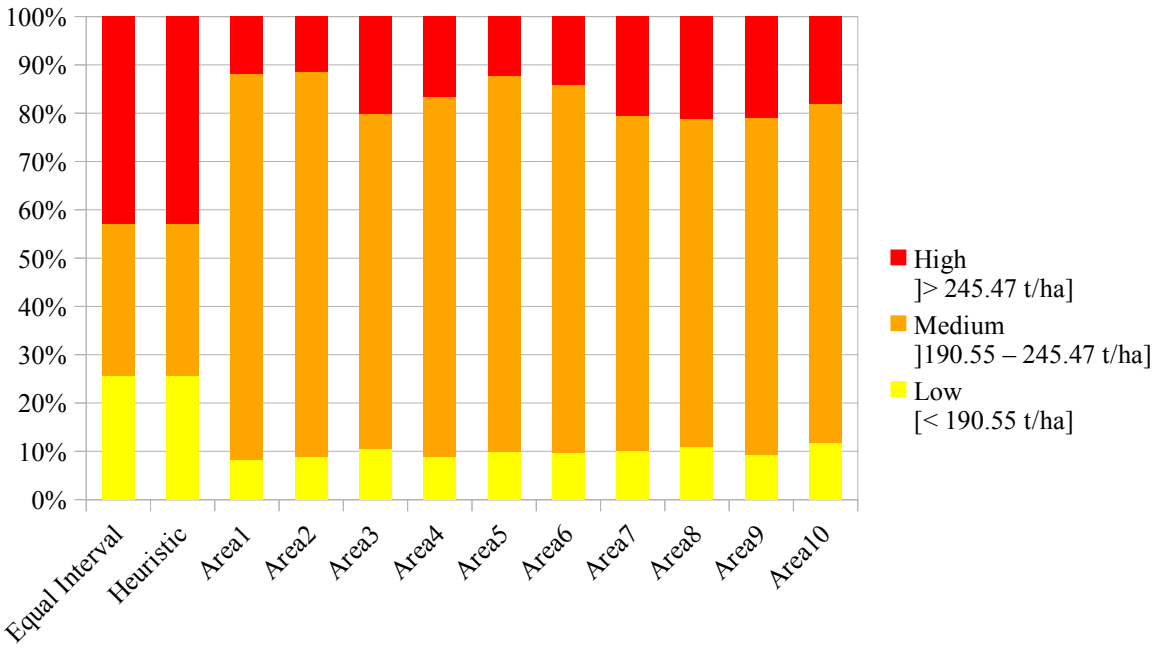


Figure 7. Graph showing the representativeness of the study area for the Unini-Log-3 test.

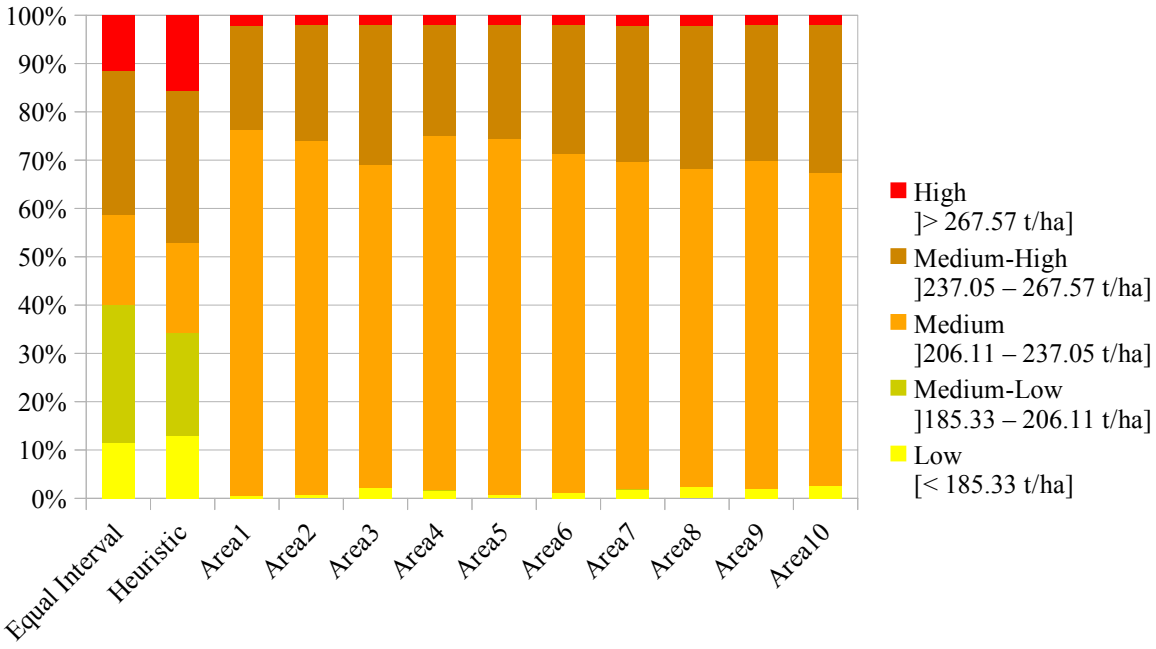


Figure 8. Graph showing the representativeness of the study area for the Unini-5 test.

The comparative analysis between the thematic produtcs generated through the Categorization Optimization, proposed heuristic, and the categorization performed by the method of equal intervals, a classic process generated the confusion matrices

contained in Tables 10 to 12. In all of them, the values are found in percentage, with emphasis on the main diagonal in bold. The confusion matrix referring to the Unini-Log-3 test was not built, since both thematic maps are identical because there were no steps between states during the execution of the algorithm, as shown in Table 5.

The confusion matrix referring to the SGC-3 test (Table 10) shows that there is low coherence between the thematic products constructed. In this case, the proposed algorithm performed significant changes in pixel values, allocating them to other categories.

On the other hand, the SGC-5 test matrix (Table 11) shows strong coherence between the thematic maps. Likewise, Table 4 shows that there was little change in the Kappa index value, ranging only from 0.95 to 0.98.

The confusion matrix referring to the Unini-5 test (Table 12) showed coherence only for the medium and high-medium categories, not for the others. In this case, the area defined for the construction of the thematic product, by the method of equal intervals, does not have pixels from the high biomass category. This analysis is in accordance with the result presented in Figure 7 where the sample areas, in general, presented few cases for this category, in contradiction to the constructed categorization models.

Table 10. Confusion matrix for the SGC-3 test.

SGC-3				
Categorization Optimizer	Equal Intervals			
		Low	**Medium**	**High**
	Low	**21.84**	21.03	20.88
	Medium	44.15	**44.55**	44.47
	High	34.01	34.43	**34.66**

Table 11. Confusion matrix for the SGC-5 test.

SGC-5

	Equal Intervals					
		Low	**Medium-Low**	**Medium**	**Medium-High**	**High**
Categorization Optimizer	**Low**	**90.89**	0	0	0	0
	Medium-Low	3.40	**48.57**	2.82	0	0
	Medium	5.66	51.43	**81.85**	0	18.75
	Medium-High	0.05	0	3.52	**100.00**	18.47
	High	0	0	11.81	0	**62.78**

Table 12. Confusion matrix for the Unini-5 test.

Unini-5						
	Equal Intervals					
		Low	**Medium-Low**	**Medium**	**Medium-High**	**High**
Categorization Optimizer	**Low**	**3.87**	0.01	0.01	0.01	0
	Medium-Low	0.04	**0.03**	0.05	0.02	0
	Medium	69.50	79.30	**78.98**	39.39	0
	Medium-High	25.14	18.90	19.07	**55.97**	0
	High	1.45	1.76	1.89	4.61	**0**

4. Conclusions

This article aimed to propose an innovative methodology at optimizing the categorization process in the construction of thematic products that guides elements linked to biomass existing in primary forests. The importance of this knowledge has a direct influence on the Sustainable Development Goals (SDGs) of the UN.

The theme referring to the estimation of AGB was approached for two study areas in the Amazon forest region and used remote sensing data and artificial intelligence techniques in an innovative way.

The results obtained show that the proposed Categorization Optimization algorithm demonstrated the ability to find new subintervals of categories that increased the Kappa agreement index. As a result, the constructed maps presented thematic accuracy superior to those obtained by classical categorization methods.

Together, the analyzes show that the adjustments made to the limits of the AGB categories did not impact the representativeness of the model for a study area, maintaining the characteristics of the region. According to Vogt et al. (2012), the labels of classes or categories should not be treated simply and automatically, but with human interference to improve the process.

From a computational perspective, the proposed heuristic enabled the identification of maximum values for the objective function in an efficient way, avoiding the high processing costs of the exhaustive search. In the extreme case, the search time decreased from 2.5 hours to 3 seconds.

In order to validate the proposed method for the construction of different types of thematic products, future tests with other databases are still needed.

In this sense, future works will present the development of heuristics that seek to analyze and compare the possibilities of maximum Kappa index found by the exhaustive search method, identifying possible advantages in selecting a specific categorization state and model.

References

Assis, F. G., F. Luiz, R. Karine Ferreira, L. Vinhas, Luis Maurano, Claudio Almeida, Andre Carvalho, Jether Rodrigues, Adeline Maciel, and Claudinei Camargo. 2019. “TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale

Thematic Mapping." *ISPRS International Journal of Geo-Information* (11): 513–546. https://doi.org/10.3390/ijgi8110513.

Bertin, J., 1977. *La Graphique et le Traitement Graphique de l'Information. Flamarion*, 278p. France.

Bueno-Crespo, A., R. Martínez-España, I. Timón, and J. Soto. 2018. "An Unsupervised Technique to Discretize Numerical Values by Fuzzy Partitions". *Journal of Ambient Intelligence and Smart Environments*, (10): 289–300 289. DOI 10.3233/AIS-180488.

Bussab, W. O., and P. A. Morettin. 2017. *Estatística Básica*. 9ª Edição. Editora Saraiva. 568 p. ISBN-10: 8547220224.

Castro-Filho, C. A. P., and E. Bias. 2021. "Comparison between Quantitative and Qualitative Theme-Feature Forest Biomass Estimation Models built over SAR Data". *International Journal of Advanced Engineering Research and Science*, (8). doi:10.22161/ijaers.87.41.

Congalton, R. G., K. Green. 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Lewis Publishers. Denver. EUA. 180 p. ISBN 0873719867.

Costa, H., G.M. Foody, and D.S. Boyd. 2017. "Using mixed objects in the training of object-based image classifications". *Remote Sens. Environ*, (190): 188–197. https://doi.org/10.1016/j.rse.2016.12.017.

Debastiani, A.B., M.M. Moura, F.E. Rex, C.R. Sanquetta, A.P.D. Corte, and N. Pinto. 2019. "Regressões Robusta e Linear para Estimativa de Biomassa Via Imagem Sentinel em uma Floresta Tropical". *BIOFIX Science Journal*. (4): 81–87. https://doi.org/10.5380/biofix.v4i2.62922.

Dent, B., J. Torquson, and T. Hodler. 2008. *Cartography: Thematic Map Design.* 6th ed. McGraw-Hill Science. 368 p. ISBN 0072943823.

Dietterich, T. G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". *Neural Comput*. (10): 1895-1923. doi: 10.1162/089976698300017197.

Diniz, C.G., A. A. de Almeida Souza, D. C. Santos, M. C. Dias, N. C. da Luz, D. R.V. de Moraes, J. S. Maia, A. R. Gomes, I. da Silva Narvaes, and D. M. Valeriano. 2015. "DETER-B: The new Amazon near real-time deforestation detection system." *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. (8): 3619–3628. DOI: 10.1109/JSTARS.2015.2437075.

Erb, K.H., T. Kastner, C. Plutzar, A. L. S. Bais, N. Carvalhais, T. Fetzel, S. Gingrich, H. Haberl, C. Lauk, M. Niedertscheider, J. Pongratz, M. Thurner, and S. Luyssaert. 2018. "S. Unexpectedly large impact of forest management and grazing on global vegetation biomass". *Nature.* (553): 73–76. https://doi.org/10.1038/nature25138.

Diretoria de Serviço Geográfico (DSG). 2008. *Infra-estrutura de dados espaciais para a Amazônia*. São José dos Campos. Encontro de Usuários de Sensoriamento Remoto das Forças Armadas (SERFA).

Foody, G. M. 2021. "Impacts of ignorance on the accuracy of image classification and thematic mapping". *Remote Sensing of Environment.* (259): 112367, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2021.112367.

Frank, E., Hall, M. A., Witten, I. H. 2016. *The WEKA Workbench. Online Appendix for* "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

Garcia, S., J. Luengo, J. A. Sáez, V. Lopez, and F. Herrera. 2013. "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning". *IEEE Transactions on Knowledge and Data Engineering.* (25): 734–750.

Google Earth. 2021. http://www.google.com.br/intl/pt-BR/earth/. Access in 22 fev. 2021.

Hastie T, R. Tibshirani, and J. Friedman. 2009. *Elements of statistical learning: data mining, inference and prediction*, 2nd edn. Springer, Berlin.

Higuchi, N., J. Santos, R. J. Ribeiro, L. Minette, and Y. Biot. 1998. "Biomassa da parte aérea da vegetação da Floresta Tropical úmida de terra-firme da Amazônia Brasileira". *Acta Amaz.*. (28): 153–166. https://doi.org/10.1590/1809-43921998282166.

Le Noë, J., S. Matej, A. Magerl, M. Bhan, K. H. Erb, and S. Gingrich. 2020. "Modeling and empirical validation of long-term carbon sequestration in forests (France, 1850–2015)". *Global Change Biology.* (26): 2421–2434. https://doi.org/10.1111/gcb.15004.

Lima, A. J. N., R. Suwa, G. H. P. M. Ribeiro, T. Kajimoto, J. Santos, R. P. Silva, C. S. A. Souza, P. C. Barros, H. Noguchi, M. Ishizuka, and N. Higuchi. 2012. "Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro, Brazil". *Forest Ecology and Management.* (277): 163-172. doi: 10.1016/j.foreco.2012.04.028.

Liu, H., F. Hussain, C. Tan, and M. Dash. 2002. "Discretization: an enabling technique". *Data Mining and Knowledge Discovery*. (4): 393–423. https://doi.org/10.1023/A:1016304305535.

MAPBIOMAS. 2021. http://mapbiomas.org/map#coverage. Access in 18 de agosto de 2021.

Martins-Bedê, F. T., C. D. C. Freitas, and L. V. Dutra. 2009. "Risk Mapping of Schistosomiasis in Minas Socioeconomic Spatial Data". *IEEE Transactions on Geoscience and Remote Sensing*. (47): 3899-3908. doi: 10.1109/IGARSS.2008.4778845.

Maslove, D.M., T. Podchiyska, and H. J. Lowe. 2013. "Discretization of continuous features in clinical datasets". *Journal of the American Medical Informatics Association*. (20): 544–553. doi:10.1136/amiajnl-2012-000929.

Mitchell, P. J., A. L. Downie, and M. Diesing. 2018. "How good is my map? A tool for semi-automated thematic mapping and spatially explicit confidence assessment." *Environmental Modelling & Software*. (108): 111-122, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2018.07.014.

Pereira, L. O., L. F. A. Furtado, E. M. L. M. Novo, S. J. S. Sant'Anna, V. Liesenberg, and T. S. E. Silva. 2018. "Multifrequency and Full-Polarimetric SAR assessment for estimating above ground biomass and leaf area index in the Amazon Várzea Wetlands". *Remote Sensing*. (10): 1–23. https://doi.org/10.3390/rs10091355.

Prodes, P. 2013. *Monitoramento da floresta Amazônica Brasileira por satélite*. Inst. Nac. De Pesqui. Espac. Proj. Prodes, 25.

Projeto RadamBrasil. 1977. *Geologia, geomorfologia, pedologia, vegetação e uso potencial da terra*. Rio de Janeiro, Departamento Nacional da Produção Mineral.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Rajbahadur, G. K., S. Wang, Y. Kamei, and A. E. Hassan. 2021. "Impact of Discretization Noise of the Dependent Variable on Machine Learning Classifiers

in Software Engineering". *IEEE Transactions on Software Engineering.* (47): 1414-1430, doi: 10.1109/TSE.2019.2924371.

Raposo, P., G. Touya, and P. Bereuter. 2020. "A Change of Theme: The Role of Generalization in Thematic Mapping". *ISPRS International Journal of Geo-Information*. (9): 371. https://doi.org/10.3390/ijgi9060371.

Rosenfeld, A., R. Illuz, D. Gottesman, and M. Last. 2018. "Using discretization for extending the set of predictive features". *EURASIP J. Adv. Signal Process.* (7). https://doi.org/10.1186/s13634-018-0528-x.

Russel, S., and P. Norvig. 2020. *Artificial Intelligence: A Modern Approach Edition, Pearson*, 1115 pages, ISBN-10: 0134610997.

Sarker, M. L. R., J. Nichol, H. B. Iz, B. B. Ahmad, and A. A. Rahman. 2012. "Potential of texture measurements of two-date dual polarization PALSAR data for the improvement of forest biomass estimation". *ISPRS Journal of Photogrammetry and Remote Sensing*. (69): 146-166. https://doi.org/10.1016/j.isprsjprs.2012.03.002.

Scipal, K., M. Arcioni, F. Fois, C-C Lin, J. Chave, J. Dall, T. LeToan, K. Papathanassiou, S. Quegan, F. Rocca, S. Saatchi, H. Shugart, L. Ulander, and M. Williams. 2010. "The BIOMASS Mission – An ESA Earth Explorer candidate to measure the BIOMASS of the Earth's forests". In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 52-55) https://doi.org/10.1109/IGARSS.2010.5648979.

Silva, R.P. 2007. *Alometria, estoque e dinâmica da biomassa de florestas primárias e secundárias na região de Manaus (AM)*. National Institute for Space Research (INPE). PhD Thesis.

Sluter, C. R., S. P. Camboim, A. L. Iescheck, L. B. Pereira, M. C. Castro, M. M. Yamada, and V. S. Araújo. 2018. "A Proposal for Topographic Map Symbols for Large-Scale Maps of Urban Areas in Brazil". *The Cartographic Journal*. (55): 362-377, DOI: 10.1080/00087041.2018.1549307.

Theodoridis, S., and K. Koutroumbas. 2008, *Pattern Recognition*. 4th ed. Academic Press. 961 p. ISBN 1597492728.

Vogt, L., P. Grobe, B. Quast, and T. Bartolomaeus. 2012. "Fiat or bona fide boundary—a matter of granular perspective". *PLoS One*. (7): 12, e48603.

Witten, I.H., E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Massachusetts. https://doi.org/10.1016/c2009-0-19715-5.

Woodhouse, I.H. 2017. *Introduction to Microwave Remote Sensing, Introduction to Microwave Remote Sensing*. Taylor & Francis Group CRC Press, Florida. https://doi.org/10.1201/9781315272573.

Wu, T., W. Dong, J. Luo, Y. Sun, Q. Huang, W. Wu, and X. Hu. 2019. "Geo-parcel-based geographical thematic mapping using C5.0 decision tree: a case study of evaluating sugarcane planting suitability". *Earth Sci Inform*. (12): 57–70. https://doi.org/10.1007/s12145-018-0360-8.

Yang, Y., and G. I. Webb. 2009. "Discretization for naive-Bayes learning: managing discretization bias and variance". *Machine Learning*. 39-74. doi: 10.1007/s10994-008-5083-5.