Universidade do Estado do Pará
Pró-Reitoria de Pesquisa e Pós-Graduação
Centro de Ciências Naturais e Tecnologia
Pós-Graduação em Ciências Ambientais – Mestrado

Thamiris das Graças Pereira

Aplicação de algoritmo computacional baseado em distância para classificações automáticas de tipologias florestais e de classes de volume de biomassa florestal nas Glebas Mamuru-Arapiuns localizadas no estado do Pará

Belém
2016

Thamiris das Graças Pereira

Aplicação de algoritmo computacional baseado em distância para classificações automáticas de tipologias florestais e de classes de volume de biomassa florestal nas Glebas Mamuru-Arapiuns localizadas no estado do Pará

Dissertação apresentada para obtenção do título de mestre em Ciências Ambientais no Programa de Pós-Graduação em Ciências Ambientais.
Universidade do Estado do Pará.
Orientador: Prof. Dr. Hélio Raymundo Ferreira Filho.
Coorientador: Prof. Dr. José Alberto Sá

Belém
2016

Dados Internacionais de Catalogação na Publicação (CIP),
Biblioteca do Centro de Ciências Naturais e Tecnologia, UEPA, Belém - PA.

P436aPereira, Thamiris das Graças

Aplicação de algoritmo computacional em distância para classificações automáticas Ed tipologias florestais e de classes de volume de biomassa florestal nas Glebas Mamuru-Arapiuns localizadas no Estado do Pará. / Tamiris das Graças; Orientador Hélio Raymundo Ferreira Filho; Coorientador José Alberto Silva de Sá. -- Belém, 2016.

56f. : il. ; 30 cm.

Dissertação (Mestrado em Ciências Ambientais) – Universidade do Estado do Pará, Centro de Ciências Naturais e Tecnologia, Belém, 2016.

1.Floresta. 2. Desmatamento. 3.Sensoriamento Remoto. I. Ferreira Filho, Hélio Raymundo. II. Sá, José Alberto Silva de. III. Título.

CDD 634.9

Thamiris das Graças Pereira

Aplicação de algoritmo computacional baseado em distância para classificações automáticas de tipologias florestais e de classes de volume de biomassa florestal nas Glebas Mamuru-Arapiuns localizadas no estado do Pará

Dissertação apresentada como requisito para obtenção do título de mestre em Ciências Ambientais no Programa de Pós-Graduação em Ciências Ambientais.
Universidade do Estado do Pará.

Data da aprovação: 29/02/2016

Banca Examinadora

_____ – Orientador

Prof. Hélio Raymundo Ferreira Filho
Doutor em Ciências de Gestão
Universidade do Estado do Pará

_____ - 1º Examinador

Prof. Arthur da Costa Almeida
Doutor em Engenharia Elétrica
Universidade Federal do Pará

_____ - 2º Examinador

Profa. Yvelyne Bianca Iunes Santos
Doutora em Engenharia de Recursos Naturais da Amazônia
Universidade do Estado do Pará

_____ - 3º Examinador

Profa. Risete Maria Queiroz Leão Braga
Doutora em Geologia e Geoquímica
Universidade Federal do Pará

_____ - 4º Examinador

Prof. Marcelo José Raiol Souza
Doutor em Engenharia de Recursos Naturais
Universidade do estado do Pará

_____ - 5º Examinador

Joner Oliveira Alves
Doutor em Engenharia Metalúrgica e de Materiais
Instituto SENAI de Inovação em Tecnologias Minerais

_____ - Suplente

Profa. Norma Ely Santos Beltrão
Doutora em Economia Agrícola
Universidade do Estado do Pará

Aos meus pais, dedico esta trajetória, tanto pelo apoio quanto pelas advertências. Sem o esforço deles não teria conseguido a educação que tenho hoje.

AGRADECIMENTOS

Primeiramente agradeço a Deus. Crer em um ser superior que olha por todos nós, fez-me ter forças para acreditar que no fim tudo daria certo e que não desistiria fácil.

A minha família pelo lar que possuo, à minha mãe e ao meu pai pelas oportunidades de estudos proporcionadas e aos meus gatos que são meus companheiros inseparáveis.

A CAPES pelo processo de seleção de projetos que proporcionou ao projeto de desenvolvimento regional recursos para sua realização e agradeço também pela bolsa que tornou possível a permanência na pesquisa com dedicação exclusiva.

Ao Ideflor-Bio que firmou parceria em 2015 com a Universidade do Estado do Pará e cedeu para essa dissertação os dados necessários para sua realização.

A Universidade do Estado do Pará que proporcionou a minha formação acadêmica (Graduação e Mestrado).

Ao Programa de Pós-Graduação em Ciências Ambientais pela oportunidade e aos professores pelos seus conhecimentos compartilhados com a turma que foram relevantes para formação de todos.

Ao meu orientador Hélio Ferreira Filho pelos dois anos de parceria e ao meu coorientador José Alberto Sá pela sua grande contribuição na parte da dissertação relacionada à inteligência Computacional.

Ao professor Arthur Almeida, que com seu Doutorado tornou esse projeto possível, e também pelas suas contribuições nos momentos mais complicados da pesquisa.

A banca examinadora formada pelos professores Arthur Almeida, Yvelyne Santos, Risete Braga, Marcelo Souza e Joner Alves, agradeço todas as contribuições dadas à dissertação.

Agradeço aos meus colegas da turma de Ciências Ambientais 2014 pela companhia e apoio que tranformou esses dois anos em um período mais animado.

Quero também agradecer ao meu grupo de trabalho que é formado por mim, Monique Farias, Nariane Quaresma e Renata Tenório. Todas as parcerias em artigos e eventos me mostraram o quanto vocês são organizadas e eficientes. Suas carreiras serão brilhantes.

Os teus atos, e não os teus conhecimentos, é que determinam o teu valor.

Johann Fichte (1762 – 1814)

RESUMO

A Amazônia já passou por diferentes etapas relacionadas a utilização de seus recursos naturais. Atualmente, existem duas formas de atuação sob os recursos da Amazônia, uma é de manter a trajetória econômica e institucional baseada na agropecuária e no extrativismo madeireiro e mineral. A outra forma de atuação sob os recursos da Amazônia são propostas globais de pagar para não desflorestar envolvendo a mercantilização do carbono. O desafio da atualidade é de intervir nos processos que geram o desmatamento e também de não manter florestas em pé improdutivas para o homem. Tecnologias são ferramentas para atingir esse desafio nos objetivos de monitoramento, gestão e desenvolvimento de novas tecnologias. O sensoriamento remoto e a inteligência computacional são dois exemplos de tecnologias aplicadas em diversas áreas do conhecimento. O trabalho tem como justificativa a utilização de sensoriamento remoto e inteligência computacional para otimização de estudos de vegetação em áreas extensas como a Amazônia. Na dissertação, as tecnologias são utilizadas na região amazônica especificamente para as formações florestais existentes nas Glebas Mamuru – Arapiuns no estado do Pará. No artigo 1, os usos dessas tecnologias são para classificação automática de tipologias florestas, em que se obteve resultados satisfatórios quanto as medidas de precisão utilizadas, que foram acurácia global 87%, índice Kappa 74% e área sub a curva do gráfico da característica de Operação do Receptor de 0,929. No artigo 2 foi realizada metodologia semelhante utilizada no artigo 1, mas com objetivo de classificação de faixas de volume de biomassa florestal da área de estudo, foram utilizadas como medidas de precisão a acurácia global, medida F, precisão, revocação e área sob a curva.

Palavras-chave: Biomassa. Amazônia. Inteligência artificial. K-vizinhos mais próximos. Medidas de precisão.

ABSTRACT

The Amazon has gone through different stages related to the use of its natural resources. Currently, there are two forms of action under the resources of the Amazon, one is to maintain the economic and institutional trajectory based on farming and timber and mineral extraction. The other form of action under the resources of the Amazon are global proposals to pay for not deforesting involving the commodification of carbon. The present challenge is to intervene in the processes that generate deforestation and also not keep forests on unproductive foot man. Technologies are tools to meet this challenge in the objectives of monitoring, management and development of new technologies. Remote sensing and computational intelligence are two examples of technologies applied in various areas of knowledge. The work is justified by the use of remote sensing and computational intelligence to optimize vegetation studies in large areas such as the Amazon. In the dissertation, the technologies are used in the Amazon region specifically for existing forest formations in Glebas Mamuru -. Arapiuns in the state of Pará in Article 1, the uses of these technologies are for automatic classification of forest types, which were obtained satisfactory results as the precision measurements used, which were overall accuracy 87%, Kappa index 74% and area under the graphic curve Receiver Operating characteristic of 0.929. Article 2 was carried out similar methodology used in Article 1, but with the purpose of classification of forest biomass volume ranges of the study area were used as precision measurements the overall accuracy, as F, precision, recall and area under the curve.

Keywords: Biomass. Amazon. Artificial intelligence. K-nearest neighbors. precision measurements.

LISTA DE TABELAS

TABELAS DO ARTIGO 1

TABELAS DO ARTIGO 2

LISTA DE QUADROS

LISTA DE QUADROS DO ARTIGO 1

LISTA DE QUADROS DO ARTIGO 2

LISTA DE FIGURAS

LISTA DE FIGURAS DO ARTIGO 1

LISTA DE FIGURAS DO ARTIGO 2

LISTA DE ABREVIATURAS E SIGLAS

K-NN	K- vizinhos mais próximos
Ideflor-bio	Instituto de Desenvolvimento Florestal e da Biodiversidade do Estado do Pará
IBGE	Instituto Brasileiro de Geografia e Estatística
Dbe	Floresta Ombrófila Densa Terras baixas Dossel emergente
Dbe + Abp	Floresta Ombrófila Densa Terras baixas Dossel emergente + Aberta com palmeiras
UEPA	Universidade do Estado do Pará
TM	Mapeador Temático
NASA	National Aeronautics and Space Administration
NDVI	Índice de Vegetação da Diferença Normalizada
ROC	Característica de Operação do Receptor
AUC	Área sob a curva
DAP	Diâmetro da Altura do Peito
DN	Digital Number
F1	Medida F
CA	Acurácia Global
TP	Verdadeiro Positivo
FN	Falso Negativo
TF	Verdadeiro Negativo
FP	Falso Positivo
PP	Número de exemplos preditos positivos
PN	Número de exemplos preditos negativos
POS	Número real de exemplos positivos na amostra
NEG	Número real de exemplos negativos na amostra
N	Número da amostra
Re	Revocação
Pr	Precisão

SUMÁRIO

1- INTRODUÇÃO GERAL

A Amazônia já passou por diferentes etapas relacionadas a utilização de seus recursos naturais, primeiro por meio de produção e exportação da borracha, segundo foi por intervenções esporádicas do governo federal, terceiro quando a região é escolhida para ser o local de ações de planejamento territorial e o quarto é caracterizado pelas ações estatais em menor escala comparados ao período anterior e pelos avanços dos agentes impulsionados pelas forças de mercado internas e também externas (PRATES & BACHA, 2011).

Atualmente, existem duas formas de atuação, que se destacam, sob os recursos da Amazônia, uma é de manter a trajetória econômica e institucional baseada no extrativismo madeireiro e mineral, e numa agropecuária, em que a produção é destinada ao mercado externo sem ou com fraca agregação de valor, e associada ao crescente desflorestamento e desterritorialização das populações tradicionais, e no extremo oposto do projeto da continuidade em derrubar a floresta e substituí-la por pastagens e lavouras, encontram-se propostas globais de pagar para não desflorestar envolvendo a mercantilização do carbono (BECKER, 2013).

O primeiro é referente ao padrão econômico voltado para a exportação que desde época da colonização é a motivação que prevalece na ocupação regional (BECKER, 2001). Mas o que é realmente necessário é enfrentar o desafio de intervir nos processos que geram o desmatamento e não manter florestas em pé improdutivas para o homem, aproximando-se do interesse nacional (BECKER, 2013).

Para se aproximar do objetivo de ter florestas produtivas ao homem e atividades para agregar à região amazônica de forma sustentável é necessário inovação tecnológica em prol do desenvolvimento sustentável e segundo Abramovay (2011), até o presente ano do seu estudo, Brasil não estava se aproximando da marca dominante da inovação tecnológica, cada vez mais direcionada para colocar ciência a serviço de sistemas produtivos poupadores de materiais, de energia, e capazes de contribuir para a regeneração da biodiversidade.

Na floresta amazônica, além da Floresta Ombrófila Densa existem outros três tipos de vegetação predominantes dentro da região florística hileiana que são Floresta Ombrófila Aberta, Floresta Estacional Sempre-verde e a Campinarana

(IBGE, 2012). Estudos relacionados à biomassa e ao carbono em formações florestais são realizados com vários objetivos, destacando-se a quantificação da ciclagem de nutrientes, a quantificação para fins energéticos e como base de informações para estudos de sequestro de carbono (SILVEIRA, 2010).

O sensoriamento remoto e os sistemas de informações geográficas desempenham papel interessante nesse vasto contexto da observação e do monitoramento da superfície terrestre (SHIMABUKURO, MAEDA & FORMAGGIO, 2009).

O sensoriamento remoto e inteligência computacional ou artificial podem ser usados juntos para segmentação e classificação como é observado nos trabalhos de Andrade, Francisco & Almeida (2014) que avaliou o desempenho de classificadores paramétrico e não paramétrico na classificação da fisionomia vegetal, utilizando assim como na dissertação o Índice de Vegetação por Diferença Normalizada (Normalized Difference Vegetation Index - NDVI) e as medidas de precisão, acurácia global e índice Kappa. E Negri, Sant'anna & Dutra (2013) que aplica modelos de aprendizado semi supervisionado na classificação de imagens de sensoriamento remoto, comparando os modelos através da acurácia global.

A maioria das abordagens de visão artificial podem ser divididas em três fases: aquisição, seleção e classificação. É importante ressaltar que, quando uma das fases não é executada de maneira correta, o resultado pode ser afetado de forma negativa (SILVA et al., 2015).

A dissertação tem como objeto de estudo as florestas existentes nas Glebas Mamuru-Arapiuns no estado do Pará. O trabalho tem como justificativa a utilização de recursos tecnológicos para otimização de estudos de vegetação e biomassa em áreas extensas como a Amazônia. E como objetivos: classificar tipologias florestais e classificar a imagem utilizando classes com faixas de volumes de biomassa florestal acima do solo com a criação de modelos computacionais.

O artigo 1 (um) teve uma abordagem relacionada à classificação das tipologias florestais por meio de um modelo construído a partir de sensoriamento remoto e inteligência computacional. O artigo 2 (dois) teve como abordagem o estudo da classificação de biomassa por classes que possuem faixas de valores de volume de biomassa por meio de sensoriamento remoto e inteligência computacional.

1.2 – REFERÊNCIAS

ABRAMOVAY, R. Desenvolvimento sustentável: qual a estratégia para o Brasil? **Novos estudos-CEBRAP**, n. 87, p. 97-113, 2010.

ANDRADE, A. C.; FRANCISCO, C. N.; ALMEIDA, M. C. Desempenho de classificadores paramétrico e não paramétrico na classificação da fisionomia vegetal. **Anais XVII Simpósio Brasileiro de Sensoriamento Remoto**. INPE. 2015.

BECKER, B. K. Amazônia: mudança climática, projetos globais e interesse nacional. **Parcerias Estratégicas**, v. 18, n. 36, p. 107-128, 2013.

BECKER, B. K. Revisão das políticas de ocupação da Amazônia: é possível identificar modelos para projetar cenários?. **Parcerias estratégicas**, v. 6, n. 12, p. 135-159, 2001.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Manual técnico da vegetação brasileira. **Manuais Técnicos em Geociências,** Vol. 1, p. 16, 2012.

NEGRI, R. G.; SANT'ANNA, S. J. S.; DUTRA, L. V. Aplicação de Modelos de Aprendizado Semissupervisionado na Classificação de Imagens de Sensoriamento Remoto. **Revista de Informática Teórica e Aplicada**, v. 20, n. 2, p. 32-55, 2013.

PRATES, R. C.; BACHA, C. C. Os processos de desenvolvimento e desmatamento da Amazônia. **Economia e Sociedade**, v. 20, n. 3, p. 601-636, 2011.

SHIMABUKURO, Y. E.; MAEDA, E. E.; FORMAGGIO, A. R. Sensoriamento Remoto e Sistemas de Informações Geográficas aplicados ao estudo dos recursos agronômicos e florestais. **Ceres**, v. 56, n. 4, 2009.

SILVA, F.; PELLI, E.; PAULA, H.; CARVALHO, H.; NOGUEIRA, L.Aplicação do pré-processamento de imagens para otimização do reconhecimento de padrões na detecção de deficiência nutricional em espécies vegetais. **Anais XVII Simpósio Brasileiro de Sensoriamento Remoto**. INPE. 2015

SILVEIRA, P. Estimativa da biomassa e carbono acima do solo em um fragmento de floresta ombrófila densa utilizando o método da derivação do volume comercial. **Floresta**, v. 40, n. 4, 2010.

2- Classificação automática de tipologias florestais mediante técnica de inteligência computacional K- Vizinhos mais Próximos baseada em dados de sensoriamento remoto por satélite: Estudo de caso nas glebas Mamuru – Arapiuns do estado do Pará.

O artigo 1 será submetido a revista IEEE América Latina

RESUMO

O uso de tecnologias para pesquisas ambientais é relevante, pois garante monitoramento e gestão de grandes áreas. O sensoriamento remoto e a inteligência computacional merecem destaque pelas possibilidades de usos para estudos e análises ambientais. E a aplicação dessas tecnologias para a vegetação do bioma amazônico é válido e importante devido à sua dimensão. O objetivo do artigo é classificar as tipologias florestais das Glebas Estaduais Mamuru-Arapiuns com integração de dados do inventário florestal e técnicas de sensoriamento remoto e inteligência computacional K-vizinhos mais próximos. Foram utilizados imagens do satélite Lansat 5 e dados de inventário da região das Glebas Mamuru – Arapiuns para criação do banco de dados e a inteligência artificial do tipo K – vizinhos mais próximos foi utilizada para criar um modelo de classificação automática a partir do banco de dados. Com uma matriz de contingência gerada pela inteligência computacional foram calculadas acurácia global, índice Kappa e a análise da Característica de Operação do Receptor. Os resultados obtidos pela classificação automática com a técnica K - vizinhos mais próximos foram classificados como bons a partir dos valores obtidos pela acurácia global 87%, índice Kappa 74% e da área sob a curva do gráfico da Característica de Operação do Receptor de 0,929.
Palavras-chave: Amazônia. Inteligência artificial. Geoprocessamento.

ABSTRACT

The use of technologies for environmental research is important as it ensures monitoring and large areas of management. Remote sensing and computational intelligence worth mentioning the possibilities of uses for studies and environmental analysis. And the application of these technologies to the vegetation of the Amazon biome is valid and important because of its size. The paper aims to classify forest types of Mamuru-Arapiuns State Glebas with data integration of forest inventory and remote sensing techniques and computational intelligence closest K-neighbors. In the methodology we used satellite imagery Lansat 5 and inventory data in the area of Glebas Mamuru - Arapiuns for database creation and artificial intelligence type K - nearest neighbors was used to create an automatic classification model from the bank data. With contingency matrix generated by computational intelligence were calculated overall accuracy, Kappa index and the analysis of Receiver Operating Characteristic. The results obtained by automatic sorting with the technique K - nearest neighbors were classified as good from the values obtained for the overall accuracy of 87%, Kappa index 74% and the area under the graph curve of the Receiver Operating Characteristic of 0.929.
Key words: Amazon. Computational intelligence. Geoprocessing.

2.1- INTRODUÇÃO

A classificação automática de tipologias florestais a partir de sensoriamento remoto e inteligência computacional tem importância relevante, pois permite análise, monitoramento e gestão de áreas de acesso custoso e árduo, garantindo economia de tempo e recursos (ALMEIDA et al., 2009).

O sensoriamento remoto tem um papel fundamental no monitoramento do uso e cobertura da terra da Amazônia, pois permite obter informações históricas e atuais para um ambiente vasto e de difícil acesso (SHIMABUKURO et al., 2005). E o conhecimento da dinâmica de uso e cobertura da terra exerce papel importante para entender os fenômenos resultantes da atividade humana que ocorrem em áreas tropicais, principalmente na região da Amazônia brasileira (GARCIA et al., 2012).

Além disso, o monitoramento ambiental em escala regional pode ser realizado a partir de técnicas de sensoriamento remoto, as quais permitem analisar a relação entre padrões espaciais da vegetação e as mudanças no balanço de radiação e dos fluxos de energia da superfície (FAUSTO et al., 2014).

Porém, o manejo de ecossistemas florestais depende de informações precisas, completas e concisas a respeito da extensão, condição e produtividade dos recursos naturais (BAFFETTA, CORONA & FATTORINI, 2012). E a integração do inventário florestal e mapeamento surgiu como uma questão importante para avaliar atributos florestais e múltiplas funções ambientais (CORONA, 2010).

Outra ferramenta que se destaca em diferentes áreas do conhecimento é a Inteligência Computacional ou Inteligência Artificial que pode ser utilizada para aprendizagem e percepção. A inteligência artificial é um dos campos mais recentes em ciências e engenharia e é aplicada para diversas finalidades como jogos de xadrez, demonstrações de teoremas matemáticos, criação de poesia, direção de um carro em estrada movimentada, diagnóstico de doenças, entre outras(NORVIG & RUSSEL, 2014).

A inteligência computacional dos K- vizinhos mais próximos (K-NN) é um método que pode ser utilizado para predizer vários atributos em inventários florestais apoiados por dados de sensoriamento remoto (MCROBERTS, 2012; MCROBERTS & TOMPPO, 2007).

O presente artigo tem como objetivo classificar as tipologias florestais das Glebas Estaduais Mamuru-Arapiuns com a integração de dados do inventário

florestal e técnicas de sensoriamento remoto e inteligência computacional K-vizinhos mais próximos.

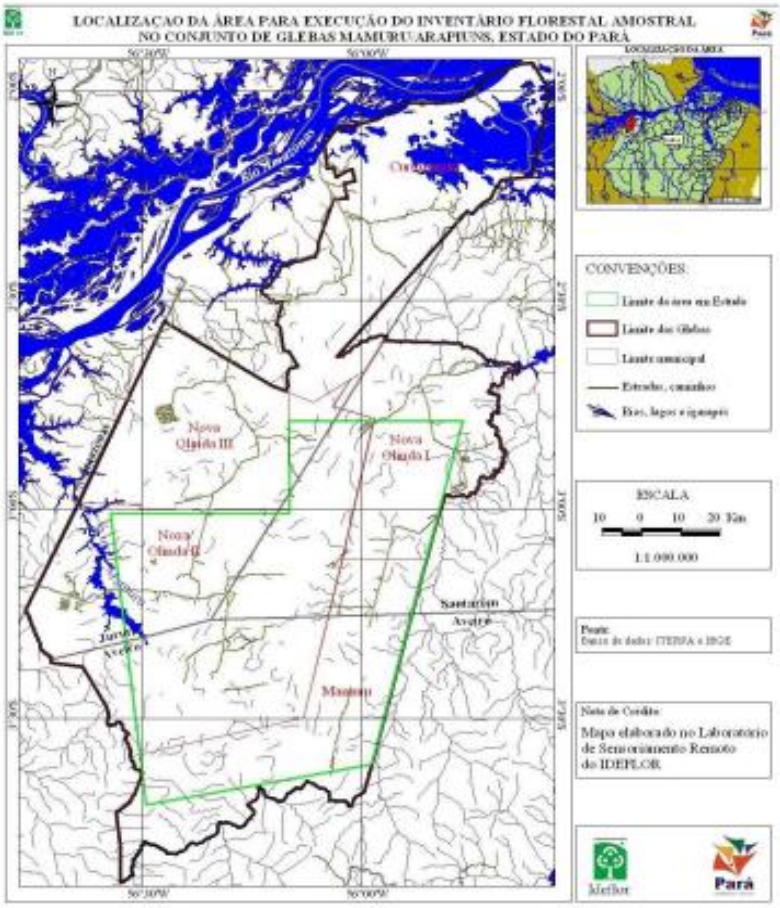
2.2- METODOLOGIA

Área de estudo

Os dados relativos ao estudo das Glebas Mamuru – Arapiuns foram obtidos junto ao Instituto de Desenvolvimento Florestal e da Biodiversidade (Ideflor-Bio), por meio de parceria com a Universidade do Estado do Pará (UEPA) formalizada em 2015.

A área deste estudo é o Conjunto de Glebas Estaduais Mamuru-Arapiuns, localizada entre os municípios de Santarém, Juruti e Aveiro, no Estado do Pará, abrangendo uma área aproximada de 600.000 hectares. O local de pesquisa fica inserido dentro dos limites das Glebas Nova Olinda I e II e a Gleba Mamuru, território sob responsabilidade Ideflor-Bio do Estado do Pará.

Figura 1: Mapa da área de estudo – Glebas Estaduais Mamuru Arapiuns



Fonte: Ideflor – Bio, 2010.

De acordo com o Instituto de Desenvolvimento Florestal do Pará – Ideflor- Bio (2010), a área de estudo apresenta dois tipos florestais predominantes classificados conforme o manual técnico da vegetação brasileira (IBGE, 2012) como Floresta Ombrófila Densa Terras baixas Dossel emergente (Dbe) e Floresta Ombrófila Densa Terras baixas Dossel emergente + Aberta com palmeiras (Dbe + Abp).

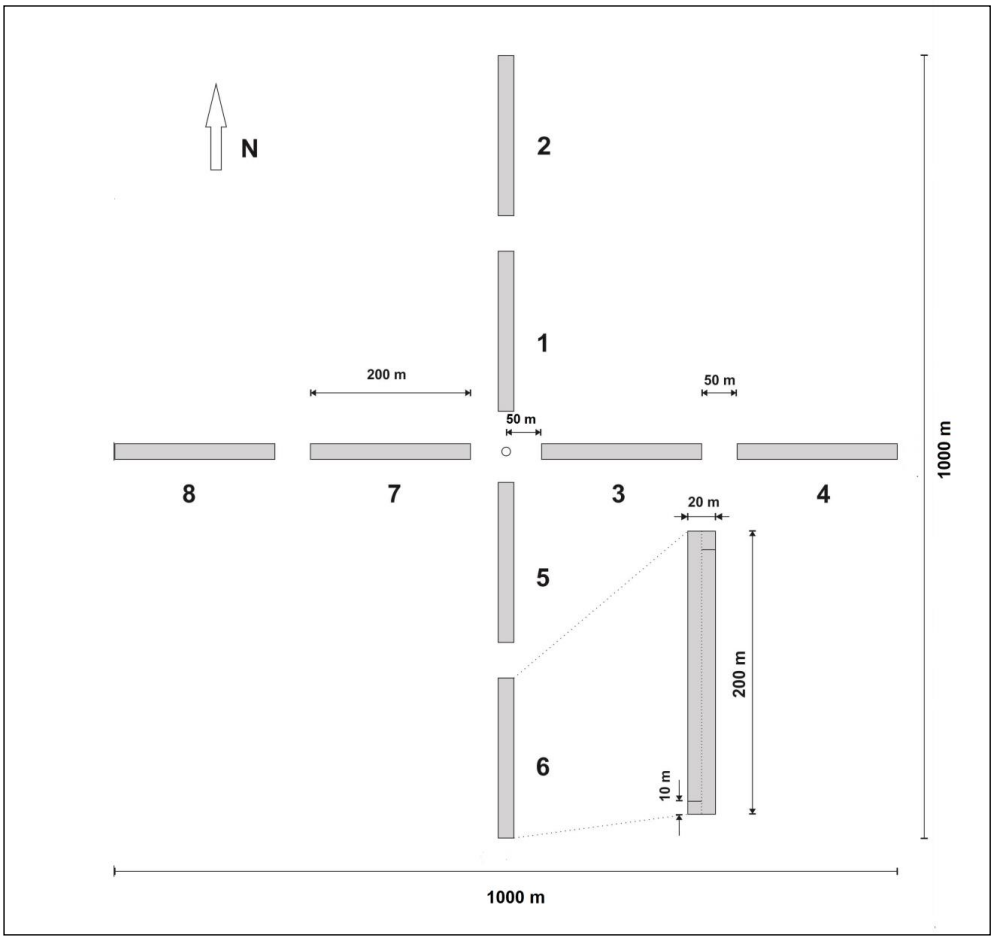
Inventário

Para o estudo a amostragem estratificada foi adotada para as tipologias florestais (estrato 1- Dbe e estrato 2 - Dbe + Abp**)**. Cada estrato possui 15 conglomerados distribuídos aleatoriamente na área. Somando-se os dois estratos há 30 conglomerados que representam as duas tipologias.

Cada conglomerado abrange uma área de 100 hectares (1.000 m x 1.000 m), composto por oito subunidades de 20 m x 200 m cada, alocadas sistematicamente a partir de um ponto central, sendo que a cada eixo cardinal (Leste-oeste; Norte-sul)

duas unidades foram alocadas, a primeira a 50 metros do ponto central, e a segunda a 50 metros da primeira (Figura 2).

Figura 2: Representação do conglomerado utilizado no inventário florestal



Fonte: Ideflor-Bio, 2010

Geoprocessamento e banco de dados

Foram obtidas imagens geradas pelo Mapeador Temático – TM do satélite Landsat 5, ano 2009, ano referente aos dados do inventario florestal, órbita 228 e ponto 62 e 63, junto à National Aeronautics and Space Administration (NASA, 2015). As imagens são compostas de sete bandas espectrais com resolução espacial de 30 m x 30 m, exceto a banda 6 (banda termal), com resolução de 120 m x 120 m.

Para o estudo foram utilizadas a banda 3 - Vermelho (0, 630 - 0, 690 цm), banda 4 - Infravermelho próximo (0, 760 - 0, 900 цm) e banda 5 - Infravermelho médio (1, 550 - 1, 750 цm), pois estão relacionadas com o comportamento espectral da vegetação que começa na região do visível e termina na região do infravermelho

médio e também porque as bandas 3 e 4 são utilizadas para o cálculo do Índice de Vegetação da Diferença Normalizada (NDVI – Normalized Difference Vegetation Index) que será utilizado no trabalho (PONZONI, SHIMABUKURO & KUPLICH, 2012).

No software QGIS 2.8.1 (2015) foi elaborado o mosaico da área de estudo e também a composição RGB (Red, Green and Blue) das imagens utilizadas. A partir do mosaico foi inserida as coordenadas geográficas do ponto central de cada conglomerado tanto do estrato 1 (Dbe) quanto do estrato 2 (Dbe + Abp), a partir desses pontos, foram extraídos o Digital Number (DN) das bandas 3, 4 e 5 das subunidades dos conglomerados de cada estrato.

Além dos valores dos pixels, o NDVI foi calculado com uso do software Microsoft Excel™ versão 2013, utilizando a seguinte fórmula proposta por Rouse et al., (1974):

NDVI = IVP – V / IVP + V (1)

Onde: IVP = valor do DN da banda 4 (infravermelho próximo) ;
V = valor do DN da banda 3 (vermelho).

O NDVI é a normalização do índice Razão Simples (Simple Ratio – SR), determinando o intervalo -1 a 1 aos seus valores (PONZONI, SHIMABUKURO & KUPLICH, 2012). O banco de dados para o trabalho é composto pelos valores das bandas 3, 4 e 5, coletados a partir do satélite Landsat 5, pelo índice de vegetação NDVI e a classe (estrato 1 ou estrato 2).

K- vizinhos mais próximos (K-Nearest Neighbor)

Existem vários classificadores automáticos baseados em inteligência computacional. Neste trabalho se utilizou o método baseado em distância chamado K-Nearest Neighbor ou K- vizinhos mais próximos (K-NN). Esta técnica não constrói representações explícitas das categorias, mas depende diretamente do cálculo da similaridade (ZAPATA-TAPASCO, et al., 2014).

O K-NN fundamenta-se em uma aprendizagem por similaridade e/ou analogia. K-vizinhos mais próximos é um tipo de classificador estatístico que tem uma execução simples e bons resultados (BARROS et al., 2011). A implementação

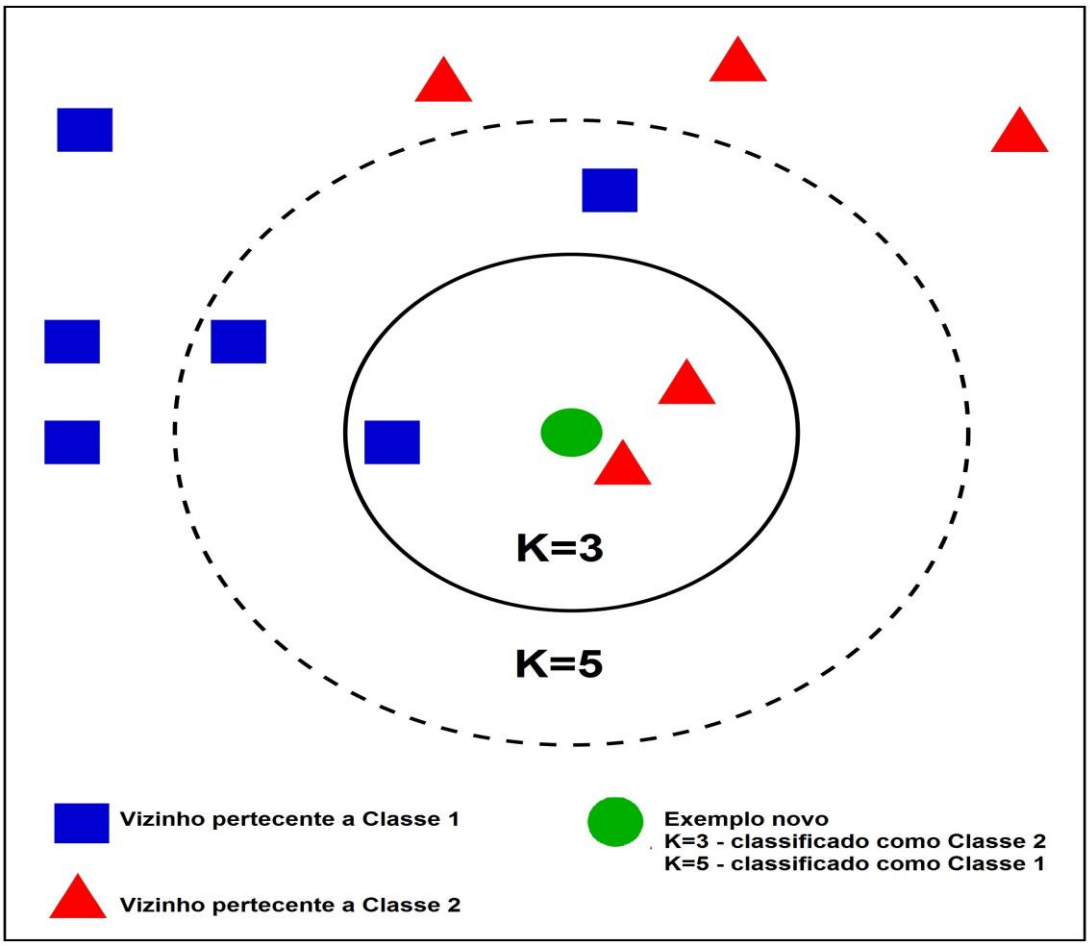
do K-NN requer duas seleções principais que são: a distância métrica e um valor para K.

Para a aplicação da técnica K-NN utilizou-se uma amostra de 500 exemplos extraídos aleatoriamente da área de estudo, sendo 250 exemplos pertencentes ao Estrato 1 (Dbe) e 250 exemplos pertencentes ao Estrato 2 (Dbe + Abp). Cada exemplo possui quatro atributos preditivos, que são: número digital (digital number) da banda 3, da banda 4 e da banda 5, além do valor do NDVI. Para a incorporação do banco de dados de armazenamento do K-NN foram utilizados 400 exemplos (sendo 200 pertencentes ao Estrato 1 e 200 pertencentes ao Estrato 2) e para o teste da inteligência computacional foram utilizados os 100 exemplos remanescentes (sendo 50 exemplos do Estrado 1 e 50 exemplos do Estrato 2). A métrica utilizada foi a distância Euclidiana. Considere os exemplos X1 = (x11, x12, ..., x1n) e X2 = (x21, x22, ..., x2n), ambos possuem *n* atributos e a distância Euclidiana entre X1 e X2 é calculada pela equação apresentada abaixo:

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad (2)$$

Para classificar um exemplo novo o método procura os K exemplos já armazenados que apresentam valores mais próximos dos valores do exemplo novo. Pois, dado certo padrão desconhecido X, sua classe é calculado pela distância entre X e os outros padrões de formação. E o padrão desconhecido X a partir das distâncias dos vizinhos mais próximos será classificado em um determinado grupo.
Além da distância entre os exemplos, existe também a relação entre o valor de K, que é o número de vizinhos mais próximos escolhido para classificação (Figura 3).

Figura 3: Esquema de classificação de um exemplo novo com valor K



Fonte: Elaborada pela autora

Segundo Barros et al.(2011), o método do vizinho mais próximo pode ser prolongado, utilizando não um, mas um conjunto de dados mais próximos para prever o valor dos novos dados, em que é conhecido como os K-vizinhos mais próximos. Ao considerar mais do que um vizinho, é fornecido imunidade a ruídos e a curva de estimação é suavizada.

O número de vizinhos mais próximos (K-NN) foi escolhido a partir de testes com o banco de dados e os valores de k testados foram 3, 5, 7, 9, 11, 13 e 15, todos ímpares, pois o K-NN compara distâncias entre o exemplo novo e os exemplos do banco de dados de armazenamento e segundo Ferrero (2009) números de vizinhos ímpares excluem empates na classificação no novo exemplo e k pares incluem empates.

A partir dos exemplos do banco de dados do K-NN e assimilação das distâncias do novo exemplo com os vizinhos mais próximos, a inteligência

computacional irá classificar o novo exemplo ou estrato 1(Dbe) ou estrato 2(Dbe + Abp).

Com utilização do software ORANGE 2.7 (2015), o banco de dados de armazenamento foi utilizado para classificação automática dos exemplos de teste com utilização da ferramenta K-NN.

A partir da matriz de contingência (ou matriz de confusão) foram calculados a acurácia global, o índice Kappa e a análise Característica de Operação do Receptor (Receiver Operating Characteristic-ROC).

Quadro 1: Fórmulas utilizadas para os cálculos da acurácia global e do índice Kappa

Nome	Fórmula	Referência
Acurácia Global	$G = \frac{\sum_{i=1}^{c} x_{ii}}{n}$	Story e Congalton (1986)
Índice de Kappa	$K = \frac{n\sum_{i=1}^{c} x_{ii} - \sum_{i=1}^{c} x_{i+} x_{+i}}{n^2 - \sum_{i=1}^{c} x_{i+} x_{+i}}$	Rosenfield e Fitzpatrick-Lins (1986)

Acurácia Global (G): a soma da diagonal principal da matriz de contingência *xii* pelo número total de amostras coletadas *n.* Índice de Kappa (K): *xii* é o valor na linha *i* e coluna *i*; *xi +* é a soma da linha *i* e *x+i* é a soma da coluna *i* da matriz de contingência; *n* é o número total de amostras e *c* o número total de classes.

O valor da acurácia é calculado levando em consideração os acertos da classificação e o índice Kappa considera toda a matriz de contingência no seu cálculo (SARMIENTO et al., 2014). O índice kappa é uma proporção de acerto depois da eliminação dos casos de acerto por acaso (ROSENFIELD & FITZPATRICK LINS, 1986; PANTALEÃO E SCOTFIEL, 2009). No Quadro 2, é exposto o grau de concordância do índice Kappa.

Quadro 2: Grau de concordância do índice Kappa

Índice Kappa	**Concordância**
< 0	Sem concordância

0,00 a 0,19	Pobre
0,20 a 0,39	Fraca
0,40 a 0,59	Moderada
0,60 a 0,79	Forte
0,80 a 1,00	Excelente

Fonte: Landis e Koch (1977).

E a análise ROC foi utilizada para investigar a sensibilidade e a especificidade dos dados testados no K-NN, ou seja, a relação entre verdadeiros positivos e falsos positivos. A sensibilidade é a porcentagem de exemplos positivos que foram classificados corretamente, já a especificidade é a porcentagem de exemplos negativos que foram classificados corretamente (NETO, RODRIGUES & MEIRA, 2014).

Uma das interpretações está relacionada a área sob a curva (Area Under Curve - AUC). A AUC é uma fração da área abaixo da curva do gráfico ROC, como cada lado do gráfico ROC varia de 0 a 1, a área do quadrado será 1 (um), ou seja, a área abaixo da curva só pode variar de 0 a 1 (PRATI, BATISTA & MONARD, 2008). No gráfico ROC a área sob a curva é a medida de desempenho do modelo (FIGUEIREDO, 2015). No quadro 3 está presente a relação entre AUC e o grau de desenpenho do método de classificação automática.

Quadro 3: Relação entre valores de área sob a curva (AUC) e o grau de desempenho do modelo de classificação

Área sob a curva (AUC)	**Desempenho**
Acima de 0,9	Excelente
0,8 – 0,9	Bom
0,7 – 0,8	Regular
0,6 – 0,7	Ruim
0,5 – 0,6	Reprovado

Fonte: Câmara, 2009.

2.3- RESULTADOS E DISCUSSÃO

Teste do número de vizinhos mais próximos

A partir dos testes dos 100 exemplos (B3, B4, B5 e NDVI) com os 400 exemplos (B3, B4, B5, NDVI e Estrato) do banco de dados de armazenamento foi possível observar que o número de vizinhos mais próximos com melhores resultados de acurácia são 9, 11 e 13. Para o trabalho foi selecionado o teste K=13, pois tem acurácia e área sob a curva (AUC) com números mais elevados. Na tabela 1 está presente os valores adquiridos a partir dos testes efetuados.

Tabela 1: Valores de Acuracia Global e AUC nos K-vizinhos mais próximos testados

	K=3	**K=5**	**K=7**	**K=9**	**K=11**	**K=13**	**K=15**
Acurácia	0,86	0,86	0,86	0,87	0,87	**0,87**	0,86
Área sob a curva (AUC)	0,908	0,912	0,9176	0,923	0,9258	**0,9292**	**0,936**

Matriz de confusão, acurácia e índice Kappa

Para a avaliação dos resultados das classificações automáticas foram utilizados os coeficientes de acurácia global, índice de Kappa, ambos gerados a partir da matriz de contingência. Pois, essa matriz compara, classe por classe, a relação entre os dados de referência conhecidos e os resultados correspondentes de uma classificação automatizada (LILLESAND et al., 2004).

No estudo, 100 exemplos foram comparados a partir da distância euclidiana aos 13 vizinhos mais próximos dos 400 exemplos do banco de dados de referência do K-NN. Dos 100 exemplos, 50 pertencem ao estrato 1 (Dbe) e 50 ao estrado 2 (Dbe + Abp), a partir da distância euclidiana os 100 exemplos foram classificados automaticamente em estrato 1 ou em estrato 2.

Na tabela 2 está a matriz de contingência da classificação automática realizada pela técnica K-NN com número de vizinhos mais próximos (K) = 13.

Tabela 2 - Matriz de Contingência

	Estrato 1	Estrato 2	
Estrato 1	45	8	53
Estrato 2	5	42	47
	50	50	100

Colunas representam classes reais e linhas representam as predições

A partir do processamento de análise dos 100 exemplos sem identificação do tipo de estrato a qual pertence, o K-NN revelou acurácia global de 0,87 ou 87% e o índice Kappa de 0,74 ou 74%.

Análise ROC

As figuras 4 e 5 são referentes aos gráficos com as curvas ROC dos estratos 1 (Dbe) e 2 (Dbe + Abp). A figura 4 representa Floresta Ombrófila Densa de Terras baixas Dossel Emergente, com curva acima da diagonal e ligeiramente acentuada para o canto superior esquerdo. A Figura 5 representa Floresta Ombrófila Densa Terras baixas Dossel emergente + Aberta com palmeiras que possui curva com características semelhantes à Figura 4

Figura 4: Gráfico ROC do Estrato 1 – Floresta Ombrófila Densa Terras baixas Dossel emergente (Dbe)

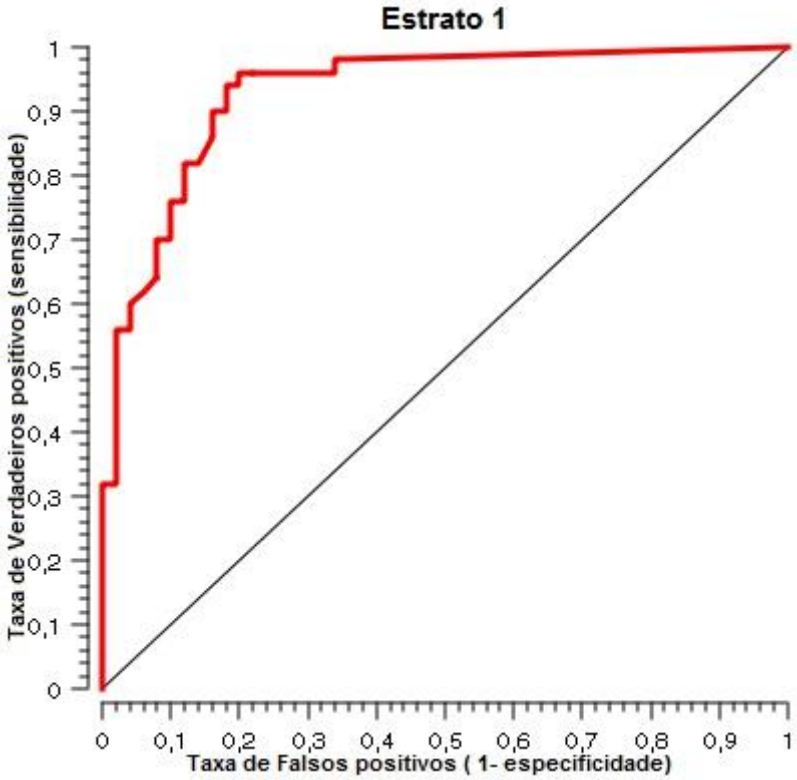
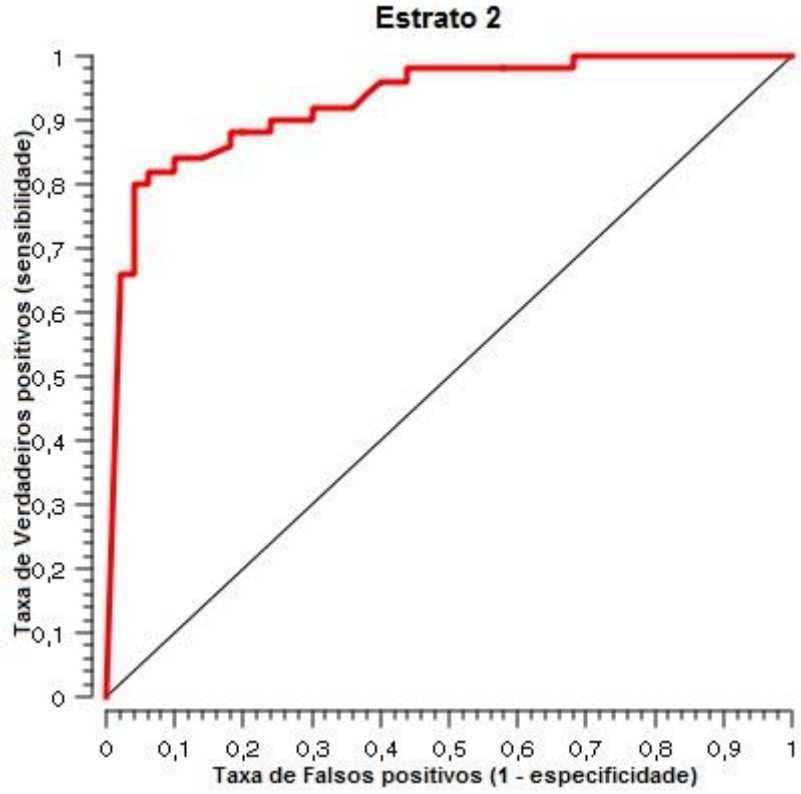


Figura 5: Gráfico ROC do Estrato 2 – Floresta Ombrófila Densa Terras baixas Dossel emergente + Aberta com palmeiras (Dbe + Abp).



Tanto a figura 4 quanto a figura 5 apresentaram gráficos com curvas acima da diagonal e também pontos próximos ao canto superior esquerdo (0,1), porém a figura 5 possui sua curva mais acentuada para o ponto (0,1), essas curvas demonstram que o modelo de classificação proposto neste trabalho é melhor que o aleatório. E a AUC da análise do modelo de classificação foi de 0,929, valor muito próximo de 1.

Nos testes efetuados com os dados obtidos a partir do inventário florestal e das imagens de satélite, o valor mais apropriado de vizinhos foi 13, por obter porcentagem maior de acurácia e também área sob a curva (AUC) com maior valor, K= 15 apresentou AUC maior que K=13, porém sua acurácia global foi menor. Segundo Ferrero (2009) não existe um único valor de K que seja apropriado para variados problemas, e que, o valor de K deve ser avaliado para cada problema particular.

Segundo a escala do índice Kappa (LANDIS & KOCK, 1977), no intervalo de 0,60 a 0,79 o grau de concordância é forte. O coeficiente de concordância kappa é frequentemente utilizado como uma medida geral de precisão (OLOFSON et al.,

2012). O índice possui porcentagem menor que a acurácia global, pois em seu cálculo é considerado os erros e os acertos da matriz de contingência.

E quanto mais próxima de 1 e mais distante de 0,50 for a AUC, maior será a acurácia do modelo (FRANKLIN, 2010). E segundo a classificação do desempenho de Câmara (2009), o modelo de classificação dos estratos 1(Dbe) e 2 (Dbe + Abp) com utilização do K-NN possui desempenho considerado bom. Além disso, a AUC também é numericamente igual à probabilidade de dados dois exemplos de classes distintas, o exemplo positivo seja ordenado primeiramente que um exemplo negativo (PRATI, BATISTA & MONARD, 2008).

No estudo todas as porcentagens apresentaram valores considerados bons, assim como no trabalho de Thessler et al. (2008) que apresentou acurácias globais nos valores de 91% e 89% com utilização de tecnologias de sensoriamento remoto e K-NN para classificação em florestas tropicais. Evidenciando que a utilização dessas tecnologias associadas a dados de campo são eficientes para classificação de tipologias florestais.

2.4- CONCLUSÃO

O modelo de classificação tem bons indicadores estatísticos como acurácia global igual a 87% e índice Kappa igual a 74%, e pode ser utilizado para classificar tipologias florestais, apesar dos erros serem aceitáveis com as porcentagens apresentadas quando relacionados aos estudos ambientais, sempre será interessante e importante a diminuição dos erros e melhoramentos do modelo para garantir mais fidelidade para com a realidade do ambiente estudado.

2.5- REFERÊNCIAS

ALMEIDA, A. C.;BARROS, P. L. C.; MONTEIRO, J. H. A.; ROCHA, B. R. P. Estimation of aboveground forest biomass in Amazonia with neural networks and remote sensing. **IEEE Latin Amer. Trans**, v. 7, n. 1, p. 27-32, 2009.

BAFFETTA, F.; CORONA, P.; FATTORINI, L. A matching procedure to improve k-NN estimation of forest attribute maps.**Forest ecology and management**, v. 272, p. 35-50, 2012.

BARROS, J. D.; OLIVEIRA JÚNIOR, J. J. D.; SILVA, S. G. D.; FARIAS, R. F. D. Characterization of bone Tissue by microwaves using wavelets and KNN. **Journal of**

Microwaves, Optoelectronics and Electromagnetic Applications, v. 10, n. 1, p. 217-231, 2011.

CÂMARA, F.P. Psiquiatria E Estatística V: Validação De Procedimentos Diagnóstica Pela Curva R.O.C. **Psychiatry on line Brasil**. Vol.14. Nº 4. 2009.

CORONA, P. Integration of forest mapping and inventory to support forest management. **iForest-Biogeosciences and Forestry**, v. 3, n. 3, p. 59-64, 2010.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B Orange: Data Mining Toolbox in Python. **Journal of Machine Learning Research** V.14. p: 2349−2353. 2013.

FAUSTO, M. A.; MACHADO, N. G.; NOGUEIRA, J. S.; BIUDES, M. S. Net radiation estimated by remote sensing in Cerrado areas in the Upper Paraguay River Basin. **Journal of Applied Remote Sensing**, v. 8, n. 1, p. 083541-083541, 2014.

FERRERO, Carlos Andres. Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia. 2009. Tese de Doutorado. Universidade de São Paulo.

FIGUEIREDO, S. M. D. M.; VENTICINQUE, E. M.; FIGUEIREDO, E. O.; FERREIRA, E. J. L. Predicting the distribution of forest tree species using topographic variables and vegetation index in eastern Acre, Brazil. **Acta Amazonica**, v. 45, n. 2, p. 167-174, 2015.

FRANKLIN, J. **Mapping species distributions:** spatial inference and prediction. Cambridge University Press, 2010.

GARCIA, C. E.; DOS SANTOS, J. R.; MURA, J. C.; KUX, H. J. H. Análise do potencial de imagem TerraSAR-X para mapeamento temático no sudoeste da Amazônia brasileira. **Acta Amazônica**, v. 42, n. 2, p. 183-192, 2012.

HAN, J.; KAMBER, M. **Data Mining**: concepts and techniques. 2. ed. San Francisco.Elsevier, 2006.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Manual técnico da vegetação brasileira**. Manuais Técnicos em Geociências, Vol. 1. 2ª Edição. 2012.

Instituto de Desenvolvimento Florestal do Pará – Ideflor- Bio. **Inventário Florestal Diagnóstico Do Conjunto De Glebas Estaduais Mamuru-Arapiuns – Pará**: Relatório Final. 2010.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, p. 159-174, 1977.

LANDSAT, N. A. S. A. Science Data Users Handbook. **2015. http://landsathandbook. gsfc. nasa. gov/inst_cal/prog_sect8_2. ht**
LILLESAND, T.M.; KIEFER, R.W.; CHIPMAN, J.W. **Remote sensing and image interpretation**. 5.ed. Madison: Wiley, 2004. 763p.

LILLESAND, T. M.; KIEFER, R. W.; CHIPMAN, J.W. **Remote sensing and image interpretation**. John wiley and Sons, 2004.

MCROBERTS, R. E. Estimating forest attribute parameters for small areas using nearest neighbors techniques. **Forest Ecology and Management**, v. 272, p. 3-12, 2012.

MCROBERTS, R. E.; TOMPPO, E. O. Remote sensing support for national forest inventories. **Remote Sensing of Environment**, v. 110, n. 4, p. 412-419, 2007.

NETO, C. G.; RODRIGUES, L. H. A.; MEIRA, C. A. A. Modelos de predição da ferrugem do cafeeiro (Hemileia vastatrix Berkeley & Broome) por técnicas de mineração de dados. Coffee Science, v. 9, n. 3, p. 408-418, 2014.

NORVIG, P.; RUSSELL, S. **Inteligência Artificial.** 3ª Edição. Elsevier Brasil, 2014.
PANTALEÃO, E; SCOFIELD,G. Comparação entre medidas de acurácia de classificação para imagens do satélite ALOS**. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto**, Natal, Brasil, 25-30 abril 2009, INPE, p. 7039-7046. 2009.

PONZONI, F. J.; SHIMABUKURO, Y. E.; KUPLICH, T. M. **Sensoriamento remoto da vegetação**. Oficina de Textos, 2012.

PRATI, R. C.; BATISTA, GEAPA; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215-222, 2008.

QGIS DEVELOPMENT TEAM et al. QGIS **Geographic Information System**. Open Source Geospatial Foundation Project. 2015.

ROSENFIELD, G. H.; FITZPATRICK-LINS, K. A coefficient of agreement as a measure of thematic classification accuracy.**Photogrammetric engineering and remote sensing**, v. 52, n. 2, p. 223-227, 1986.

ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. **NASA special publication**, v. 351, p. 309, 1974.

SARMIENTO, C. M.; RAMIREZ, G. M.; COLTRI, P. P.; LIMA, L. F.; NASSUR, O. A. C.; SOARES, J. F. Comparação de classificadores supervisionados na discriminação de áreas cafeeiras em Campos Gerais-Minas Gerais. **Coffee Science**, v. 9, n. 4, p. 546-557, 2014.

SHIMABUKURO, Y.E.; DUARTE, V.; MOREIRA, M.A.; ARAI, E.; RUDORFF, B.F.T.; ANDERSON, L.O.; ESPÍRITO-SANTO, F.D.B.; FREITAS, R.M.; AULICINO, L.C.M.; MAURANO, L.E.; ARAGÃO, J.R.L. Detecção de áreas desflorestadas em tempo real: conceitos básicos, desenvolvimento e aplicação do Projeto DETER.**São José dos Campos: INPE**, 2005.

STORY, M.; CONGALTON, R. G. Accuracy assessment - A user's perspective. **Photogrammetric Engineering and remote sensing**, v. 52, n. 3, p. 397-399, 1986.

THESSLER, S.; SESNIE, S.; BENDAÑA, Z. S. R.; RUOKOLAINEN, K.; TOMPPO, E.; FINEGAN, B. Using k-nn and discriminant analyses to classify rain forest types in a Landsat TM image over northern Costa Rica. **Remote Sensing of Environment**, v. 112, n. 5, p. 2485-2494, 2008.

ZAPATA-TAPASCO, A.; MORA-FLÓREZ, J.; PÉREZ-LONDOÑO, S. Hybrid methodology based on knn regression and boosting classification techniques for locating faults in distribution systems. **Ingeniería y competitividad**, v. 16, n. 2, p. 165-177, 2014.

3- Utilização de técnicas de sensoriamento remoto e K-vizinhos mais próximos para classificação em intervalos de valores de Biomassa Florestal na região Amazônica: estudo de caso nas Glebas Mamuru- Arapiuns, Pará

Resumo

Em relação às florestas, o método mais tradicional de levantamento de aspectos de uma vegetação são os inventários florestais que são elaborados baseados em levantamentos de campo. E atualmente, metodologias que unem sensoriamento e inteligência computacional em áreas florestais são realizadas de diferentes formas para estimar e classificar variáveis florestais. O objetivo do trabalho é classificar de forma automatizada valores de quantidade de biomassa presente no Conjunto de Glebas Mamuru-Arapiuns com utilização de inventário florestal e técnicas de sensoriamento remoto e de inteligência computacional do tipo K-vizinhos mais próximos. Na metodologia foram utilizados dados de sensoriamento remoto e de inventário florestal para a criação de um banco de dados para aplicação da inteligência computacional do tipo K-vizinhos mais próximos, e assim, foram criadas classes de faixas de volume de biomassa florestal e um método de classificação automática foi desenvolvido. Entre os tipos de distâncias métricas utilizadas não houve predominância de uma métrica específica. Os valores de vizinhos mais próximos com melhores resultados foram número de vizinhos igual a 3 para as classes de volume de biomassa o estrato 1 e 5 para o estrato 2. Nos resultados houve comparações entre os valores de acurácia global, medida F, precisão, revocação e área sob a curva.

Palavras-chave: Medida F. Amazônia. Inteligência Computacional. Análise ROC.

Abstract

With regard to forests, the more traditional method of raising aspects of vegetation are forest inventories that are developed based on field surveys. And currently, methodologies that combine sensing and computational intelligence in forest areas are carried out in different ways to estimate and classify forest variables. The objective is to sort of way automated quantity of biomass present values in Glebas Set Mamuru-Arapiuns with use of forest and remote sensing inventory and computational intelligence type nearest K-neighbors. In the methodology we used remote sensing data and forest inventory for the creation of a database for the application of computational intelligence type nearest K-neighbors, and so classes were created from forest biomass volume ranges and a method of automatic classification was developed. Among the types of metric distances used there was no predominance of a specific metric. The values of nearest neighbors with best results were number of neighbors is equal to 3 for biomass volume of classes 1 and the layer 5 to layer 2. In comparisons between the results was overall accuracy values as F, precision, recall and area under the curve.

Key words: F score. Amazônia. Inteligência Computacional. ROC analysis.

3.1- INTRODUÇÃO

Os estudos de uso e cobertura das terras por meio dos dados de sensoriamento remoto e ferramental do Sistema de Informações Geográficas - SIG têm proporcionado rapidez nos resultados de mapeamento da superfície terrestre e avaliação das mudanças ocorridas ao longo do tempo (RODRIGUES & HOTT, 2010).

O mapeamento é realizado por meio da classificação de imagens orbitais que pode ser realizada através de análise visual ou automática (SARMIENTO et al., 2014). E a partir dessas geotecnologias, grandes extensões de terras podem ser mapeadas quanto ao seu uso e cobertura, possibilitando a tomada de decisões para planejamento de ações (RODRIGUES & HOTT, 2010).

Em relação às florestas, o método mais tradicional de levantamento de aspectos de uma vegetação são os inventários florestais. Estes são baseados em levantamentos de campo em sua maioria são de difícil realização e caros, principalmente quando realizados em grandes áreas e, em particular, quando as florestas são localizadas em áreas remotas e de difícil acesso (ALVES et al., 2013).

Os sistemas computacionais que procuram explorar a inteligência artificial, também chamada de inteligência computacional, baseiam-se na inteligência humana em realizar determinadas tarefas, aprender novos procedimentos e decisões, entender linguagens e resolver problemas com as técnicas do raciocínio (URNAU, KIPPER & FROZZA, 2014). E a utilização de técnicas de sensoriamento remoto pode ser uma ferramenta muito interessante para a estimativa de variáveis florestais, a um custo relativamente baixo, com precisão aceitável e com tempo relativamente reduzido (ALVES et al., 2013).

Metodologias que unem sensoriamento e inteligência computacional em áreas florestais são realizadas de diferentes formas como demonstra os trabalhos de Alves et al. (2013) que utiliza a inteligência computacional K- vizinhos mais próximos em imagens de satélite para estimar padrões florestais, Sarmiento et al. (2014) aplica mais de um tipo de inteligência computacional em imagens de satélite para comparar classificadores em áreas cafeeiras, Garofalo et al. (2015) faz análise comparativa de classificação de imagens do satélite Landsat 8, Amaral et al. (2012) realiza em seu estudo um aprimoramento de classificação de séries temporais por geoprocessamento e Santos et al. (2015) que analisa a fragmentação da paisagem por sensoriamento remoto.

A inteligência computacional utilizada no trabalho é a técnica do K-vizinhos mais próximos (KNN). O algoritmo de classificação KNN, proposto por Fukunaga e Narendra (1975), é uma técnica empregada no reconhecimento de padrões, baseada na técnica do vizinho mais próximo (nearest neighbor- NN), que utiliza os 'k' vizinhos mais próximos do padrão de consulta, ao invés de apenas o vizinho mais próximo (XU et al., 2013).

O presente trabalho tem como objetivo classificar em intervalos de forma automatizada o volume de biomassa florestal existente no Conjunto de Glebas Mamuru-Arapiuns, localizadas no estado do Pará, com utilização de inventário florestal e técnicas de sensoriamento remoto e de inteligência computacional do tipo K-vizinhos mais próximos.

3.2- METODOLOGIA

Área de estudo

Os dados do trabalho foram adquiridos através de parceria entre o Instituto de Desenvolvimento Florestal e da Biodiversidade (Ideflor-Bio) e a Universidade do Estado do Pará (UEPA) no ano de 2015.

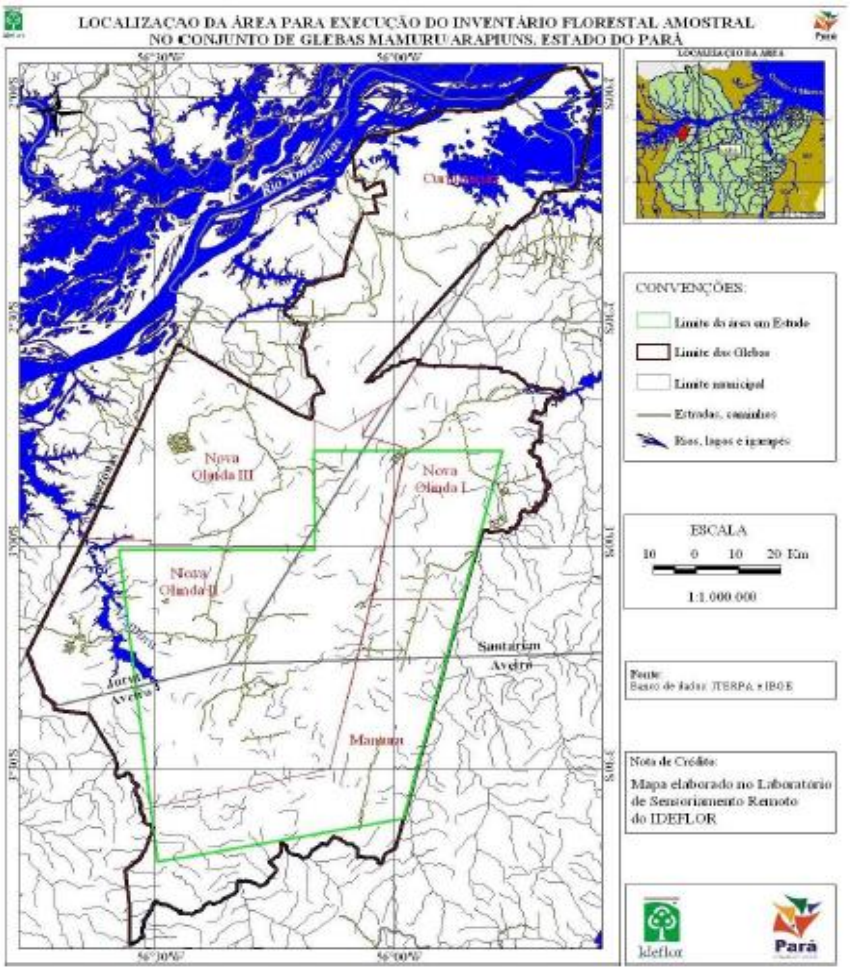
A área de estudo é o Conjunto de Glebas Estaduais Mamuru-Arapiuns, localizada entre os municípios de Santarém, Juruti e Aveiro, no Estado do Pará, abrangendo uma área aproximada de 600.000 hectares. O local de pesquisa fica inserido dentro dos limites das Glebas Nova Olinda I e II e a Gleba Mamuru. O conjunto de glebas Mamuru-Arapiuns, na região do Baixo Amazonas, foi a primeira área a passar pelo processo de concessão florestal no estado e o local está sob responsabilidade do Instituto de Desenvolvimento Florestal do Pará – Ideflor- Bio.

De acordo com o Ideflor- Bio (2010), a área de estudo apresenta dois tipos florestais predominantes conforme o manual técnico da vegetação brasileira (IBGE, 2012), sendo estes: Floresta Ombrófila Densa Terras baixas Dossel emergente (Dbe); Floresta Ombrófila Densa Terras baixas Dossel emergente + Aberta com palmeiras (Dbe + Abp).

Na área das glebas existem dois tipos de solos que são Latossolo Amarelo e Gleissolo Háplico. E o clima da região é do tipo Amw de Köppen, caracterizado como quente e úmido e apresenta duas estações bem definidas: uma chuvosa, de janeiro a julho e outra seca, de agosto a dezembro. A temperatura anual varia entre

25°C e 28°C com média anual de precipitação pluviométrica em torno de 1.900 mm (Ideflor-Bio, 2010).

Figura 1 – Mapa de localização da área de estudo



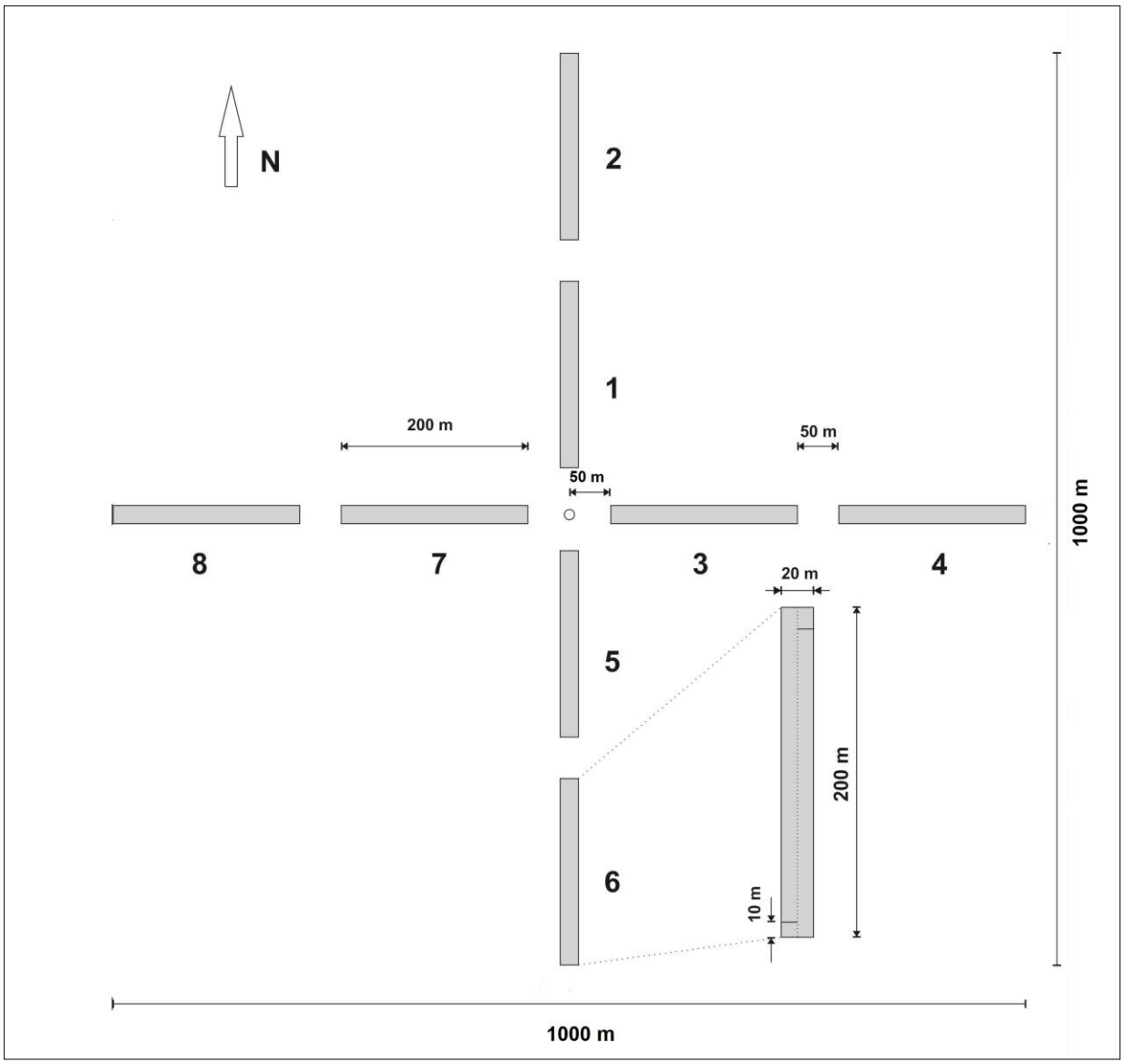
Fonte: Ideflor-Bio (2010).

Inventário

A Amostragem Estratificada foi adotada para as tipologias florestais predominantes da área (estrato 1- Dbe e estrato 2 - Dbe + Abp**)**, onde cada estrato possui 15 conglomerados, totalizando 30 conglomerados utilizados no estudo.

Cada conglomerado abrange uma área de 100 hectares (1.000 m x 1.000 m), composto por oito subunidades de 20 x 200 m cada, alocadas sistematicamente a partir de um ponto central, sendo que a cada eixo cardinal (Leste-oeste; Norte-sul) duas unidades foram alocadas, a primeira a 50 metros do ponto central, e a segunda a 50 metros da primeira (Figura 2).

Figura 2: Ilustração da amostragem estratificada por conglomerados



Fonte: Ideflor-Bio (2010)

Com esta metodologia, foi calculado o volume em m³ da biomassa presente em cada subunidade de cada conglomerado (Tabela 1), e para o cálculo foi utilizada uma equação de volume. O volume que está presente no relatório do Ideflor-Bio foi calculado a partir da seguinte fórmula proposta por Silva et al.,(1984):

$$\ln V = -7{,}62812 + 2{,}18090 \ln DAP \quad (1)$$

Onde: V= Volume;
DAP = Diâmetro na Altura do Peito;

Tabela 1: Volumes de biomassa apresentados no relatório do inventário florestal

Estrato	Conglomerado	Subunidade	Volume (m³)
1	1	1	341,3
1	1	2	331,3
1	1	3	266,9
1	1	4	321,1
1	1	5	275,3
1	1	6	396,7
1	1	7	363,6
1	1	8	325,4

Fonte: Adaptação de Ideflor-Bio (2010).

Geoprocessamento e banco de dados

Foram obtidas imagens geradas pelo Mapeador Temático (TM) do satélite Landsat 5, ano 2009, pois os dados do inventário florestal são referentes a este ano, órbita 228 e ponto 62 e 63, junto à National Aeronautics and Space Administration - NASA (2015). As imagens são compostas de sete bandas espectrais com resolução espacial de 30 m x 30 m, exceto a banda 6 (banda termal), com resolução de 120 m x 120 m.

Para o estudo foi utilizadas a banda 3 - Vermelho (0, 630 - 0, 690 цm), banda 4 - Infravermelho próximo (0, 760 - 0, 900 цm) e banda 5 - Infravermelho médio (1, 550 - 1, 750 цm), pois as três bandas estão relacionadas com o comportamento espectral da vegetação e também porque as bandas 3 e 4 são utilizadas para o cálculo do Índice de Vegetação da Diferença Normalizada (NDVI – Normalized Difference Vegetation Index) que será utilizado no trabalho (PONZONI, et al., 2012).

No software QGIS (2012) foi elaborado o mosaico da área de estudo e também a composição RGB (Red, Green and Blue) das imagens utilizadas. A partir do mosaico foi inserida as coordenadas geográficas do ponto central de cada conglomerado tanto do estrato 1 (Dbe) quanto do estrato 2 (Dbe + Abp), a partir desses pontos, foram extraídos os DN's (digtal numbers), que são os valores existentes nos pixels. das bandas 3, 4 e 5 das subunidades dos conglomerados de cada estrato.

Além dos valores dos pixels, o NDVI foi calculado com uso do software Excel™ da Microsoft Office, utilizando a seguinte fórmula proposta por Rouse *et al*., 1973:

$$NDVI = (IVP - V) / (IVP + V) \quad (2)$$

Onde: IVP = DN da banda 4 (infravermelho próximo) ;
V = DN da banda 3 (vermelho).

O NDVI foi calculado utilizando o valor DN (digital number). Esse índice de vegetação é a normalização do índice Razão Simples (Simple Ratio – SR), determinando o intervalo -1 a 1 aos seus valores (PONZONI *et al*., 2012).

O volume presente no relatório do Ideflor-Bio é correspondente a cada subunidade e o volume foi dividido proporcionalmente levando em consideração os valores das bandas 3, 4 e 5 e do NDVI (Tabela 2). Cada subunidade apresenta 7 (sete) exemplos, e cada exemplo possui os valores (DN) das bandas 3, 4,5, o valor do NDVI, o estrato ao qual pertence e o valor de volume dividido proporcionalmente.

Tabela 2: Exemplo de como os volumes se apresentaram diluídos em cada subunidade

Estrato	Cong.	Sub.	Banda 3	Banda 4	Banda 5	NDVI	Volume (m³)
1	1	2	23	106	73	0,643	47,026
1	1	2	23	108	74	0,649	47,424
1	1	2	24	106	74	0,631	46,102
1	1	2	24	104	76	0,625	45,680
1	1	2	23	116	72	0,669	48,901
1	1	2	23	109	70	0,652	47,618
1	1	2	23	114	73	0,664	48,548

Para criar classes de volumes de biomassa, os valores foram colocados em intervalos, e a cada intervalo lhe foi dada uma nomenclatura. No estrato 1, exemplos com volume entre 20m³ e 40 m³ foram classificados como FV1 e exemplos com volume entre 40m³ e 60m³ foram classificados como FV2. No estrato 2, exemplos com volume entre 20m³ e 40m³ foram classificados como FVA e exemplos com volume entre 40m³ e 60m³ foram classificados como FVB (Tabela 3).

Tabela 3: Exemplo de como os volumes dos estratos foram classificados

Banda 3	Banda 4	Banda 5	NDVI	Volume
22	107	74	0,659	FV1
22	105	79	0,654	FV2
21	98	67	0,647	FV2
21	88	61	0,615	FV1
22	109	71	0,664	FV2
22	87	63	0,596	FVA

22	97	70	0,630	FVB
22	99	72	0,636	FVB
22	97	67	0,630	FVB
23	105	73	0,641	FVA

O banco de dados para o trabalho é composto pelos valores das bandas 3, 4 e 5 do satélite Landsat 5, pelo índice de vegetação NDVI e pela classificação intervalar do volume dos estratos.

K- vizinhos mais próximos (K-Nearest Neighbor)

Existem vários classificadores automáticos baseados em inteligência computacional. Neste trabalho se utilizou um método baseado em distância. A implementação do K-NN requer duas seleções principais que são: a distância métrica e o um valor para K. Sendo que K é o número de vizinhos mais próximos utilizados para classificar um exemplo novo em uma determinada classe.

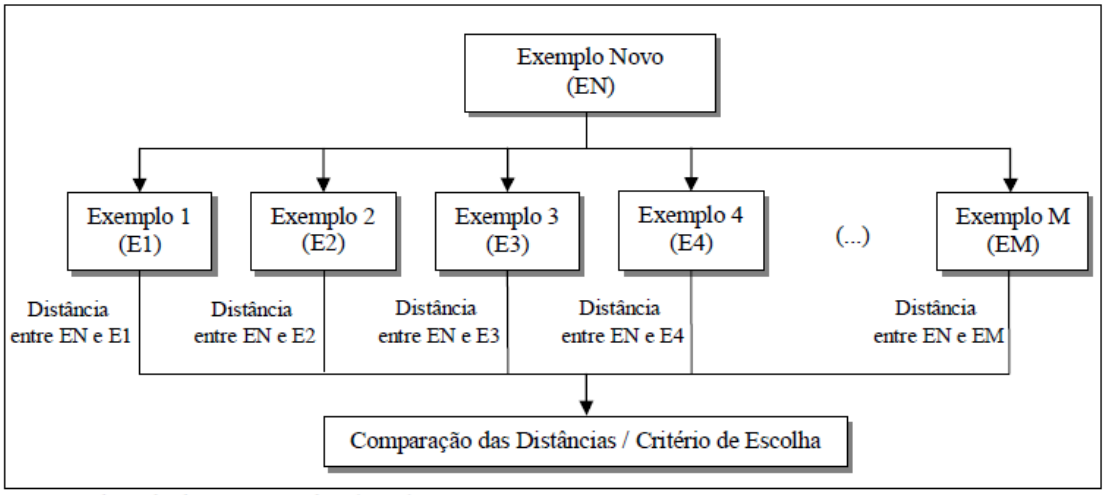
Para a aplicação da técnica K-NN utilizou-se uma amostra de 700 exemplos extraídos aleatoriamente da área de estudo, sendo 350 exemplos pertencentes ao Estrato 1 (Dbe) e 350 exemplos pertencentes ao Estrato 2 (Dbe + Abp). Cada exemplo possui quatro atributos preditivos, que são: número digital (digital number) da banda 3, da banda 4 e da banda 5, além do valor do NDVI. Para a incorporação do banco de dados de armazenamento do K-NN foram utilizados 560 exemplos (sendo 280 pertencentes ao Estrato 1 e 280 pertencentes ao Estrato 2) e para o teste da inteligência computacional foram utilizados os 140 exemplos, ou seja, 20% da massa total (sendo 70 exemplos do Estrado 1 e 70 exemplos do Estrato 2). As métricas utilizadas foram a distância Euclidiana e a distância de Manhattan. Considere os exemplos X1 = (x11, x12,..., x1n) e X2 = (x21, x22,...,x2n), ambos possuem *n* atributos e as distâncias Euclidiana (3) e de Manhattan (4) entre X1 e X2 é calculada pelas equações apresentadas abaixo:

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (3)$$

$$d(X_1, X_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}| \quad (4)$$

Para classificar um exemplo novo o método procura os K exemplos já armazenados que apresentam valores mais próximos dos valores do exemplo novo. Pois, dado certo padrão desconhecido X, sua classe é calculado pela distância entre X e os outros padrões de formação. E o padrão desconhecido X a partir das distâncias dos vizinhos mais próximos será classificado em um determinado grupo. A Figura 3 representa o esquema de como um exemplo novo é classificado pelo método K-NN.

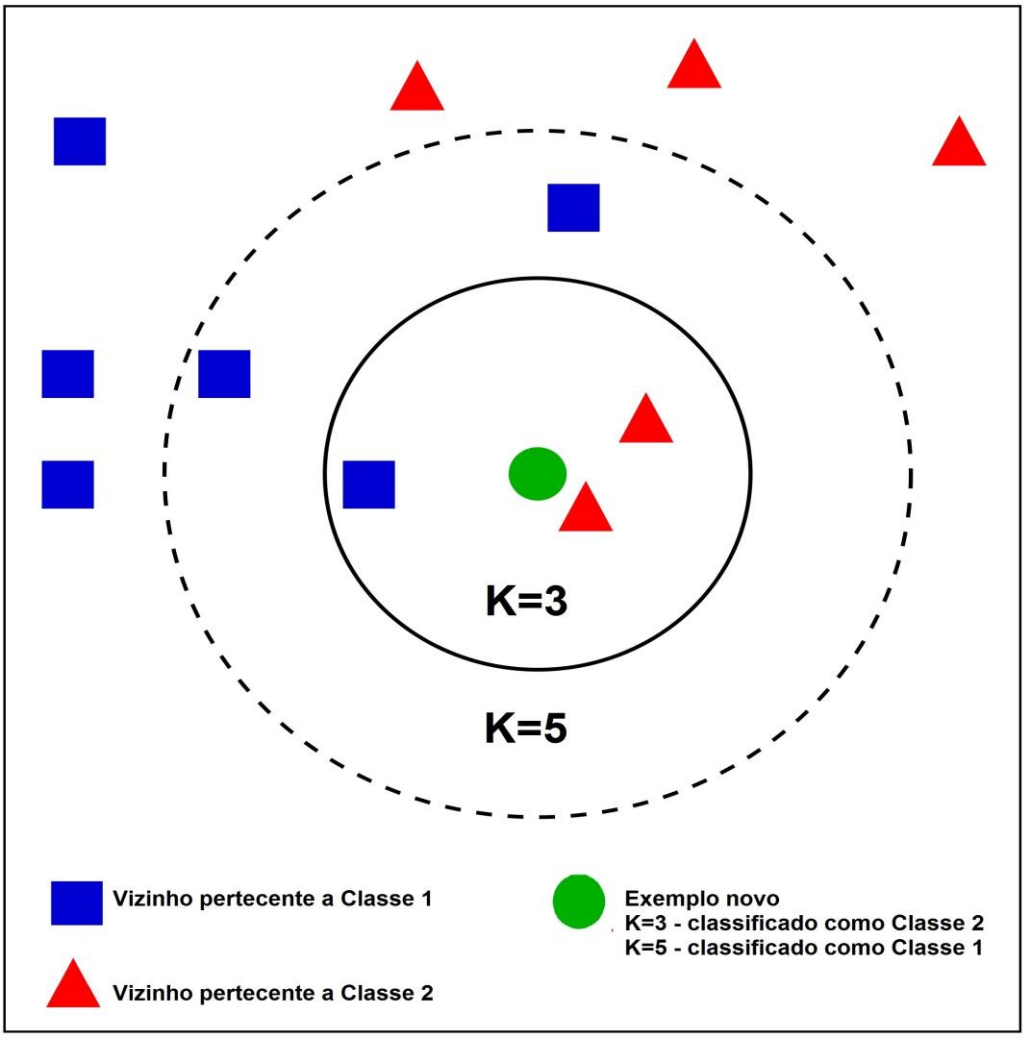
Figura 3: Esquema de classificação de um exemplo novo pelo método K-NN



Fonte: Adaptação de Han e Kamber, 2006.

Segundo Barros et al.(2011), o método do vizinho mais próximo pode ser prolongado, utilizando não um, mas um conjunto de dados mais próximos para prever o valor dos novos dados, em que é conhecido como os K-vizinhos mais próximos. Ao considerar mais do que um vizinho, é fornecido imunidade a ruídos e a curva de estimação é suavizada. Na figura 4 é exibida a forma como a classificação é executada conforme o número de vizinhos mais próximos (K).

Figura 4: Esquema de classificação conforme a escolha do número de vizinhos mais próximos (K)



Fonte: Elaborada pela Autora

O número de vizinhos mais próximos (K-NN) foi escolhido a partir de testes com o banco de dados e os valores de k testados foram 3, 5, 7, 9 e 11, todos ímpares, pois o K-NN compara distâncias entre o exemplo novo e os exemplos do banco de dados de armazenamento e segundo Ferrero (2009) números de vizinhos ímpares excluem empates na classificação no novo exemplo e k pares incluem empates.

A partir dos exemplos do banco de dados do K-NN e assimilação das distâncias do novo exemplo com os vizinhos mais próximos, a inteligência computacional irá classificar o novo exemplo nas seguintes classes FVA ou FVB (Estrato 1) ou FV1 ou FV2 (Estrato 2).

Com utilização do software ORANGE 2.7 (2015), o banco de dados de armazenamento foi utilizado para classificação automática dos exemplos de teste com utilização da ferramenta K-NN.

A partir da matriz de contingência (Quadro 1) foram calculados a acurácia global (CA), revocação (recall), medida-F(F1), precisão (precision) e a análise Característica de Operação do Receptor (Receiver Operating Characteristic-ROC).

Quadro 1: Esquema de uma Matriz de Contingência

		Real	
Predição	TP	FP	PP
	FN	TN	PN
	POS	NEG	N

Na tabela 4, TP corresponde a verdadeiros positivos, FP corresponde a falsos positivos, FN corresponde a falsos negativos e TN corresponde a verdadeiros negativos. PP e PN são referentes ao número de exemplos preditos respectivamente como positivos e negativos. POS e NEG ao número real de exemplos positivos e negativos na amostra. N é o número de elementos da amostra.

A acurácia global é um parâmetro de qualidade do modelo que se baseia nos exemplos classificados corretamente, ou seja, seu cálculo leva em consideração a diagonal descendente da matriz. A Acurácia Global (CA) a soma da diagonal principal da matriz de contingência pelo número total de amostras.

$$CA = \frac{|TP| + |TN|}{N} \quad (5)$$

Onde: TP = Verdadeiros Positivos
TN = Verdadeiros Negativos
N = Número total da amostra

As medidas de desempenho mais comumente usadas na classificação plana são as noções recuperação de informação clássica (RI) de precisão e revocação. Precisão para uma categoria Y denotado como Pr, mede a porcentagem de atribuições corretas entre todos os documentos atribuídos a Y. A revocação Re dá a percentagem de atribuições corretas entre todos os documentos que devem ser atribuídos a Y(SUN & LIM, 2001). Assim, foram calculados a revocação, precisão e medida- F de cada uma das classes do estudo (FVA; FVB; FV1;FV2) conforme o trabalho de Sun & Lim (2001).

$$Re = \frac{|TP|}{|TP| + |FN|} \quad (6)$$

Onde: TP = Verdadeiros Positivos
FN = Falsos Negativos

$$Pr = \frac{|TP|}{|TP| + |FP|} \quad (7)$$

Onde: TP = Verdadeiros Positivos
FP = Falsos Positivos

Segundo Faceli *et al.* (2011), a revocação e a precisão não são discutidas isoladamente, mas são combinadas em uma única medida,chamada de medida F, que é a média harmônica ponderada da precisão e a revocação. No presente trabalho a medida utilizada será F1 em que precisão e revocação tem igual valor de peso no cálculo.

$$F1 = 2.\frac{Pr.Re}{Pr+Re} \quad (8)$$

Onde: Pr = Precisão
Re = Revocação

E a análise ROC foi utilizada para investigar a sensibilidade e a especificidade dos dados testados no K-NN, ou seja, a relação entre verdadeiros positivos e falsos positivos. A sensibilidade é a porcentagem de exemplos positivos que foram classificados corretamente, já a especificidade é a porcentagem de exemplos negativos que foram classificados corretamente (NETO, RODRIGUES & MEIRA, 2014).

A sensibilidade e a especificidade foram calculadas a partir das fórmulas presente no trabalho de Fawcett (2004). Onde a razão de TP (9) é a sensibilidade que é a mesma fórmula de revocação e a razão de TN (10) é a especificidade. A taxa de falsos positivos equivale a (1 – especificidade).

$$TP \text{ ou } Sen. = \frac{|TP|}{POS} \quad (9)$$

Onde: TP = Verdadeiros positivos
POS = Total real de exemplos positivos

$$TN \text{ ou } Esp. = \frac{|TN|}{NEG} \quad (10)$$

Onde: TN = Verdadeiros negativos
NEG= Total real de exemplos negativos

Além da análise dos pontos (x,y) dos eixo x (taxa de verdadeiros positivos) e y (taxa de falsos positivos) referentes às matrizes de contingência dos estrato 1 e estrato 2. Foi realizada também a interpretação da área sob a curva (Area Under Curve - AUC). E a AUC é uma fração da área abaixo da curva do gráfico ROC, como cada lado do gráfico ROC varia de 0 a 1, a área do quadrado será 1 (um), ou seja, a área abaixo da curva só pode variar de 0 a 1 (PRATI, BATISTA & MONARD, 2008).

No quadro 2 está presente a relação entre AUC e o grau de desempenho do método de classificação automática.

Quadro 2: Relação entre valores de área sob a curva (AUC) e o grau de desempenho do modelo de classificação

Área sob a curva (AUC)	Desempenho
Acima de 0,9	Excelente
0,8 – 0,9	Bom
0,7 – 0,8	Regular
0,6 – 0,7	Ruim
0,5 – 0,6	Reprovado

Fonte: Câmara, 2009.

3.3- RESULTADOS E DISCUSSÃO

Número de K vizinhos mais próximos

O estrato 1(Dbe) possui 350 exemplos. O volume do estrato foi dividido em duas classes que são FV1 e FV2. Na classe FV2 estão inseridos exemplos (B3, B4, B5 e NDVI) com valores de volume de biomassa entre 20m³ e 40 m³ e FV2 com valores entre 40m³ e 60m³. Para os testes de seleção do número de vizinhos mais próximos (K) foi utilizado 20% do total de exemplos referente ao estrato 1, ou seja, 70 exemplos (35 da classe FV1 e 35 da classe FV2). Para os testes foram utilizadas as duas métricas de distância e cada uma com K igual a 3, 5, 7, 9 e 11.

Tabela 4: Valores de, K, AUC, F1, PRECISÃO e REVOCAÇÃO referente aos testes com a distância Euclidiana no estrato 1 (Dbe)

	Distância Euclidiana				
K	3	5	7	9	11
AUC	0,871	0,857	0,857	0,843	0,829
CA	0,871	0,857	0,857	0,843	0,829
F1	0,87	0,853	0,853	0,836	0,824
PRECISÃO	0,882	0,879	0,879	0,875	0,848
REVOCAÇÃO	0,857	0,829	0,829	0,8	0,8

Com os testes realizados com a utilização da distância Euclidiana (Tabela 4) foi possível verificar que em todos os parâmetros de análise o teste com K= 3 possui valores maiores em relação aos testes com K=5, K=7 K=9 e K=11.

Tabela 5: Valores de, K, AUC, F1, PRECISÃO e REVOCAÇÃO referente aos testes com a distância de Manhattan no estrato 1 (Dbe)

Distância Manhattan					
K	3	5	7	9	11
AUC	0,871	0,843	0,829	0,843	0,843
CA	0,871	0,843	0,829	0,843	0,843
F1	0,866	0,841	0,824	0,836	0,836
PRECISÃO	0,906	0,853	0,848	0,875	0,875
REVOCAÇÃO	0,829	0,829	0,8	0,8	0,8

Assim como nos testes com a distância euclidiana, o teste com a distância de Manhattan (Tabela 5) que possui melhores resultados é K=3. Comparando K=3 pelas equações de distâncias utilizadas, o teste com a distância euclidiana possui AUC e CA iguais e F1 e revocação maiores que o teste com a distância de Manhattan.

No estrato 2 (Dbe + Abp), o procedimento foi o mesmo realizados com o estrato 1(Dbe). Sendo que as faixas agora possuem outra denominação que são FVA e FVB.

Tabela 6: Valores de, K, AUC, F1, PRECISÃO e REVOCAÇÃO referente aos testes com a distância de Euclidiana no estrato 2 (Dbe + Abp)

Distância Euclidiana					
K	3	5	7	9	11
AUC	0,743	0,829	0,814	0,786	0,786
CA	0,743	0,82	0,814	0,786	0,786
F1	0,757	0,842	0,831	0,805	0,8
PRECISÃO	0,718	0,78	0,762	0,738	0,75
REVOCAÇÃO	0,8	0,914	0,914	0,886	0,857

Com os testes realizados com a utilização da distância Euclideana (Tabela 6) foi possível verificar que em todos os parâmetros de análise o teste com K= 5 possui valores maiores em relação aos testes com K=3, K=7 K=9 e K=11.

Tabela 7: Valores de, K, AUC, F1, PRECISÃO e REVOCAÇÃO referente aos testes com a distância de Manhattan no estrato 2 (Dbe + Abp)

Distância Manhattan					
K	3	5	7	9	11
AUC	0,771	0,843	0,8	0,8	0,786
CA	0,771	0,843	0,8	0,8	0,786
F1	0,784	0,857	0,821	0,821	0,8
PRECISÃO	0,744	0,786	0,744	0,744	0,75
REVOCAÇÃO	0,829	0,943	0,914	0,914	0,857

Assim como nos testes com a distância euclidiana, o teste com a distância de Manhattan (Tabela 7) que possui melhores resultados é K=5. Comparando K= 5 pelas equações de distâncias utilizadas, o teste com a distância de Manhattan os valores de AUC, CA, F1, precisão e revocação maiores que o teste com a distância Euclidiana.

Nos testes realizados com FV1 e FV2 tanto o teste com K= 3 e distância Euclidiana quanto o teste com K=3 e distância de Manhattan apresentaram acurácia global (CA) e área sob a curva (AUC) com valores iguais, o que diferia um teste do outro eram os valores de precisão, revocação e medida-F (F1), enquanto no primeiro caso a revocação e F1 era maiores, no segundo caso a precisão era maior.

Como F1 era maior no teste com a distância Euclideana, este foi selecionado como o melhor modelo para classificar FV1 e FV2. No trabalho de Lichtnow et al. (2010) é exemplificado como a medida – F (F1) é uma medição que auxilia na escolha de um melhor experimento por fazer um balanceamento entre precisão e revocação, em seu artigo a precisão aumentou de um experimento para o seguinte, mas a revocação diminuiu, mas como o aumento da precisão foi mais significativo que a diminuição da revocação, o segundo experimento conseguiu um valor de F1 maior, e isso indica uma melhoria nos resultados.

Com os testes referente as classes FVA e FVB, apesar dos melhores resultados serem aqueles com K=5 com ambas as distâncias, todas as medições para avaliar a qualidade do modelo de classificação foram maiores com a distância de Manhattan. E segundo Zhang et al. (2015), as medidas de precisão e de revocação são usadas em conjunto para avaliar a qualidade de classificação. Ao comparar os dois resultados de classificação, se uma classificação apresenta

valores mais elevados em relação a ambas as medidas do que o outro, em seguida, a sua qualidade é mais elevada do que a do outro.

Matriz de Contingência

A partir dos testes foi elaborada a matriz de contingência do melhor resultado. Na Tabela 8, apresenta a matriz do estrato 1 (Dbe) com número de vizinhos mais próximos igual a três (K=3) e com distância entre o exemplo novo e o banco de dados de armazenamento calculada a partir da distância Euclideana.

Tabela 8: Matriz de contingência do estrato 1 (Dbe) com as classes de faixas de volume FV1 e FV2

		Real		
Predição		FV1	FV2	∑
	FV1	31	5	36
	FV2	4	30	34
	∑	35	35	70

A acurácia global do estrato 1 obtida a partir da matriz de contingência foi de aproximadamente 87%.Na Tabela 9, apresenta-se a matriz de contingência do estrato 2 com número de vizinhos mais próximos igual a cinco (K=5) e com distância entre o exemplo novo e o banco de dados de armazenamento calculada a partir da distância de Manhattan. E a acurácia global do estrato 2 obtida a partir da matriz de contingência foi de 84%.

Tabela 9: Matriz de contingência do estrato 2 (Dbe+ Abp) com as classes das faixas de volume FVA e FVB

		Real		
Predição		FVA	FVB	∑
	FVA	26	2	28
	FVB	9	33	42
	∑	35	35	70

Análise ROC

O gráfico ROC é baseado na probabilidade de detecção ou taxa de verdadeiros positivos, e na probabilidade de falsos alarmes ou taxa de falsos positivos Para construir o gráfico ROC deve ser plotado a taxa de falsos positivos no eixo dos ordenadas – eixo x – e a taxa de verdadeiros positivos no eixo das

abscissas – eixo y. Um modelo de classificação é representado por um ponto no espaço ROC e para se obter o ponto no espaço ROC correspondente a um modelo de classificação, calcula-se a taxa de verdadeiros e falsos positivos (PRATI, BATISTA & MONARD, 2008)

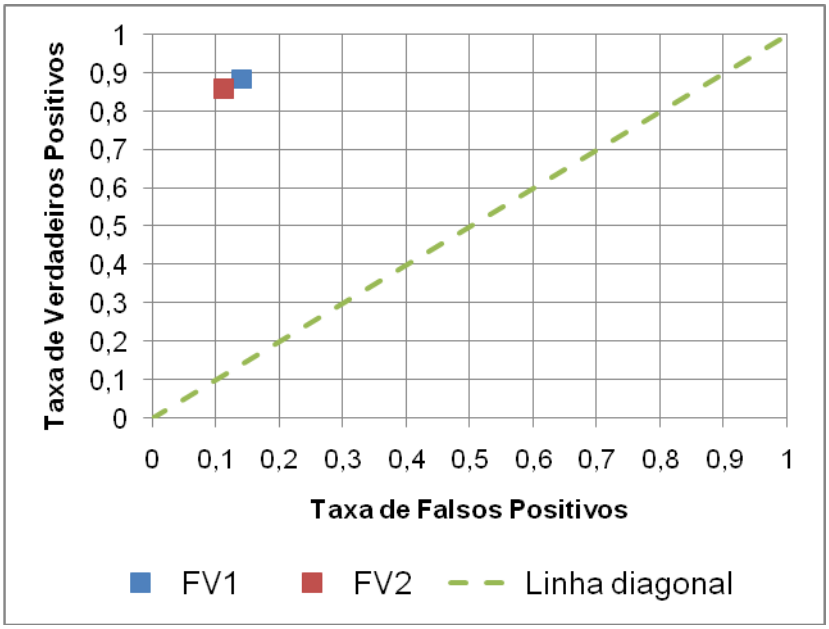
Os valores dos pontos (x, y) do gráfico ROC da faixa de volume FV1 e FV2 estão na Tabela 10. E a área sob a curva (AUC) das faixas de volume FV1 e FV2 foi de 0,871.

Tabela 10: Valores dos pontos cartesianos do gráfico ROC do estrato 1 nas classes de faixas de volume FV1 e FV2

	Gráfico ROC	
	X (taxa de FP)	Y (taxa de TP)
FV1	0,14285714	0,8857143
FV2	0,11428571	0,8571429

A partir dos pontos cartesianos foi elaborada a Figura 5 que corresponde ao modelo de classificação das classes FV1 e FV2.

Figura 5: Gráfico ROC das faixas de biomassa do Estrato 1



Alguns pontos merecem destaque no gráfico que são os pontos (0,0), (0,1), (1,1) e (1,0). Modelos que correspondem ao ponto (0,0) não apresentam nenhum falso positivo, mas também não conseguem classificar nenhum verdadeiro positivo. A estratégia inversa, de sempre classificar um novo exemplo como positivo, é

representada pelo ponto (1,1). O ponto (0,1) representa o modelo perfeito, em que todos os exemplos positivos e negativos são corretamente classificados e o ponto (1,0) representa o modelo que sempre faz predições erradas (PRATI, BATISTA & MONARD, 2008). E tanto o ponto da classe FV1 quando o ponto da classe FV2 se encontram nos intervalos entre 0,1 e 0,2 no eixo x e entre 0,8 e 0,9 no eixo y, ou seja, pontos próximos ao ponto (0,1) que são considerados modelos de classificação próximos ao modelo padrão.

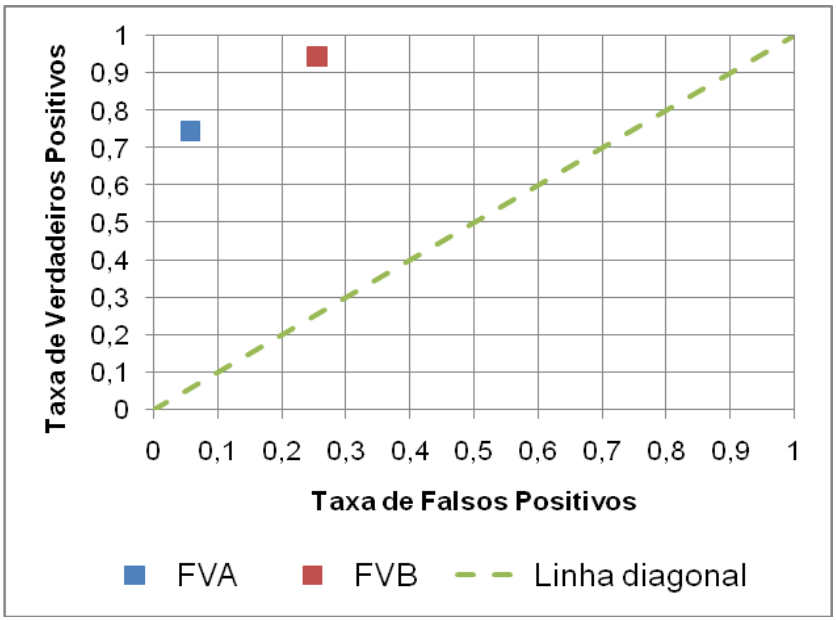
Os valores dos pontos (x, y) do gráfico ROC da faixa de volume FVA e FVB estão na Tabela 11. E a área sob a curva (AUC) das faixas de volume FVA e FVB foi 0,843.

Tabela 11: Valores dos pontos cartesianos do estrato 2 das classes nas faixas de volume FVA e FVB

	Gráfico ROC	
	X (taxa de FP)	Y (taxa de TP)
FVA	0,05714286	0,7428571
FVB	0,25714286	0,9428571

A partir dos pontos cartesianos foi elaborada a Figura 6 que corresponde ao modelo de classificação das classes FVA e FVB.

Figura 6: Gráfico ROC das classes de faixas de biomassa do estrato 2



O ponto da classe FVA encontra-se nos intervalos 0 e 0,1 no eixo x e 0,7 e 0,8 no eixo y, já o ponto da classe FVB se encontra nos intervalos entre 0,2 e 0,3 no

eixo x e entre 0,9 e 0,1 no eixo y. O modelo de classificação da classe FVA possui seu ponto mais próximos do eixo y comparado com o modelo da classe FVB que além de estar mais distante com eixo y, também possui um valor mais no eixo x. E segundo Prati, Batista & Monard (2003) um ponto em um gráfico ROC domina outro se esse ponto estiver acima e a esquerda do outro, isto é, tem uma taxa mais alta de verdadeiros positivos e uma taxa mais baixa de falsos positivos.

O modelo de classificação de FVA possui desempenho bom por causa da AUC e é mais conservativo que o modelo de FVB pelo seu ponto no gráfico ROC, pois comparando FVA e FVB o modelo FVA é mais próximo ao canto inferior esquerdo podem ser considerado conservativo, pois faz uma classificação positiva somente se tem grande segurança na classificação. Como conseqüência, o modelo comete poucos erros falsos positivos, mas freqüentemente tem taxa menor de verdadeiros positivos comparado ao modelo FVB (PRATI, BATISTA & MONARD, 2008).

A área sob a curva (AUC) dos dois modelos de classificação está acima de 0,8 e pelos intervalos de desempenho de modelos proposto por Câmara (2009) o desempenho de ambos é classificado como bom. A AUC é um importante e amplamente utilizado índice de curva ROC na análise de diagnóstico ou classificação binária. Ele resume a curva ROC em termos de sensibilidade e especificidade e muitas vezes é tratado como uma medida para avaliar os testes de diagnóstico ou classificadores. É importante ressaltar que pesquisadores também fazem uso de AUC como um objetivo de encontrar bons classificadores (YU, CHANG & PARK, 2014).

3.4- CONCLUSÃO

O objetivo de classificar as tipologias florestais (Dbe; Dbe + Abp) em faixas de biomassa foi possível com a metodologia proposta. Para as medidas de precisão utilizadas, o modelo é satisfatório para uso na área de estudo para monitoramento e gestão via sensoriamento remoto.

No estrato 1 (Dbe), o número de vizinhos mais próximos (K) foi igual a três com utilização da métrica euclidiana, acurácia global de aproximadamente 87% e AUC igual a 0,871. No estrato 2, o número de vizinhos mais próximo (K) foi igual a cinco, acurácia global de 84% e AUC igual a 0,843.

. Todos os parâmetros utilizados para analisar o modelo de classificação tiveram resultados de desempenho bom em relação a AUC e acurácia razoável, pois é sempre interessante chegar o mais próximo de 100%.

Essas faixas podem estipular ganhos e perdas de biomassa, mas de maneira menos precisa, sendo necessários estudos que possibilitem estimativas mais exatas.

3.5 - REFERÊNCIAS

ALVES, M. V. G., CHIAVETTA, U., KOEHLER, H. S.,MACHADO, S. A.; KIRCHNER, F. F. Aplicação de k-nearest neighbor em imagens multispectrais para a estimativa de parâmetros florestais. **Floresta**, V. 43, N. 3, P. 351-362, 2013.

AMARAL, B. F.; GONÇALVES, R. R. V.; ROMANI, L. A. S.; SOUSA, E. P. M. Aprimorando a classificação semissupervisionada de séries temporais extraídas de imagens de satélite. **Symposium *on knowledge* discovery, mining and learning**.2012.

BARROS, J. D.; OLIVEIRA JÚNIOR, J. J. D.; SILVA, S. G. D.; FARIAS, R. F. D. Characterization of bone Tissue by microwaves using wavelets and KNN. **Journal of Microwaves, Optoelectronics and Electromagnetic Applications**, v. 10, n. 1, p. 217-231, 2011.

CÂMARA, F.P. Psiquiatria E Estatística V: Validação De Procedimentos Diagnóstica Pela Curva R.O.C. **Psychiatry on line Brasil**. Vol.14. Nº 4. 2009.

CERDA, J.; CIFUENTES, L. Uso de curvas roc en investigación clínica: aspectos teórico-prácticos. **Revista Chilena de Infectología**, V. 29, N. 2, P. 138-141, 2012.

FACELI, K.; LORENA, A.C.; GAMA, J.;CARVALHO, A. **Inteligência artificial**: uma abordagem de aprendizado de máquina. GRUPO GEN-LTC, 2011.

FAWCETT, T. ROC graphs: notes and practical considerations for researchers. **Machine learning**, V. 31, N. 1, P. 1-38, 2004.

FERRERO, Carlos Andres. Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia. 2009. Tese de Doutorado. Universidade de São Paulo.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithms for computing k-nearest neighbors. **IEEE transactions on computers**, New York, V. 24, N. 7, P. 750-753, 1975.

GAROFALO, D. F. T., MESSIAS, C. G., LIESENBERG, V., BOLFE, É. L., FERREIRA, M. C. Comparative analysis of digital classifiers of landsat-8 images for thematic mapping procedures. **Pesquisa Agropecuária Brasileira**, V. 50, N. 7, P. 593-604, 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Manual técnico da vegetação brasileira**. Manuais Técnicos em Geociências, Vol. 1. 2ª Edição. 2012.

Instituto de Desenvolvimento Florestal do Pará – Ideflor- Bio. **Inventário Florestal Diagnóstico Do Conjunto De Glebas Estaduais Mamuru-Arapiuns – Pará**: Relatório Final. 2010.

LICHTNOW, D., DE OLIVEIRA, J. P. M., DE LIMA, J. V., PROENÇA JR, M. L., DE BARROS, R. M., & ADANIYA, M. H. A. Técnicas de extração de informação para avaliação da qualidade de páginas web com o uso de ontologias. **Cadernos de Informática**, *5*(1), 77-88. 2010.

NETO, C. G.; RODRIGUES, L. H. A.; MEIRA, C. A. A. Modelos de predição da ferrugem do cafeeiro (Hemileia vastatrix Berkeley & Broome) por técnicas de mineração de dados. Coffee Science, v. 9, n. 3, p. 408-418, 2014.

PONZONI, F. J.; SHIMABUKURO, Y. E.; KUPLICH, T. M. **Sensoriamento remoto da vegetação**. Oficina de Textos, 2012.

PRATI, R. C., BATISTA, G. E. A. P. A., MONARD, M. C. Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise ROC. **Proc. Of the workshop on advances & trends in ai for problem solving**. 2003. P. 28-33.

PRATI, R. C.; BATISTA, GEAPA; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, V. 6, N. 2, P. 215-222, 2008.

REZENDE, SOLANGE O.; MARCACINI, RICARDO M.; MOURA, MARIA F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, V. 7, P. 7-21, 2011.

RODRIGUES, C. A. G.; HOTT, M. C. Dinâmica da vegetação natural no nordeste do estado de São Paulo, entre 1988 e 2003. **Revista Árvore**, V. 34, N. 5, P. 881-887, 2010.

ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. **NASA special publication**, v. 351, p. 309, 1974.

SANTOS, A. A.; MACHADO, M. M. M. Análise da fragmentação da paisagem do Parque Nacional da Serra da Canastra e de sua Zona de Amortecimento-MG. **Raega - O Espaço Geográfico em Análise**, V. 33, P. 75-93, 2015.

SARMIENTO, C. M., RAMIREZ, G. M., COLTRI, P. P., LIMA, L. F., NASSUR, O. A. C., & SOARES, J. F. Comparação de Classificadores Supervisionados na Discriminação de Áreas Cafeeiras em Campos Gerais-Minas Gerais. **Coffee Science**, V. 9, N. 4, P. 546-557, 2014.

SILVA, J. N. M.; Carvalho J.O.P.; Lopes, J.C.A.;Carvalho, M.S.P. Equações de volume para a Floresta Nacional do Tapajós. **Boletim de Pesquisa Florestal**, v. 8, p. 50-63. 1984.

SUN, AIXIN; LIM, EE-PENG. Hierarchical text classification and evaluation. **Proceedings IEEE international conference** on. IEEE, 2001. P. 521-528.

URNAU, EDUARDO; KIPPER, LIANE MAHLMANNN; FROZZA, REJANE. Desenvolvimento de um sistema de apoio à decisão com a técnica de raciocínio baseado em casos. **Perspectivas em Ciência da Informação**, V. 19, N. 4, P. 118-135, 2014.

XU, Y. ET AL. Coarse to fine k nearest neighbor classifier**. Pattern recognition letters**, Northholland, V. 34, P. 980-986, FEB. 2013.

YU, W.; CHANG, YUAN-CHIN, I.; PARK, E. A Modified Area Under the ROC Curve And its Application to Marker Selection and Classification. **Journal of the Korean Statistical Society**, V. 43, N. 2, P. 161-175, 2014.

ZHANG, X., FENG, X., XIAO, P., HE, G., & ZHU, L. Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. **ISPRS Journal Of Photogrammetry And Remote Sensing**, V. 102, P. 73-84, 2015.

4- CONCLUSÃO GERAL

Com o presente trabalho foi possível atingir os objetivos dos artigos da dissertação e os modelos são viáveis para utilização. No primeiro artigo, o objetivo de classificar as tipologias florestais (Dbe; Dbe+Abp) com utilização de sensoriamento remoto e inteligência computacional do tipo K-NN embasados em dados de inventário florestal da foi atingido. Os resultados da acurácia global de 87%, índice Kappa de 74% e AUC igual a 0,929 garantiram uma avaliação positiva do modelo de classificação automática que classifica em estrato 1 e 2.

No artigo dois, o objetivo de classificação por faixas de biomassa também foi atingido com resultados. Para o estrato 1 (Dbe), o número de vizinhos mais próximos (K) foi igual a três com utilização da métrica euclidiana, acurácia global de aproximadamente 87% e AUC igual a 0,871, para o estrato 2, o número de vizinhos mais próximo (K) foi igual a cinco, acurácia global de 84% e AUC igual a 0,843.

É importante ressaltar a importância de progredir em estudos de classificação automática na área de estudo, e também de futuros trabalhos com estimativas de biomassa florestal nas Glebas Mamuru- Arapiuns com uso de tecnologias dos tipos inteligência computacional e geoprocessamento.

Universidade do Estado do Pará
Centro de Ciências Naturais e Tecnologia
Programa de Pós-Graduação em Ciências Ambientais – Mestrado
Tv. Enéas Pinheiro, 2626, Marco, Belém-PA, CEP: 66095-100
www.uepa.br/paginas/pcambientais