Universidade Federal do Paraná

Razer Anthom Nizer Rojas Montaño

Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal

Curitiba PR
2016

Razer Anthom Nizer Rojas Montaño

Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Informática no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Eduardo Todt.

Co-orientador: Prof. Dr. Carlos Sanquetta.

Curitiba PR
2016

Montaño, Razer Anthon Nizer Rojas
M765 Aplicações de técnicas de aprendizado de máquina na mensuração florestal / Razer Anthon Nizer Rojas Montaño. – Curitiba, 2016.
102 f.: il., tabs, grafs.

Orientador: Eduardo Todt
Co-orientador: Carlos Sanquetta
Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

1. Inventário florestal. 2. Redes neurais. 3. Florestas. I. Todt, Eduardo. II. Sanquetta, Carlos Roberto. III.Título. IV. Universidade Federal do Paraná.

CDD 006.3

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS EXATAS
Programa de Pós Graduação em INFORMÁTICA
Código CAPES: 40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de **RAZER ANTHOM NIZER ROJAS MONTAÑO**, intitulada: "**Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal**", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua aprovação.

Curitiba, 25 de Novembro de 2016.

EDUARDO TODT
Presidente da Banca Examinadora (UFPR)

ANA PAULA DALLA CORTE
Avaliador Externo (UFPR)

CARLOS ROBERTO SANQUETTA
Avaliador Externo (UFPR)

JAIME WOJCIECHOWSKI
Avaliador Externo (UFPR)

JULIO CÉSAR NIEVOLA
Avaliador Externo (PUC/PR)

LUCAS FERRARI DE OLIVEIRA
Avaliador Interno (UFPR)

Aos meus filhos Guilherme, Brunna e Juan Pablo.

Agradecimentos

À Deus, pela oportunidade deste desafio e por fornecer as forças necessárias para vencê-lo.

À minha mãe, Nádia Nizer, que nunca poupou esforços para nos dar do melhor e nos direcionar para uma vida reta e produtiva.

À minha avó, Camila Nizer (em memória), que sonhava em ver os netos doutores e onde estiver, encarnada ou desencarnada, pode saborear este momento.

Aos meus filhos Guilherme, Brunna e Juan Pablo, que são a fonte de energia e o porto revigorante para continuar vencendo os desafios.

Ao meu orientador Eduarto Todt e ao meu co-orientador Carlos Sanquetta, sem os quais esta tese não seria possível. Fontes de conhecimento, direcionamento e amizade.

Ao meu amigo e irmão Jaime Wojciechowski, que foi o grande amparo em momentos difíceis, pela amizade e companheirismo.

Ao meu amigo e irmão Everton Pascke, pelo companheirismo, amizade e suporte em diversos momentos.

Aos meus amigos de longa data, Eduardo Sant'ana, Bruno Ribas, Willian Zalewski, Ricardo Oliveira, Adil Calomeno, Daniel Wandarti, Rodolfo Rodovalho e Clariane Menezes, que estiveram comigo por muitos anos, acompanharam todo o trajeto, por todo o apoio e suporte nas horas em que mais precisei.

Aos inúmeros amigos, colegas de trabalho, alunos, pois é o relacionamento interpessoal que faz com que cresçamos como pessoa.

Aos pesquisadores citados neste trabalho por terem criado condições para o desenvolvimento deste projeto.

Ao DINF por fornecer toda a infraestrutura necessária para o desenvolvimento deste projeto.

À Resistência Curitiba, por permitir momentos de descontração e amizade.

Resumo

Vive-se em um mundo onde a escassez de recursos naturais leva a um uso cada vez mais racional destes, seja água, recursos minerais, biológicos, hídricos e energéticos. Especificamente quando se trata de madeira, o homem evoluiu muito no manejo de florestas plantadas, aplicando técnicas de cultivo e planejamento de corte para a exploração. Estas técnicas fazem uso da mensuração florestal para medição e estimação de valores dendrométricos importantes, como altura, volume e biomassa. A estimativa de valores dendrométricos é de extrema importância, pois não é viável o abate de toda uma população para que sejam observados os exatos valores. Assim, pesquisadores fazem uso de ferramentas estatísticas em mensuração florestal há anos, com bastante sucesso em suas estimações. Recentemente, com o avanço da área da inteligência artificial, técnicas de aprendizado de máquina têm se mostrado também capazes de competir com os métodos estatísticos de regressão, abrindo assim um leque de opções aos pesquisadores. Inserido neste contexto, o objetivo deste trabalho é aplicar técnicas de aprendizado de máquina para resolução de problemas de mensuração florestal, mostrando também que este se encaixa como um processo de descoberta de conhecimento, inserido na área da Ciência da Computação. Foram realizados experimentos com dados de árvores de Acácia-negra para avaliação de biomassa e relação hipsométrica, Pinus para estimativa de volumes e relação hipsométrica, e com uma amostra de árvores de florestas tropicais de várias regiões do mundo, para estimativa de biomassa. Compararam-se modelos alométricos clássicos com Redes Neurais Artificiais (RNA), Máquinas de Vetores de Suporte (SVM) e *Random Forest* (RF), e em todos os testes realizados o modelo que obteve melhor correlação é uma técnica de aprendizado de máquina. Para a predição de volumes de Pinus, o melhor modelo foi SVM com correlação de 99,19%. Para estimativa de biomassa da Acácia-negra, SVM obteve a melhor correlação com 98,6%. Para estimativa de biomassa de florestas tropicais, o melhor modelo foi RNA com correlação de 98,06%. Para relação hipsométrica da Acácia-negra, o modelo de SVM obteve a melhor correlação, de 97,73%. Já para relação hipsométrica de Pinus, o modelo de melhor predição foi RNA com correlação de 98,02%. O teste de Friedman mostrou a presença de diferença estatística entre os métodos e, embora o pós-teste de Nemenyi não tenha conseguido evidenciá-la, foi detectada uma tendência a uma separação entre os métodos. Os resultados obtidos mostram fortemente que os modelos de aprendizado de máquina são uma alternativa competitiva frente aos métodos clássicos, superando-os nos experimentos aqui realizados.

Palavras chave: Inventário Florestal, Mensuração Florestal, Descoberta de Conhecimento em Bases de Dados, Aprendizado de Máquina, Redes Neurais, Máquinas de Vetores de Suporte, Árvores de Modelos, Florestas Aleatórias.

Abstract

We live in a world where scarcity of natural resources leads to an increasing rational use of these resources, like water, minerals, biological or energetical. Specifically wood, man evolved much in the management of planted forests by applying cultivation techniques and cutting planning for exploration. These techniques make use of forest measurement to estimation of important dendrometric values such as height, volume and biomass. The estimation of dendrometric values is extremely important because is not feasible to cut down the whole population to note the exact values. Thus, researchers make use of statistical tools for measuring forest for years with success in their estimates. Recently, with the advancement of artificial intelligence, machine learning techniques have been able to compete with the statistical regression methods, thereby opening a range of options to researchers. Within this context, the objective of this work is to use machine learning techniques to resolution of forest measurement problems, showing that this is a process of knowledge discovery, from the area of Computer Science. Experiments were performed with data of Acacia-negra trees for evaluation of biomass and hypsometric relation, Pinus to estimate volumes and hypsometric relation, and with a sample of tropical forest trees of various regions to estimate biomass. Classical allometric models were compared with Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forest (RF), and in all tests the model that obtained the best correlation was a machine learning technique. For the prediction of Pinus volumes, the best model was SVM with correlation of 99.19 %. To estimate Acacia-negra biomass, SVM showed the best correlation with 98.6 %. To estimate biomass of tropical forest trees, the best model was RNA with a correlation of 98.06 %. For hypsometric relation of Acacia-negra, the SVM model obtained the best correlation, of 97.73 %. As for the hypsometric relation of Pinus, the best prediction model was RNA with a correlation of 98.02The Friedman test showed the presence of statistical difference between the methods and, although the Nemenyi post-hoc test was not able to show it, it was detectecd a tendency towards a separation of methods. The results obtained strongly show that machine learning models are a competitive alternative in comparison to the classical methods, surpassing them in the experiments carried out here.

Keywords: Forest Inventory, Forest Measuring, Knowledge Discovery in Databases Machine Learning, Neural Networks, Support Vector Machines, Model Trees, Random Forest.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AM	:	Aprendizado de Máquina
DAP	:	Diâmetro à Altura do Peito
IA	:	Inteligência Artificial
KDD	:	*Knowledge Discovery in Database*, descoberta de conhecimento em bases de dados
MD	:	Mineração de Dados
MLP	:	*Multilayer Perceptron*, rede neural do tipo Perceptron com várias camadas
RF	:	*Random Forest*, floresta aleatória
RNA	:	Redes Neurais Artificiais
SMO	:	*Sequential Minimal Optimization*, algoritmo de treinamento de SVMs
SVM	:	*Support Vector Machines*, máquinas de vetores de suporte

Lista de Símbolos

β_0	:	primeiro coeficiente de uma equação de regressão
β_1	:	segundo coeficiente de uma equação de regressão
β_2	:	terceiro coeficiente de uma equação de regressão
ln	:	logarítmo neperiano

Capítulo 1

Introdução

Quando se analisa o cenário florestal e de recursos naturais, até meados dos anos 60 esses recursos eram usados de forma intensiva e sem a devida atenção aos aspectos ecológicos ou de sustentabilidade [May et al., 2003]. Tinham-se disponíveis florestas em abundância e um mercado sem rigor em relação a medidas de quantidades comercializadas. Da mesma forma, não havia preocupações sobre a sustentabilidade dos plantios e/ou florestas nativas. À medida que a escassez de madeira começou a crescer, os preços se elevaram e a relação entre o vendedor e o consumidor passou a exigir maior rigor no dimensionamento dos produtos.

Com a escassez desses recursos e redução da biodiversidade, o melhoramento dos procedimentos de quantificação e avaliação dos produtos florestais se tornou chave para o sucesso dos negócios e manutenção sustentável de florestas. Assim, as florestas passaram a ser consideradas capazes de oferecer rendimentos permanentes, desde que sua produção não comprometesse o equilíbrio entre o estoque em crescimento e o volume comercialmente aproveitável [Scolforo, 1998; Scolforo, 2011].

O conjunto de técnicas adotadas para a administração de áreas florestais, de forma a manter este equilíbrio entre a obtenção de recursos e a sustentação do ecossistema é conhecido como Manejo Florestal. Dentro do manejo, uma das etapas envolve o inventário florestal, que é o procedimento realizado para se obter informações qualitativas e quantitativas de uma área florestal.

Neste contexto, as quantificações, que são um dos pilares do inventário florestal, são baseadas na medição de variáveis dendrométricas de árvores e populações. Estas variáveis são medidas por meio de instrumentos ou abate de árvores, sendo que algumas medições são de fácil obtenção e outras de difícil obtenção, sendo necessário, em alguns casos, o abate da árvore.

Isso ocorre, por exemplo, com a relação hipsométrica, que determina a relação entre o diâmetro à altura do peito (dap) de uma árvore e da sua altura total (h_t). É fácil medir o dap, mas a altura total demanda tempo e instrumentos de precisão, ainda estando sujeita a erros, caso não se possa abater a árvore. Assim, a partir de uma amostra de medição de dap e h_t, consegue-se criar uma relação em que, dado o dap de outra árvore da população, seja feita uma estimativa de sua h_t.

Economicamente, não é viável fazer medições de todas as variáveis dendrométricas das árvores de uma população e, portanto, admite-se que estimativas possam ser feitas e assume-se determinada taxa de erro. O método mais empregado para estimativa de variáveis dendrométricas é a regressão, por meio de pesquisas de modelos nas várias áreas da mensuração florestal.

A regressão, portanto, é a ferramenta padrão de ajuste de modelos para as várias tarefas da mensuração florestal. Entretanto, devido à sua baixa flexibilidade e rigidez, especialistas na área buscam alternativas que possam aproveitar o poder computacional dos dias de hoje, bem

como fazer uso das várias informações disponíveis em suas bases de dados, o que não é possível com a maioria dos modelos baseados em regressão.

Recentemente, estudos na área de Inteligência Artificial (IA), mostraram que técnicas de Aprendizado de Máquina (AM, ou do inglês ML - *Machine Learning*) se adequam ao tipo de predição feito na mensuração florestal. Alguns trabalhos [Diamantopoulou, 2005; Castellanos et al., 2007; da Silva et al., 2009; da Silva Binoti, 2012; Binoti et al., 2014a] mostram que as predições obtidas através do uso destes procedimentos podem ser melhores que as obtidas com a clássica regressão.

Assim, tendo em vista que a mensuração florestal usa uma base forte em estatística e matemática para o ajuste de modelos [Sanquetta et al., 2009], que as técnicas de aprendizado de máquina já vêm sendo utilizadas nesta área, em especial as RNAs [da Silva Binoti, 2012; Binoti et al., 2014a], que algoritmos supervisionados para regressão apresentam uma grande flexibilidade de parametrização [Faceli et al., 2011; Rezende, 2003] e que modelos baseados em árvores, como *Random Forest*, conseguem modelar características particulares de porções dos dados em um mesmo modelo, aspecto que mesmo a regressão não-linear tem dificuldade, muitas vezes aproximando bem certas predições e outras não [Breiman, 2001], este trabalho apresenta as seguintes hipóteses:

- As técnicas de AM podem substituir os modelos comumente usados na área de Mensuração Florestal;
- As técnicas de AM geram modelos mais precisos para predição de volumes e a avaliação de biomassa.

Esta pesquisa tem cunho multidisciplinar, na área da Ciência da Computação usando como estudo de caso problemas específicos da área de Engenharia Florestal. Como contribuições, podem-se destacar:

- Possibilidade de definição de uma nova abordagem em mensuração florestal baseada em aprendizado de máquina;
- Comparação dessas técnicas com regressão para abrir um rol de possibilidades aos engenheiros, bem como a possibilidade de dar mais flexibilidade e robustez em suas estimativas;
- A aplicação de técnicas de aprendizado de máquina em bases de dados da engenharia florestal, que leva à pesquisa de novos métodos e algoritmos especializados para resolução dos problemas específicos da área.

1.1 Objetivos

Este trabalho tem como objetivo aplicar técnicas de aprendizado de máquina para representação de relações hipsométricas, predição de volumes e de biomassa de árvores confrontando os resultados com os modelos tradicionais de regressão. Através destas comparações, deve-se verificar se os modelos obtidos por meio de técnicas de aprendizado de máquina são competitivos com os gerados por regressão.

Assim, considerando os atuais métodos utilizados em mensuração florestal e as técnicas de aprendizado de máquina atualmente disponíveis, o objetivo geral desta tese é:

- Apresentar o aprendizado de máquina como um novo método para predição de volumes, relação hipsométrica e estimativa de biomassa, mostrando que o processo de Mensuração Florestal é um processo de Descoberta de Conhecimento (KDD - *Knowledge-discovery in databases*) e gerar modelos melhores que os ajustados por regressão;

Os objetivos específicos são:

- Apresentar alternativas para os problemas de mensuração apresentados (relação hipsométrica, predição de volumes e estimação de biomassa) avaliando técnicas de Aprendizado de Máquina, abrindo um leque de novas opções de pesquisa;
- Confrontar as técnicas utilizadas, usando-se medidas estatísticas de qualidade das aproximações, verificando se há superioridade dos modelos de aprendizado de máquina;
- Mostrar que Mensuração Florestal é um processo de KDD;
- Propor a adoção de técnicas de aprendizado de máquina na área de Engenharia Florestal como alternativa às tradicionais técnicas de regressão em relações hipsométricas, predição de volumes e estimativa de biomassa.

1.2 Estrutura do Trabalho

Esta tese está organizada em cinco capítulos. Neste primeiro foram apresentados os objetivos do trabalho, hipóteses e assertivas que sustentaram o desenvolvimento deste trabalho. O Capítulo 2 apresenta a fundamentação teórica do trabalho, tanto na área de Aprendizado de Máquina, como na Mensuração Florestal e Dendrometria. Também são mostrados os principais trabalhos envolvendo as duas áreas, aplicações e tendências. O Capítulo 3 apresenta os materiais e métodos usados no estudo. No Capítulo 4 são mostradas as configurações dos experimentos realizados, os resultados alcançados e sua discussão. Considerações finais e trabalhos futuros são apresentados no Capítulo 5.

Capítulo 2

Fundamentação Teórica

Este trabalho envolve duas áreas da ciência, a saber a Engenharia Florestal e a Ciência da Computação. Mais especificamente, a Mensuração Florestal e o Aprendizado de Máquina.

Assim, neste capítulo são descritas as bases teóricas para as pesquisas apresentadas nos capítulos seguintes.

2.1 Inventário Florestal e Mensuração Florestal

Os povoamentos florestais equiâneos são uma alternativa à demanda de madeira no Brasil e muitas empresas investem em medidas para melhorar a produção e aumentar os lucros. Neste contexto, determinar o volume de madeira em um povoamento é o objetivo central do Inventário Florestal.

O inventário florestal é uma técnica importante para o conhecimento de uma determinada área [Sanquetta et al., 2009], no que tange os recursos existentes, tanto quantitativos quanto qualitativos. Para tal, faz-se necessário o uso de técnicas para determinação de quais recursos se tem disponível nestas áreas, com a máxima precisão possível.

Segundo Pellico Neto e Brena [Netto e Brena, 1997], o inventário florestal visa obter informações qualitativas e quantitativas dos rescursos florestais em uma determinada área. Já Van Laar e Akça [van Laar e Akça, 2007] acrescentam como objetivo também a obtenção de informações sobre os recursos e o ambiente físico da floresta, em um ponto específico no tempo, a um custo razoável. Sob um ponto de vista mais amplo, Rondeux [Rondeux, 1999] afirma que o inventário florestal também deve responder a questões relacionadas com a biodiversidade de plantas e variáveis específicas que determinam a dinâmica do local.

Um inventário completo pressupõe a medida de todos os indivíduos da população. Mas esta medida completa é restrita a aplicações específicas, como florestas com alto valor econômico, científico ou cultural, ou exigências legais ou em áreas de pequenas dimensões [Hush et al., 2002; Sanquetta et al., 2009].

Assim sendo, deve-se distinguir entre determinação e estimativa das medições, conforme Sanquetta e outros [Sanquetta et al., 2004]. A determinação é feita de forma direta e implica na medição real dos valores dendrométricos, o que não é possível em grandes áreas florestais por conta da inviabilidade financeira e de tempo.

Nos casos em que as determinações não são viáveis, são feitas estimativas baseadas em medições diretas de um subconjunto de indivíduos e extrapoladas através de cálculos estatísticos para toda a população ou área.

Isto posto, o inventário florestal é baseado em estimativas extrapoladas através de dados obtidos por medição direta de uma amostra. Leva-se em consideração o tamanho e qualidade da

amostra (representatividade) e a margem de erro que os métodos estatísticos podem gerar. Estas incertezas são consideradas na análise do resultado tanto em termos comerciais como no aspecto de planejamento da atividade florestal.

2.1.1 Mensuração Florestal

Prodan [Prodan, 1997] define a mensuração florestal como o ramo da ciência florestal responsável por determinar/estimar valores dendrométricos de árvores, povoamentos e florestas, de seu crescimento e sub-produtos florestais.

Dendrometria, do grego (DENDRO=árvore, METRIA=medida), tem como objetivo a determinação ou estimativa de recursos florestais de povoamentos ou da própria árvore. Assumindo-se que, por exemplo para um inventário florestal, estimar os volumes produzidos por estes povoamentos, da forma mais precisa possível é crucial para a indústria e procedimentos imprecisos podem gerar muito prejuízo às empresas.

Um dos grandes objetivos é o Rendimento Sustentável [Dillewijn, 1968], quando há um equilíbrio entre a retirada de produtos florestais e o crescimento, mantendo a capacidade da floresta ou plantio de forma a fornecer permanentemente e racionalmente seus produtos florestais.

Para tal, a área da mensuração florestal envolve a Estimativa de Volumes, Relação Hipsométrica, Avaliação de Biomassa e Carbono, Distribuição Diamétrica, Modelos de Crescimento e Produção, Modelos de Afilamento e Sortimento.

A estimativa de volumes envolve a predição do volume de madeira de uma área baseando-se em medições de *dap* e altura de uma amostra desta área. Quanto mais precisa a estimativa, melhor a avaliação do volume para fins comerciais.

A relação hipsométrica é a relação que existe entre o *dap* e a altura das árvores. Como a medição da altura envolve ou o abate ou a utilização de equipamentos sensíveis, é consideravelmente mais demorada e mais custosa que a medição do *dap*. Assim, quanto menos árvores tiverem suas alturas determinadas de forma direta, mais barato e mais rápido será o processo em questão, por exemplo, estimativa de volume.

Biomassa é a quantidade de material vegetal por unidade em uma determinada área, expressa em unidade de massa. Estima-se que 45% da biomassa vegetal seja carbono, que é capturado da atmosfera a partir da fotossíntese. A avaliação de biomassa e carbono é importante pois as florestas são grandes absorventes de carbono da atmosfera, seu cultivo e preservação ajuda a diminuir a concentração de gases do efeito estufa emitidos. Quanto maior a quantidade de biomassa, maior é a emissão de gases do efeito estufa pelo desmatamento.

A distribuição diamétrica é uma avaliação dendrométrica usada para a definição de estoque de crescimento, sendo base para a tomada de decisões econômicas e silviculturais. Os modelos de crescimento e produção são importantes para o manejo de florestas e áreas plantadas, de forma a planejar a colheita, rotação, espécies adequadas a determinada área, etc.

Levando em consideração que as árvores possuem uma geometria específica, as funções de afilamento (ou funções de forma) são usadas para o cálculo do diâmetro de uma árvore a qualquer altura, ou a altura em um dado diâmetro. Este cálculo é importante para a estimação de volumes dos fustes de árvores em distintas seções de comercialização, permitindo assim a obtenção de volumes nos variados produtos finais a que se destina, fundamental para o cálculo do sortimento.

O sortimento é o estudo do maior aproveitamento do povoamento florestal, determinando-se assim a qualidade e quantidade de madeira existente. É a determinação do volume dos vários produtos que o povoamento pode oferecer. Uma vez escolhidas as seções de toras para os

diversos produtos, o cálculo do volume de cada produto é feito através da integração da função de afilamento em cada seção.

Neste trabalho três aspectos da mensuração florestal são tomados como base experimental, a saber: Relação Hipsométrica, Predição de Volumes e Avaliação de Biomassa.

Medições

As medições das variáveis podem ser feitas de forma direta, indireta ou estimativa. Na forma direta a variável de interesse é medida diretamente, como por exemplo o diâmetro à altura do peito e comprimento de toras. Na forma indireta são retirados valores que não estão ao alcance fácil, em geral com a ajuda de instrumetos óticos, como altura da árvore em pé e área basal. Nas estimativas, os dados são medidos para alguns indivíduos e depois extrapolados para a população ou área total.

O procedimento usual na área é a observação de alguns espécimes e o uso de relações hipsométricas, equações volumétricas e de afilamento, ajustadas de tempos em tempos, para a estimativa dos outros indivíduos, bem como de diferentes espécies ou regiões.

As principais medidas usadas para levantamentos florestais são a altura total e os diâmetros medidos em diversas alturas em relação ao solo. O Diâmetro à Altura do Peito (dap) é a medida do diâmetro da árvore a 1,3m e a Altura Total (h_t) é a altura total da árvore, da base ao topo.

2.2 Inteligência Artificial e Descoberta do Conhecimento

A IA é uma área de estudo preocupada com a construção e representação de entidades inteligentes. Seu início tem um marco em 1943, quando Warren McCulloch e Walter Pitts simularam uma rede de neurônios artificiais, que respondiam a estímulos de seus vizinhos [McCulloch e Pitts, 1943]. Nesta época, foi sugerido que este modelo poderia representar o aprendizado de máquina.

O termo "Inteligência Artificial"foi usado pela primeira vez em 1956, quando vários pesquisadores da área se reuniram para apresentar seus trabalhos em Dartmouth College, NH, USA. Os trabalhos apresentados eram na área de teoria dos autômatos, redes neurais e estudo da inteligência. O encontro foi idealizado por John McCarthy, Marvin Minsky, Claude Shannon e Nathaniel Rochester. Nesta época destacam-se Allen Newell e Hervert Simon, com o *Logic Theorist*, um programa que demostrava teoremas matemáticos.

No início da IA houve muito entusiasmo e otimismo por parte dos pesquisadores, principalmente devido aos grandes progressos na área. Mas logo muitas dificuldades começaram a aparecer, devido à grande complexidade dos projetos propostos, que eram muito diferentes das tarefas simples até então implementadas. Outra dificuldade foi a *explosão combinatória* encontrada em diversos problemas, bem como o poder de processamento da época.

A primeira abordagem em IA era o uso de passos fundamentais do raciocínio para encontrar uma solução completa de um problema. Este era o *método fraco*, pois usava pouca informação do domínio, não permitindo a resolução de problemas difíceis [Russell e Norvig, 2010].

Para contornar este problema, a solução foi inserir conhecimento nos programas, de tal forma que começaram a ser chamados de *Sistemas Especialistas*. Na década de 80, estes sistemas se tornaram comerciais [Russell e Norvig, 2010].

A partir de 1987, a IA começou a ser tratada com base em metodologia científica, o que levou os pesquisadores a compreender melhor os problemas e suas complexidades. Mesmo assim, com poucos avanços no foco principal da IA, que era criar máquinas pensantes como

humanos, deu-se início ao *inverno da IA*, onde empresas faliram e departamentos especializados nesta área foram exintos. Deu-se início à pesquisa específica em IA, para resolver problemas pontuais (conhecido com o *IA estreita*) [Russell e Norvig, 2010].

Em meados dos anos 80, John Hopfield retoma os estudos sobre Redes Neurais Artificiais (RNA), iniciados nos anos 70 [Hopfield, 1982]. A evolução destas pesquisas, bem como a difusão da tecnologia em empresas e aplicações, abriram um leque de possibilidades de estudo no ramo da engenharia do conhecimento e cognição, que mais tarde evoluiu para o aprendizado de máquina, que é a área da IA responsável pelo desenvolvimento de teorias computacionais com foco no aprendizado automatizado.

2.2.1 Descoberta de Conhecimento

Com a crescente evolução da computação e da eletrônica, tecnologias possibilitaram um aumento considerável no processamento e armazenamento de dados. Esse aumento de dados extrapola a capacidade de pesquisadores e analistas em analisar estes dados, tanto em termos de tempo como quantidade.

Diante desta deficiência em analisar e processar grandes volumes de dados, diversas pesquisas têm sido direcionadas para o desenvolvimento de técnicas de descoberta automática de conhecimento em bases de dados. Essa área de pesquisa é a Descoberta de Conhecimento em Bases de Dados (KDD do inglês *Knowledge-discovery in databases*).

O KDD é um processo semiautomático pelo qual se analisa e se produz conhecimento de um conjunto bruto de dados, a um custo computacional aceitável. Segundo [Fayyad et al., 1996b], é o processo não trivial de indentificar padrões novos, válidos, pontencialmente úteis e explicáveis, a partir de uma base de dados. É formado pelas seguintes etapas:

- Identificação do Problema: estudo do domínio e definição de objetivos a serem alcançados pelo processo;
- Pré-processamento: transformação dos dados em formatos utilizáveis pela técnica escolhida, por exemplo, redução de dimensionalidade, combinação de atributos, etc;
- Extração de Padrões: cumprimento dos objetivos definidos, compreendendo a escolha da tarefa, escolha do algoritmo e a extração de padrões propriamente dito;
- Pós-processamento: refinamento do conhecimento obtido nas etapas anteriores, de maneira a torná-lo usável;
- Utilização do Conhecimento: interpretação dos resultados e efetiva utilização destes.

A Figura 2.1 apresenta o processo de KDD e sua característica iterativa. O processo engloba todo o ciclo que os dados percorrem desde o início até se transformarem em informação significativa, sendo portanto uma sequência finita de operações onde o resultado de cada etapa depende do resultado das etapas anteriores. KDD também possui característica interativa, onde o usuário pode controlar o fluxo das atividades conforme sua necessidade [Fayyad et al., 1996a].

De acordo com os objetivos do processo deve-se escolher a característica da tarefa para a fase de extração de padrões. As tarefas podem ser preditivas ou descritivas [Faceli et al., 2011].

Tarefas preditivas encontram uma função (modelo ou hipótese), a partir de dados de treinamento, que possa prever um rótulo ou um valor que expresse um novo exemplo baseado em suas características de entrada. Estes métodos levam a modelos supervisionados, onde, no

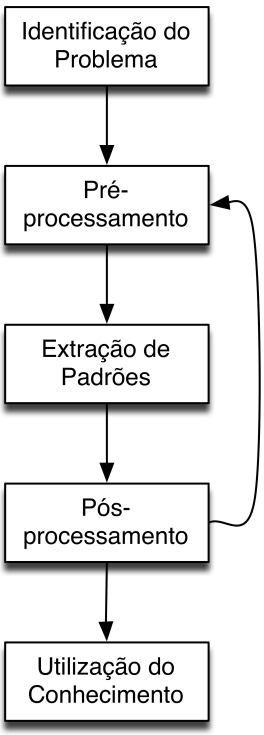


Figura 2.1: Processo iterativo de KDD, apresentando suas etapas e sequenciamento.

Fonte: O autor.

treinamento, já se conhece as classes ou valores de saída destas informações, simulando um agente externo que avalia o aprendizado.

As tarefas preditivas são:

- Classificação: é a predição de um valor categórico, uma classe, por exemplo, se há ou não risco em fazer um empréstimo para um cliente;
- Regressão: é a predição de um valor real, por exemplo, qual será o lucro com o empréstimo efetuado.

Tarefas descritivas dizem respeito a explorar ou descrever um conjunto de dados. Estas tarefas não têm um atributo como saída e, portanto, não são técnicas supervisionadas. Como exemplo tem-se a categorização de grupos de objetos semelhantes.

As tarefas descritivas são:

- Associação: é a busca por padrões frequentes de associações entre os atributos de um conjunto de dados;
- Agrupamento: é a identificação de porções de dados, em grupos ou *clusters*, que possuem características semelhantes, conforme determinados critérios.

Neste trabalho são usadas de forma mais efetiva as técnicas de aprendizado de máquina, que são aplicadas na etapa de extração de padrões do processo de KDD. Mais especificamente, as tarefas preditivas de regressão, que podem substituir a regressão clássica comumente usada para resolver problemas de estimação de valores dendrométricos na área de Engenharia Florestal.

2.2.2 Aprendizado de Máquina

A partir da década de 1970, técnicas baseadas em inteligência artificial começaram a ser usadas de forma mais ampla na solução de problemas reais [Russell e Norvig, 2010]. Neste ponto, a IA deixou de ter um viés essencialmente teórico para ganhar um aspecto prático.

Uma grande gama de problemas práticos resolvidos pelas técnicas de IA eram os que faziam uso de conhecimento de um determinado domínio (por exemplo, Medicina) para que fossem feitas inferências de regras, descoberta de diagnósticos e detecção de infecções. Os programas desenvolvidos eram chamados de Sistemas Especialistas ou Sistemas Baseados em Conhecimento.

Estes sistemas eram dependentes da aquisição do conhecimento através de um especialista na área, processo este sujeito à subjetividade, falta de cooperação, intuição do especialista, informações incertas, entre outros fatores críticos para a qualidade das informações coletadas.

Com a crescente complexidade dos problemas e o grande volume de dados gerados, tornou-se necessário o desenvolvimento de ferramentas mais sofisticadas que pudessem resolver problemas sem dependência e intervenção humana. Estas ferramentas deveriam ser capazes de usar as experiências passadas e criar hipóteses para resolver um determinado problema.

O processo de obtenção de conclusões genéricas a partir de um conjunto particular de exemplos é conhecido como indução, e é um processo de inferência lógica.

Induzir hipóteses a partir de um conjunto de dados, que é a experiência passada de algum problema, é chamado de aprendizado de máquina [Faceli et al., 2011]. Aprendizado de máquina, portanto, é a área da IA responsável por desenvolver técnicas computacionais sobre o aprendizado e construir sistemas capazes de adquirir conhecimento de forma autônoma [Rezende, 2003]. Segundo Mitchell [Mitchell, 1997], aprendizado de máquina é "A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência."

Existem vários algoritmos de AM e cada um se utiliza de uma forma (ou viés) de representação dos dados e de busca. Por exemplo, um algoritmo de árvore de decisão usa uma estrutura de árvore para representação e como viés de busca utiliza a preferência por árvores com poucos nós.

Algoritmos de AM também são organizados de acordo com o paradigma de aprendizado adotado para lidar com a tarefa. As tarefas de aprendizado podem ser divididas em preditivas e descritivas, conforme descritas anteriormente.

Assim, o aprendizado pode ser categorizado em dois tipos:

- Aprendizado supervisionado: onde se conhecem os atributos de saída do conjunto de treinamento e, com isso, pode-se avaliar a capacidade do resultado induzido de predizer os valores para novos elementos;

- Aprendizado não supervisionado: quando não se conhecem atributos de saída e deseja-se agrupar ou encontrar regras de associação de dados de um conjunto.

A Figura 2.2 apresenta a hierarquia de aprendizado, baseando-se nos tipos de tarefas de aprendizado. Em relação aos diferentes tipos de tarefas, domínios e objetivos relacionados dos algoritmos de AM, existem vários critérios para auxiliar na escolha do algoritmo mais adequado para cada situação. De acordo tipo de conceito utilizado para induzir uma determinada hipótese, os algoritmos de aprendizado podem ser estruturados em paradigmas [Rezende, 2003], a saber:

- **Paradigma simbólico:** realiza o processo de aprendizagem utilizando representações simbólicas por meio da análise de exemplos e contra-exemplos. Essas representações estão, geralmente, na forma de expressão lógica, árvore de decisão ou rede semântica.

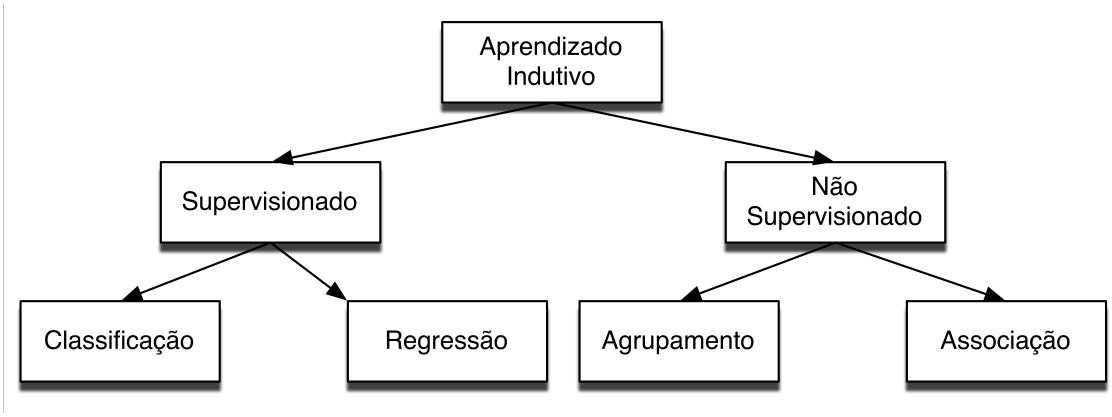


Figura 2.2: Hierarquia de Aprendizado

Fonte: O autor, adaptado de [Faceli et al., 2011].

Um exemplo é o algoritmo C4.5 [Quinlan, 1993], que se baseia na indução de árvores de decisão;

- **Paradigma baseado em exemplos:** armazena os exemplos e utiliza medidas de similaridade para identificar os casos mais similares ao exemplo a ser analisado. Um exemplo é o algoritmo k-vizinhos mais próximos (k-NN) [Aha et al., 1991];
- **Paradigma estatístico:** utiliza modelos estatísticos para encontrar uma aproximação do conceito induzido. Dentre as abordagens existentes, destacam-se o aprendizado Bayesiano [Mitchell, 1997] e máquinas de vetores de suporte (SVM do inglês *Support Vector Machines*) [Cortes e Vapnik, 1995];
- **Paradigma conexionista:** baseia-se em construções matemáticas inspiradas em conexões neuronais do sistema nervoso humano. Como exemplo tem-se as RNAs [Haykin, 1999];
- **Paradigma evolutivo:** baseia-se em uma população de elementos de classificação que competem para fazer a predição. Desse modo, os elementos que apresentam desempenho mais fraco, segundo algum critério, não são considerados [Back, 1996]. As técnicas desenvolvidas nesse paradigma são inspiradas pela teoria de Darwin, na qual os indivíduos mais adaptados sobrevivem. Como exemplo tem-se os algoritmos genéticos (AG) [Faceli et al., 2011].

Dentre estes paradigmas, para este trabalho são de interesse os Modelos Preditivos de Regressão e são estudados as Redes Neurais Artificiais (Paradigma Conexionista), Máquinas de Vetores de Suporte para Regressão (Paradigma Estatístico) e Métodos baseados em árvores, *Random Forest* (Floresta Aleatória) (Paradigma Simbólico). A escolha destes modelos se deu pelo fato de cada um seguir um paradigma diferente, o que é interessante para efeitos comparativos.

Modelos Preditivos

Um algoritmo de AM preditivo tem como entrada um conjunto de exemplos rotulados e gera um *estimador*, que pode ser usado em um conjunto diferente de dados (não rotulados).

Se os rótulos são do tipo discreto (classes, categorias), então o problema é de classificação e o estimador gerado pelo algoritmo é um *classificador*. Se os rótulos são um conjunto valores, então é um problema de regressão e o estimador gerado é um *regressor*. Um classificador (ou

regressor) é uma função onde, dado um exemplo não rotulado, atribui uma classe (ou um valor) a este exemplo [Dietterich, 1998].

Assim, um classificador tem como objetivo analisar um conjunto de dados e decidir quais fazem parte de qual classe, dentre as disponíveis. Já um regressor tem como objetivo aproximar uma função a partir deste novo conjunto de dados. Ambos são ajustados através dos dados iniciais rotulados.

Neste trabalho, como os dados utilizados são rotulados através de um conjunto de valores reais, o processo a ser utilizado é o de regressão.

RNAs - Redes Neurais Artificiais

As RNAs são uma tentativa de representação de um modelo computacional do sistema nervoso, a fim de simular a capacidade de aprendizado do cérebro humano. Os primeiros estudos surgiram juntamente com o advento dos primeiros computadores eletrônicos. Em 1943, McCulloch e Pitts [McCulloch e Pitts, 1943] propuseram um modelo matemático de um neurônio artificial, que executava operações simples. Eles mostraram que a combinação de vários neurônios aumenta o poder de processamento do conjunto como um todo, mas estas redes iniciais não tinham capacidade de aprendizado.

Em [Rosenblatt, 1958] foi apresentado o estudo sobre redes Perceptron, mas as redes neurais foram deixadas de lado por conta de Minsky e Papert [Minsky e Papert, 1969] que em 1969 mostraram que as redes perceptron eram limitadas a funções linearmente separáveis. Só nos anos 80 é que as atenções se voltaram para as RNAs e em 1989 [Cybenko, 1989] é mostrado que uma rede neural com uma camada intermediária pode implementar qualquer função contínua e com duas camadas pode implementar qualquer função.

Existem diversos tipos de redes neurais que podem ser usados para classificação e predição, mas neste trabalho optou-se por *Multilayer Perceptron* (MLP, uma rede perceptron com várias camadas ocultas) por ser a mais estudada e usada [Lippmann, 1989; Bishop, 1995].

Uma RNA é um sistema computacional composto por unidades simples de processamento, altamente conectadas. Estas unidades, ou neurônios, computam funções matemáticas e seus resultados são processados em conjunto nas demais camadas da rede, conforme ilustra a Figura 2.3. As conexões simulam sinapses biológicas e possuem pesos associados às suas entradas. Esses pesos são ajustados a medida que o conjunto todo é treinado, isto é, aprendendo o conhecimento adquirido [Braga et al., 2007].

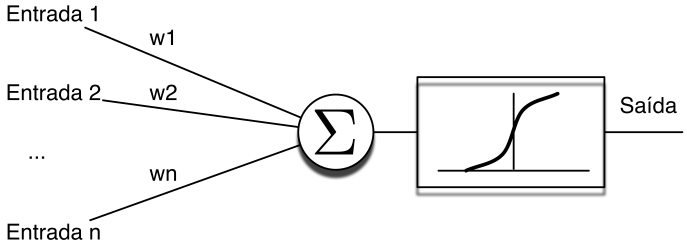


Figura 2.3: Estrutura de um neurônio apresentando os seus valores de entrada, seus pesos, o somatório realizado e a função de ativação que resulta na sua saída.

Fonte: O autor, adaptado de [Faceli et al., 2011].

Um neurônio recebe valores e retorna um resultado. Os valores de entrada são ponderados, combinados (somados) e passam por uma função matemática f_a. Assim, se o vetor

$x = [x_1, x_2, ..., x_m]^t$ é a entrada de um neurônio e o vetor $w = [w_1, w_2, ..., x_m]^t$ representa os pesos aplicados a cada entrada, o resultado do neurônio $f'(x)$ é:

$$u = \sum_{i=1}^{m} x_i \times w_i \tag{2.1}$$

$$f'(x) = f_a(u) \tag{2.2}$$

A função de ativação f_a pode ser de vários tipos, mas três são comumente utilizadas: linear, limiar e sigmoidal [Haykin, 2001]. A função linear implica em retornar um múltiplo de u. A função limiar, empregada no neurônio de McCulloch e Pitts define quando o resultado é 1 ou 0 (ou -1), através de um valor limite (limiar). A função sigmoidal representa uma aproximação contínua e diferenciável da função limiar.

Os neurônios podem ser dispostos em uma ou mais camadas. Neste último caso, um neurônio recebe como entrada as saídas dos neurônios da camada anterior e sua saída é colocada na próxima camada. A camada que dá o resultado é denominada camada de saída. As demais são camadas ocultas. A Figura 2.4 ilustra a estrutura de uma RNA.

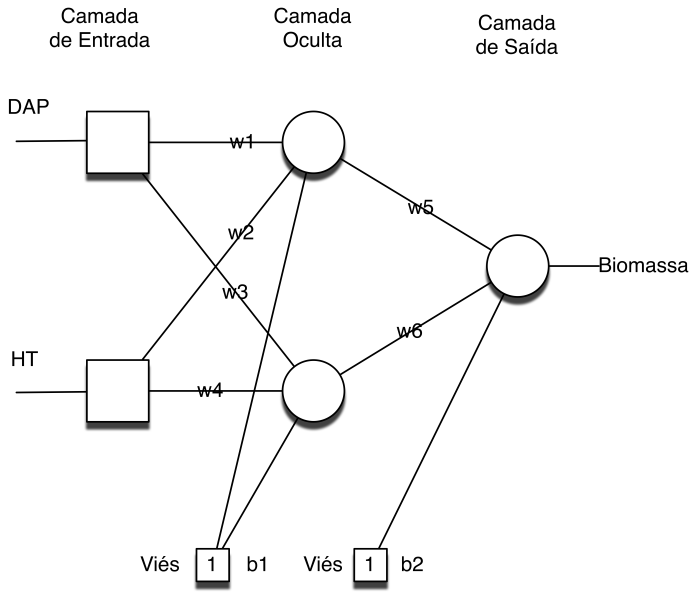


Figura 2.4: Estrutura de uma RNA, aprensentando a camada de entrada, que recebe os valores da base de dados, uma camada oculta de processamento e a camada de saída, que retorna o resultado estimado.

Fonte: O autor.

Para resolver problemas não linearmente separáveis deve-se usar uma ou mais camadas ocultas, conforme [Cybenko, 1989]. Redes multicamadas, como a MLP, usam funções não lineares em suas camadas ocultas, como a função sigmoidal. No caso de problemas de classificação, cada neurônio de saída corresponde a uma classe. No caso de problemas de regressão, a saída não pode ser discretizada e é retornado um valor decimal. Em geral, as MLPs possuem cada neurônio de uma camada conectado a todos os neurônios da próxima camada, mas não havendo conexões entre neurônios da mesma camada.

Um dos algoritmos para treinamento de redes multicamadas é o *back-propagation* [Rumelhart et al., 1986], que é implementado nas redes MLP do WEKA. Ele é constituído por

duas fases, uma para frente (*forward*) e uma para trás (*backward*). Na fase para frente o objeto é apresentado à rede, os neurônios calculam seus valores com as ponderações específicas e a função de ativação produz seu valor de saída. Isso é feito até que os neurônios de saída tenham seus valores calculados. O resultado computado é comparado com o resultado esperado e esta diferença é o erro cometido pela rede.

Na fase para trás (*backward*) usa-se o erro cometido pela rede para ajustar os pesos dos neurônios. O processo matemático é o de descobrir o quanto cada neurônio influencia no erro total da rede. Isto é, dado um peso w_i de alguma conexão na rede, calcula-se a derivada parcial do erro total da rede com respeito a w_i ($\frac{\partial E_{total}}{\partial w_i}$).

Para cada peso w_i, calcula-se o novo valor aplicando-se a fórmula:

$$w_i^+ = w_i - \eta * \frac{\partial E_{total}}{\partial w_i} \tag{2.3}$$

onde η é a taxa de aprendizado, um fator multiplicativo aplicado à rede.

Para aumentar a velocidade de treinamento da rede, bem como reduzir a sua instabilidade, é adicionado um termo chamado *momentum*, que pode variar de 0,0 (não utilizado) até 1,0. Este termo adicionado ao treinamento leva em consideração as alterações passadas para a convergência da rede, útil em casos em que a rede encontra um mínimo local e o erro para de diminuir, estacionando em um valor maior que o aceitável.

No processo de treinamento como um todo, os pesos são atualizados após toda a amostra de treinamento ser apresentada à rede. Assim, pode-se definir também o número de épocas ou ciclos no treinamento, que indica a quantidade máxima de vezes que a amostra será entrada na rede. Um número grande pode levar ao *overfitting*, já um número pequeno pode gerar uma rede sem a capacidade de generalização.

Para evitar a aplicação excessiva de ciclos, a partir do conjunto de treinamento, pode-se definir um conjunto de validação, para o qual a estimativa do erro é calculada a cada quantidade definida de épocas. A partir do momento que o erro obtido começa a crescer, o processo de treinamento para, indicando que a rede pode começar a perder generalização.

Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*) surgiram com os estudos iniciados em [Vapnik e Chervonenkis, 1971] e foram consolidados em [Vapnik, 1995], e são baseados na teoria de aprendizado estatístico. SVMs podem ser usados para classificar dados linearmente separáveis e também podem ser estendidos para gerar fronteiras não lineares. Há também uma formulação de SVM que pode ser usada para problemas de regressão.

SVMs lineares com margens rígidas definem fronteiras lineares para dados que são linearmente separáveis, isto é, encontrando um hiperplano que separa os objetos das classes. A equação do hiperplano é:

$$h(x) = w \times x + b \tag{2.4}$$

onde X é o espaço de entrada, $x \in X$ é o vetor de entrada, w é um vetor de pesos ajustáveis, $w \times x$ é o produto escalar entre os vetores w e x e $b/\|w\|$ corresponde à distância do hiperplano em relação à origem, com $b \in \Re$. Essa equação divide a entrada X em duas regiões, conforme a Figura 2.5, $w \times x + b > 0$ e $w \times x + b < 0$ e uma função sinal $g(x) = sgn(h(x))$ é usada para obtenção da classificação, da seguinte forma:

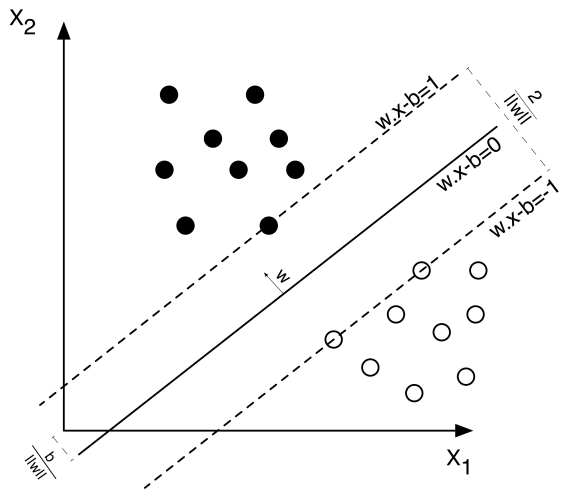


Figura 2.5: Hiperplano para classificação de dados lineares em duas dimensões.

Fonte: O autor, adaptado de [Faceli et al., 2011].

$$g(x) = sgn(h(x)) = \begin{cases} +1 & se\ w \times x + b > 0 \\ -1 & se\ w \times x + b < 0 \end{cases} \tag{2.5}$$

Como é possível obter infinitos hiperplanos equivalentes através de $h(x)$, usa-se o hiperplano canônico, o qual w e b são escalados para satisfazer a equação:

$$|w \times x_i + b| = 1 \tag{2.6}$$

Que implica nas seguintes inequações:

$$y_i(w \times x_i + b) - 1 \geq 0, \forall x_i, y_i \in X \tag{2.7}$$

Assim, a maximização da margem de separação dos objetos em relação a $w \times x + b = 0$ é obtida pela minimização de $\|x\|$, através do seguinte problema de otimização:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2.8}$$

com as restrições:

$$y_i(w \times x_i + b) - 1 \geq 0, \forall i = 1, ..., n \tag{2.9}$$

Estas restrições garantem que os dados de treinamento não aparecem entre as margens de separação, e por isso são chamados de Margens Rígidas. Pelas características da função objetivo, ela possui somente um mínimo global e pode ser resolvida através de uma função Lagrangiana, da seguinte forma:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i(y_i(w \times x_i + b) - 1) \tag{2.10}$$

Esta função deve ser minimizada, o que implica em maximizar as variáveis α_i, enquanto w e b devem ser minimizados. Derivando L em relação a b e w e igualando a zero, obtém-se:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.11}$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{2.12}$$

e substituindo estas equações na Lagrangiana, obtém-se:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{2.13}$$

com as restrições:

$$\begin{cases} \alpha_i \geq 0, \forall i = 1, ..., n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{2.14}$$

Esta formulação é chamada dual e o problema original é chamado primal. A forma dual possui restrições mais simples e permite representar o problema de otimização em termos de produtos internos entre objetos, útil para resolver problemas não-lineares.

Sejam α^* a solução do problema dual e w^* e b^* as soluções da forma primal. Pode-se obter w^* a partir de α^* através da equação 2.12. O parâmetro b^* é definido por α^* e por condições provenientes da teoria de otimização com restrições. Estas restrições afirmam que no ponto ótimo o produto entre as variáveis duais de Lagrange e as restrições deve ser nulo. Tem-se:

$$\alpha_i^* (y_i (w^* x_i + b^*) - 1) = 0, \forall i = 1, ..., n \tag{2.15}$$

Nesta equação, α_i^* pode ser diferente de 0 somente para os objetos que se encontram sobre os hiperplanos H_1 e H_2, que se encontram mais próximos ao hiperplano separador, exatamente sobre as margens. Para os demais, a condição é obedecida com $\alpha_i^* = 0$ e não participam do cálculo de $w*$. Os exemplos que possuem $\alpha_i^* > 0$ são denominados Vetores de Suporte (*support vectors*) e são os objetos mais informativos do conjunto de treinamento. O valor de b^* é calculado pela seguinte equação, onde n_{SV} é o número de vetores de suporte e SV representa o conjunto dos vetores de suporte:

$$b^* = \frac{1}{n_{SV}} \sum_{x_j \in SV} \frac{1}{y_j} - w^* x_j = \frac{1}{n_{SV}} \sum_{x_j \in SV} \left(\frac{1}{y_j} - \sum_{x_j \in SV} \alpha^* y_i x_i x_j \right) \tag{2.16}$$

E como resultado final, obtém-se o seguinte classificador $g(x)$:

$$g(x) = sgn(h(x)) = sgn\left(\sum_{x_i \in SV} y_i \alpha^* x_i x + b^* \right) \tag{2.17}$$

Como em situações reais é difícil encontrar aplicações com dados linearmente separáveis, as SVMs devem ser estendidas para lidar com conjuntos de treinamentos mais gerais. Assim, deve-se permitir que dados possam violar as restrições impostas, basicamente adicionando-se

variáveis de folga ξ_i, para todo $i = 1, .., n$, relaxando assim as restrições impostas ao problema de otimização primal. Obtém-se então as seguintes restrições:

$$y_i(wx_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1, ..., n \tag{2.18}$$

Isso permite que alguns objetos permaneçam entre os hiperplanos H_1 e H_2, e também ocorram alguns erros de classificação. Por isso, estas máquinas são conhecidas como SVMs com Margens Suaves. Um erro no treinamento é indicado quando $\xi_i > 1$, assim a soma de todos os ξ_i representa um limite no número de erros de treinamento. Para minimizar este erro, a função objeto é reformulada como:

$$Minimizar_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \tag{2.19}$$

A constante C impõe um peso à minimização dos erros em relação à minimização da complexidade do modelo. A somatória dos erros pode ser vista como uma minimização dos erros marginais, pois um valor de $\xi_i \in (0, 1]$ indica um objeto entre as margens. O problema de otimização apresentado assim possui as mesmas propriedades do anterior e, portanto, pode ser resolvido através de uma função Lagrangiana, tornando suas derivadas parciais nulas. Como resultado, o problema dual é:

$$Maximizar_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{2.20}$$

com as restrições:

$$\begin{cases} 0 \leq \alpha_i \leq C, \forall i = 1, ..., n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{2.21}$$

Este problema é o mesmo das máquinas com margens rígidas, exceto pela restrição em α_i que são limitados por C. Sejam α^* a solução do problema dual e w^*, b^* e ξ^* as soluções da forma primal. O vetor w^* é determinado da mesma forma. As variáveis ξ_i^* podem ser calculadas pela seguinte equação [Cristiani e Shawe-Taylor, 2000]:

$$\xi_i^* = \max\left\{0, 1 - y_i \sum_{j=1}^{n} y_i \alpha_j^* x_j x_i + b^*\right\} \tag{2.22}$$

As condições são:

$$\begin{aligned} &\alpha_i^*(y_i(w^* x_i + b^*) - 1 + \xi_i^*) = 0, \forall i = 1, ..., n \\ &(C - \alpha_i^*)\xi_i^* = 0 \end{aligned} \tag{2.23}$$

Os pontos x_i para os quais $\alpha_i^* > 0$ são os vetores de suporte (SV), isto é, os objetos que participam da formação do hiperplano separador. No caso da formulação com margens suaves, dois tipos de vetores de suporte podem ser encontrados. Se $\alpha_i^* < C$, então $\xi_i^* = 0$ portanto esses SVs se encontraram sobre as margens e são denominados livres. Os SVs para os quais $\alpha_i^* = C$ podem representar três casos:

- erros, se $\xi_i^* > 1$;
- pontos corretamente classificados, porém entre as margens, se $0 < \xi_i^* \leq 1$;
- pontos sobre as margens, se $\xi_i^* = 0$.

O último caso ocorre raramente e os demais são chamados de *limitados*. Para calcular b^*, computa-se a média da equação a seguir, sobre todos os SVs x_j entre as margens, isto é, os que possuem $\alpha_i^* < C$.

$$b^* = \frac{1}{n_{SV}} \sum_{x_j \in SV} \frac{1}{y_j} - w^* x_j = \frac{1}{n_{SV}} \sum_{x_j \in SV} \left(\frac{1}{y_j} - \sum_{x_i \in SV} \alpha^* y_i x_i x_j \right) \tag{2.24}$$

Como resultado final, a mesma função de classificação 2.17 é obtida.

Problemas não-lineares

Em casos de problemas não lineares, não é possível dividir os dados de forma satisfatória com o uso de um hiperplano. Para que SVMs possam trabalhar com este tipo de dados, o conjunto de treinamento é mapeado para um espaço de maior dimensão, chamado de espaço de características [Hearst et al., 1998]. Isso é feito através de uma função de mapeamento que transforma os dados de seu espaço original para o espaço de características, conforme ilustra a Figura 2.6.

Esta manipulação é baseada no Teorema de Cover [Cover, 1965] sobre a separabilidade dos padrões, que atesta:

> "Um problema complexo de classificação de padrões colocado em um espaço de dimensão elevada tem mais chance de ser resolvido linearmente (ou seja, de ser um problema descrito por classes linearmente separáveis) do que quando colocado em um espaço de dimensão baixa."

A escolha apropriada desta função faz com que um conjunto não-linearmente separável seja classificado por uma SVM linear.

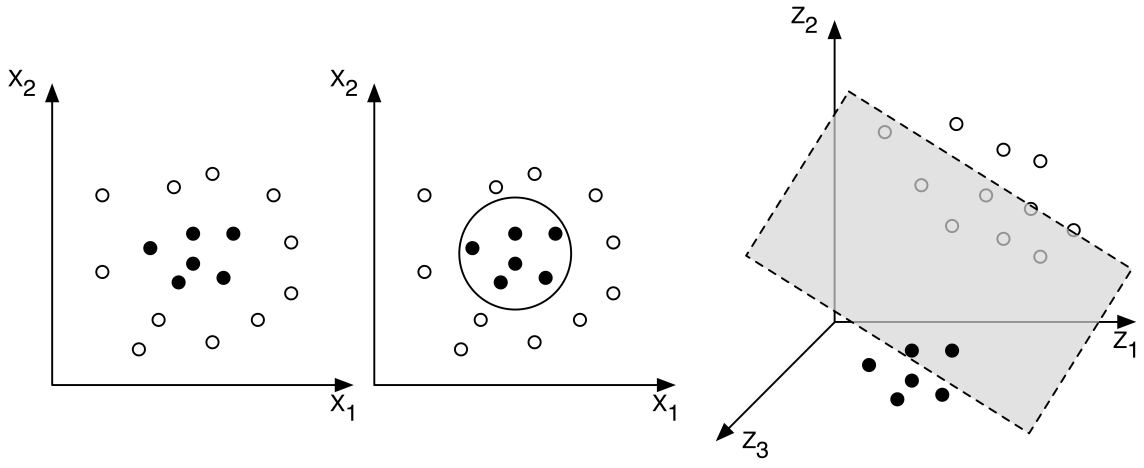


Figura 2.6: Transformação dos dados não lineares para o espaço de características. Percebe-se que na dimensão original não há hiperplano separador, mas ao se transformar os dados o hiperplano pode ser obtido.

Fonte: O autor, adaptado de [Faceli et al., 2011].

Para realizar o mapeamento, aplica-se a função de mapeamento Φ sobre os dados, obtendo-se assim o seguinte problema de otimização:

$$Maximizar_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\Phi(x_i)\Phi(x_j)) \tag{2.25}$$

E o classificador extraído se torna:

$$g(x) = sgn(h(x)) = sgn\left(\sum_{x_i \in SV} y_i \alpha_i^* \Phi(x_i)\Phi(x) + b^* \right) \tag{2.26}$$

e b^* é calculado como:

$$b^* = \frac{1}{n_{SV_{\alpha^* < C}}} \sum_{x_j \in SV_{\alpha_j^* < C}} \left(\frac{1}{y_j} - \sum_{x_i \in SV} \alpha^* y_i \Phi(x_i)\Phi(x_j) \right) \tag{2.27}$$

Visto que a dimensão do espaço de características pode ser muito alta, a computação de Φ pode ser onerosa ou até mesmo inviável. Mas, pelas equações vistas anteriormente, a única informação necessária sobre o mapeamento é a realização de produtos escalares entre os objetos no espaço de características, que pode ser obtido através de funções denominadas *kernels*.

Uma função *kernel* é uma função que recebe dois pontos x_i e x_j no espaço original e calcula o produto escalar desses dois objetos no espaço de características [Herbrich, 2001].

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \tag{2.28}$$

Pode-se aplicar a função de *kernel* sem se conhecer o mapeamento Φ, garantindo simplicidade no seu cálculo, desde que sejam seguidas condições estabelecidas pelo teorema de Mercer [Mercer, 1909]. O teorema de Mercer afirma que qualquer função *kernel* positiva definida satisfaz a equação:

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geq 0 \tag{2.29}$$

Isto é, possui um mapeamento Φ. Utilizar uma função *kernel* evita a necessidade de se trabalhar diretamente no espaço de características, evitando o mapeamento direto dos elementos usando-se Φ. Como resultado, obtém-se somente o produto escalar dos objetos x_i e x_j quando mapeados para este espaço. Três são as funções mais utilizadas e que satisfazem o teorema de Mercer: Polinomial, RBF (*Radial Basis Function*) e Sigmoidal, que podem ser vistas na Tabela 2.1.

Tabela 2.1: Funções de Kernel usadas em SVMs

Tipo de Kernel	Função $K(x_i, x_j)$	Parâmetros
Polinomial	$(\delta(x_i x_j) + \kappa)^d$	δ, κ, d
RBF	$exp(-\sigma\|x_i - x_j\|^2)$	σ
Sigmoidal	$tanh(\sigma(x_i x_j) + \kappa)$	σ, κ

Para se obter um classificador usando SVMs, deve-se escolher a função de *kernel*, bem como seus parâmetros e a constante de regularização C.

Problemas de Regressão

Para problemas de Regressão, que é o caso a ser aplicado neste trabalho, o problema de otimização deve ser reformulado, como visto em [Smola e Schölkopf, 2004]. Basicamente, ao invés de se encontrar um hiperplano separador no espaço de características, encontra-se uma função que obtenha os valores de treinamento, em uma variação de no máximo ε, e que seja o mais plana possível.

O algoritmo ε-SVR (*Support Vector Regression*) [Vapnik, 1995] encontra uma função $h(x)$ que resulte em saídas contínuas para os dados de treinamento, desviando no máximo ε de seu valor desejado. No caso de funções h lineares, deve-se buscar uma função com pequeno w, minimizando-se $\|w\|$ através do seguinte problema de otimização:

$$Minimizar_{w,b}\frac{1}{2}\|w\|^2 \tag{2.30}$$

com restrições:

$$\begin{cases} y_i - wx_i - b \leq \varepsilon_i \\ wx_i + b - y_i \leq \varepsilon_i \end{cases} \tag{2.31}$$

Então procura-se a função linear que aproxime os pares (x_i, y_i) de treinamento com precisão ε, conforme ilustra a Figura 2.7.

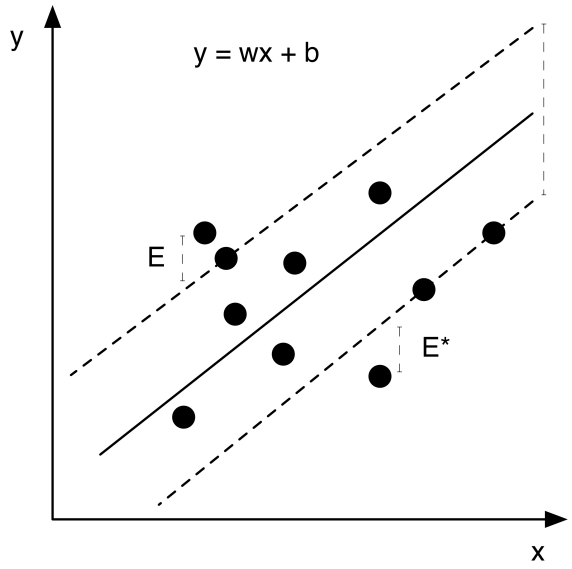


Figura 2.7: Procedimento de Regressão via SVM para o caso em duas dimensões.

Fonte: O autor, adaptado de [Faceli et al., 2011].

Para SVMs com margens suaves, adicionam-se variáveis de folga e obtém-se o problema dual através do uso de uma Lagrangiana, tornando nulo o resultado das derivações parciais e substituindo-se as expressões resultantes na equação original. Para regressões não lineares,

aplicam-se funções de *kernel*, as mesmas para classificação, e o problema de otimização final é dados por:

$$Maximizar_{\alpha,\bar{\alpha}} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\bar{\alpha}_i)(\alpha_j-\bar{\alpha}_j)K(x_i,x_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i-bar\alpha_i)+\sum_{i=1}^{n}y_i(\alpha_i-\bar{\alpha}_i) \quad (2.32)$$

com restrições:

$$\begin{cases} \sum_{i=1}^{n}(\alpha_i-\bar{\alpha}_i)=0 \\ \alpha_i,\bar{\alpha}_i \in [0,C] \end{cases} \quad (2.33)$$

Nestas equações, α_i e $\bar{\alpha}_i$ representam as variáveis de Lagrange e K a função de *kernel*.

Algoritmo *Sequential Minimal Optimization*

O algoritmo de Otimização Sequencial Mínima (SMO do inglês *sequential minimal optimization*) é um algoritmo para treinamento de SVM [Platt, 1999]. O objetivo é melhorar o desempenho no treinamento de SVM, decompondo o problema de programação quadrática subjacente em uma série de sub-problemas, resolvendo-os de forma analítica com apenas duas variáveis. Outra vantagem é que a quantidade de espaço utilizada pelo algoritmo é linear ao tamanho do conjunto de treinamento, podendo tratar grandes quantidades de dados.

Assim, para os experimentos realizados neste trabalho, foi usado o SMO como algortimo de treinamento das SVMs obtidas.

Árvores de Regressão e Árvores de Modelos

Esta seção descreve as técnicas de AM baseadas na indução de árvores. Estes são os conceitos e princípios básicos usados nas Florestas Aleatórias, que serão vistas em seguida. As árvores possuem dois tipos de nós:

- Nós de Decisão: são os nós internos, usados para se caminhar na árvore até atingir as folhas. Estes nós em geral possuem condicionantes que fazem com que um caminho seja seguido, conforme os dados sendo aplicados;
- Nós Folha: são nós externos ou nós objetivo. Com o caminhamento na árvore através dos nós de decisão, usando os dados de entrada como direcionadores, os nós folha representam os resultados da predição. Podem ser valores reais ou uma função a ser aplicada sobre os valores utilizados para se chegar neste nó.

Conforme o problema sendo abordado (classificação ou regressão), os nós folha da árvore induzida representam valores de natureza diferente. Para problemas de classificação, os nós folha retém um valor categórico, e as árvores são conhecidas como Árvores de Decisão, por exemplo, ID3 [Quinlan, 1979], ASSISTANT [Cestnik et al., 1987], C4.5 [Quinlan, 1993] e CART [Breiman et al., 1984]. Os estudos sobre modelos baseados em árvores de regressão iniciaram com os trabalhos de [Morgan e Sonquist, 1963]. Nestas árvores as folhas armazenam valores escalares ou um modelo linear e são conhecidas como Árvores de Regressão ou Árvores de Modelos, respectivamente. Podem-se destacar os trabalhos de [Breiman et al., 1984; Karalič e Cestnik, 1991; Quinlan, 1992; Karalic, 1992]. As árvores são definidas como:

- Árvore de Regressão: onde uma árvore é induzida e em suas folhas residem valores reais, que são as estimativas dos valores para os dados usados para se caminhar na árvore;
- Árvore de Modelos: onde uma árvore é induzida e em suas folhas residem modelos, em geral, lineares de regressão, no qual são aplicados os dados usados para se caminhar na árvore e se obter o valor real predito.

Uma árvore de regressão é similar a uma árvore de decisão, sendo que o primeira técnica é usada para problemas de regressão e a segunda para problemas de classificação. Os algoritmos são muito similares em vários aspectos, mas possuem em a mesma metodologia de construção. Uma das grandes vantagens nas técnicas de árvores de regressão é a capacidade de dividir o espaço de instâncias em subespaços e cada subespaço é analisado para gerar um modelo específico.

A Figura 2.8 ilustra os nós de uma árvore de regressão, onde os nós retangulares representam nós de decisão e os demais nós folha.

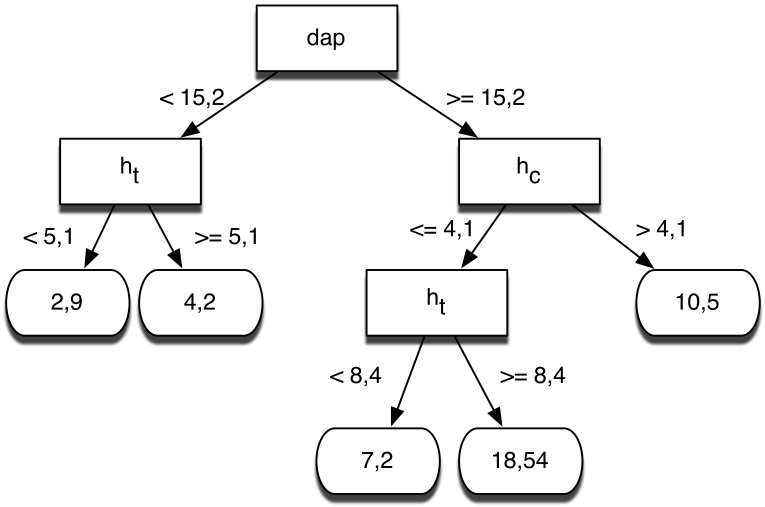


Figura 2.8: Exemplo de uma Árvore de Regressão, onde os nós internos (retangulares) são nós de decisão e os nós folha possuem um valor que deve ser devolvido como predição.

Fonte: O autor.

O algoritmo para construir árvores de regressão é simples. Escolhe-se um atributo que maximiza algum critério específico e são criadas duas partições de dados. Nesta divisão, o algoritmo é invocado recursivamente em cada partição conseguida. O algoritmo termina através de algum critério de parada estabelecido. O Algoritmo 1 mostra os passos para geração de uma árvore de regressão, chamado de $GeraArvore(D)$.

Em algoritmos de árvore de regressão, a função de custo a ser minimizada é usualmente o Erro Quadrático.

Construir a árvore de decisão minimal é um problema NP-Completo [Rivest, 1987] e, portanto, os algoritmos usam heurísticas para sua construção. Em geral executam uma pesquisa verificando um passo à frente, bem como a técnica de *hill-climbing* sem *backtracking*. Estas técnicas permitem a criação da árvore em tempo linear ao número de dados de treinamento.

Árvores de modelos são árvores construídas a partir da técnica desenvolvida por Quinlan [Quinlan, 1992], que combina regressão e indução de árvores. Assim como as árvores de regressão, as árvores de modelos predizem um valor numérico para uma determinada entrada, mas em suas folhas são armazenadas equações de regressão linear.

Algoritmo 1 Algoritmo de Indução de Árvores

if critério de parada(D) = Verdadeiro **then**
  return Um nó folha rotulado com a constante que minimiza a função perda
end if
Escolha o atributo que maximiza o critério de divisão em D
for all partição dos exemplos D_i baseado nos valores do atributo escolhido **do**
  Induz uma subárvore $Arvore = GeraArvore(D_i)$
end for
return Árvore contendo um nó de decisão baseado no atributo escolhido, e descendentes $Arvore_i$

Um critério de divisão é usado para escolher qual atributo é melhor para dividir uma porção da base de treinamento T. O desvio padrão é tratado como medida do erro naquele nodo e cada atributo é testado pelo cálculo da redução esperada do erro. Para divisão dos nós, o atributo escolhido é aquele que maximiza a redução esperada de erro.

A redução do desvio padrão (SDR), que é a redução do erro esperada, é calculada pela equação a seguir.

$$SDR = sd(T) - \sum \frac{|T_i|}{T} sd(T_i) \tag{2.34}$$

onde T_i corresponde a T_1, T_2, etc, que são conjuntos resultantes da divisão do nodo, de acordo com o atributo escolhido.

As funções lineares nas folhas são partes que, quando combinadas, formam uma função não-linear. O processo de divisão termina quando o desvio padrão é um pouco menor que o desvio padrão do conjunto original

A Figura 2.9 ilustra os nós de uma árvore de modelos. O nós folha possuem uma função, um modelo, que ao ser atigindo deve ser aplicado aos valores usados para se chegar até este nó. A aplicação deste modelo resulta em um valor real que é o resultado predito pela árvore para os valores de entrada.

Random Forest

Floresta Aleatória (RF do inglês *Random Forest*) é uma técnica de AM que agrupa várias árvores de modelos treinadas a partir de um conjunto de dados para predição, de forma que o resultado final é obtido através da consolidação dos resultados das árvores. Para treinamento das árvores, conjuntos diferentes de dados são usados. Para o teste, os dados de entradas são inseridos em todas as árvores geradas e o resultado final é a consolidação dos resultados de cada predição [Breiman et al., 1984; Breiman, 2001].

A técnica utilizada para consolidar o resultado da floresta depende do tipo de problema sendo tratado. Se o problema for de regressão, a média entre os resultados das árvores é computado e dado como saída da floresta. Em caso de classificação, o resultado final é obtido através de votação dos resultados de cada árvore da floresta. Neste trabalho o foco é em problemas de regressão e, portanto, a consolidação dos resultados é feita através da média.

Em AM, o conceito de combinação de classificadores (ou regressores) criados com amostras diferenciadas da base de treinamento e cujo resultado é obtido através da consolidação dos resultados de cada classificador (regressor) é conhecido como agregação de amostras (do inglês *bootstrap aggregating* ou *bagging*).

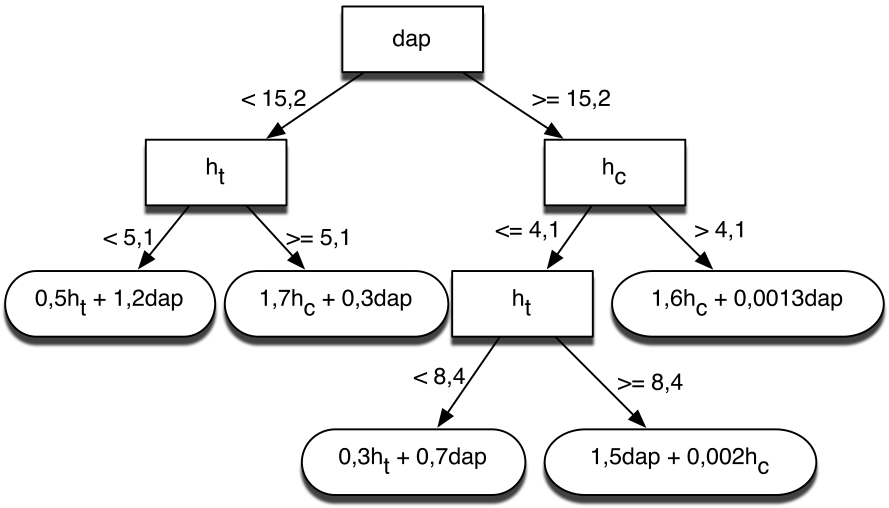


Figura 2.9: Exemplo de uma Árvore de Modelos, onde os nós internos (retangulares) são nós de decisão e os nós folha possuem um modelo linear referente aos dados de entrada.

Fonte: O autor.

Dado um conjunto de treinamento $X = x_1, x_2, ..., x_n$ com respostas $Y = y_1, y_2, ..., y_n$, *bagging* seleciona uma amostra aleatória do conjunto de treinamento, B vezes, e ajusta árvores com estas amostras.

Algoritmo 2 Algoritmo de Treinamento: *bagging*

for all $b = 1, ..., B$ **do**
  Obtenha n amostras de X e Y, chamados X_b e Y_b
  Treina uma árvore de regressão f_b sobre X_b e Y_b
end for

Cada árvore usada no *Random Forest* é uma árvore de regressão treinada considerando um número parametrizável de atributos, tomados de forma aleatória em cada treinamento, para sua construção. Esta árvore é conhecida como *Random Tree* [Breiman, 2001].

Após o treinamento da floresta, as predições $\hat{f}$ de valores não usados x' são feitas através da Equação 2.35, que para regressão é a média dos valores obtidos em cada árvore da floresta.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x') \tag{2.35}$$

A Figura 2.10 ilustra o funcionamento de uma floresta aleatória, onde os valores de entrada são inseridos em todas as árvores da floresta. Cada árvore é gerada de forma aleatória e resulta em um valor real predito diferente. Os valores obtidos são então consolidados através da média, que é o valor final predito pela floresta.

O intuito é diminuir a variância do conjunto sem aumentar o viés, levando a um melhor modelo do que se fossem usadas árvores isoladas. Isso significa que enquanto predições usando-se somente uma árvore são bastante sensíveis a ruído no conjunto de treinamento, a média de várias árvores não é, por causa da não correlação entre elas.

Além de se obter amostras do conjunto de treinamento, o algoritmo de Floresta Aleatória também obtém amostras do conjunto de características, para a criação de cada árvore. Este

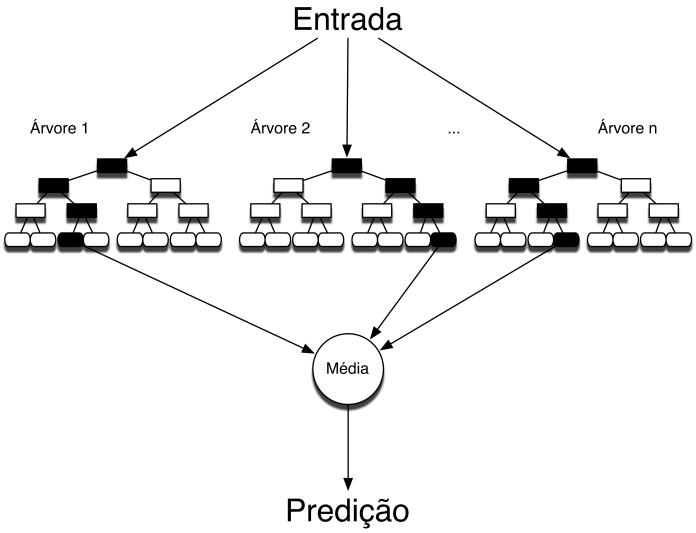


Figura 2.10: Esquema de funcionamento de uma Floresta Aleatória. A entrada é apresentada a todas as árvores da floresta e o resultado da predição é a média dos resultados de cada árvore.

Fonte: O autor.

processo é chamado de *bagging* de características (do inglês *feature bagging*). O objetivo é diminuir a correlação entre as árvores geradas e para árvores de regressão, sendo p o número de características, [Hastie et al., 2003] recomenda $p/3$ como número de características em cada árvore.

2.3 Treinamento e Avaliação de Modelos

A estimativa de acurácia de um classificador ou regressor é importante não só para predizer seu futuro poder de estimação, mas também para que possa ser feita uma escolha entre vários modelos, ou até uma combinação entre eles [Zhang, 1992; Wolpert, 1992].

Quando um modelo de classificação/regressão é construído utilizando-se uma base de treinamento, qualquer estimativa de desempenho ou acurácia será otimista se for aplicada a esta mesma base. Assim, para um resultado mais realista, deve-se aplicar o modelo treinado a um conjunto de instâncias que não foi utilizado na construção do modelo. Estas instâncias constituirão a base de teste.

Exitem várias formas de se particionar uma base de dados para se compor um conjunto de treinamento e de testes. Neste trabalho são detalhados o método de particionamento e da validação cruzada, apresentados a seguir.

2.3.1 Particionamento

O método de particionamento (do inglês *holdout*) [Theodoridis e Koutroumbas, 2008] é a maneira mais simples de se testar a acurácia de um regressor ou classificador. A base de dados é dividida em dois conjuntos disjuntos, um conjunto de treinamento e um conjunto de teste. Sobre o conjunto de treinamento é feita a geração do modelo e sobre o conjunto de teste é feita a avaliação deste modelo treinado.

A escolha, mesmo que aleatória, dos conjuntos de treinamento e teste pode levar à alta variância nos resultados, o que inviabiliza este processo em bases de dados pequenas. Não há

padrão na determinação do tamanho das bases, mas usam-se, geralmente, 70% para treinamento e 30% para testes, ou 80% para treinamento e 20% para testes [Witten e Frank, 2005].

2.3.2 Validação Cruzada

Na validação cruzada (*cross-validation*) [Duda et al., 2000; Kohavi, 1995], os dados são separados em k grupos (*k-fold cross-validation*), para os quais um destes grupos é deixado para teste e os demais para treinamento.

A cada passo, um grupo é tomado para teste e os $k - 1$ restantes são usados para treinamento. O processo se repete k vezes, até que todos os grupos tenham sido usados uma vez para teste. Em cada iteração, um modelo é treinado com os grupos de treinamento e testado com o grupo de teste. Ao final, calcula-se a média das correlações, que são usadas aqui como medida de qualidade.

O modelo final, no entanto, é treinado com todos os dados da base, mas sua medida de qualidade foi obtida através da validação cruzada. O Algoritmo 3 detalha estes passos. Deve-se atentar ao fato de que o modelo gerado é treinado com todos os elementos da base, mas a medida de acurácia (correlação neste caso) é dada pela validação cruzada.

Algoritmo 3 Algoritmo para validação cruzada de k grupos

Arranjar a base com m elementos de forma aleatória
Dividir a base em k grupos (k partes de aproximadamente m/k indivíduos)
for all $i = 1, ..., k$ **do**
  Treinar o modelo com todos os elementos que não pertencem ao grupo i
  Testar o modelo gerado com todos os elementos do grupo i
  Computar a medida de acurácia r_i, no caso deste trabalho a correlação
end for
Computar a acurácia total: $r = \frac{\sum_{i=1}^{k} r_i}{m}$
Treinar um modelo M com todos os m dados da base
return M e r

Em geral, a partição em 10 partes (também conhecida em inglês como *10-fold cross-validation*) é usada, mesmo que haja poder computacional para uma separação em mais partes [Kohavi, 1995].

A grande vantagem deste método é que todos os dados são usados tanto para treinamento como para teste, o que o torna indicado para os casos em que a base de dados é menor.

2.4 Ferramentas

Nesta seção são descritas a ferramenta de mineração de dados WEKA e o ambiente estatístico R.

2.4.1 WEKA

O *Waikato Environment for Knowledge Analysis* (WEKA) é um software gratuito e de código aberto, sob a licença *GNU General Public Licence* (GPL), desenvolvido pela Universidade

de Waikato (Nova Zelândia)[1]. O WEKA é um conjunto de implementações de algoritmos para várias tarefas de mineração de dados [Witten e Frank, 2005].

O software foi desenvolvido em linguagem Java e contém uma interface ao usuário para interagir com arquivos de dados e produzir resultados visuais. Também é disponibilizada uma interface de programação (API) geral, tornando possível incorporar o WEKA a aplicativos personalizados

O WEKA possui um formato de arquivos de entrada de dados, conhecido como ARFF. Nestes arquivos são em formato texto e descrevem todas as informações das bases de dados, tanto informações numéricas como categóricas. Apesar de ser o formato padrão de arquivo, o WEKA também trabalha com arquivos CSV (*comma-separated values*).

Dentre os algoritmos implementados no WEKA, destacam-se os de classificação, mas estão implementados também algoritmos de agrupamento, descoberta de regras de associação, entre outros.

2.4.2 R

R[2] é um ambiente de desenvolvimento e uma linguagem usada para cálculos estatísticos e análise de dados de forma gráfica. Foi criado na Universidade de Auckland, Nova Zelândia e está disponível sob a licença GPL e as versões binárias estão disponíveis para as principais plataformas, como Windows, Java e Linux.

Uma das principais características do R é sua alta capacidade de expansão, baseada no uso dos pacotes, bibliotecas que implementam funções específicas para diversas áreas de estudo. Um conjunto básico de pacotes é incluído inicialmente com a instalação de R, mas muitos outros estão disponíveis na rede de distribuição do R.

R é largamente utilizado entre estatísticos e analistas de dados para desenvolver aplicativos de diversas áreas da estatística e análise de dados.

2.5 Uso de Aprendizado de Máquina na Mensuração Florestal

Técnicas de AM são ferramentas poderosas para classificação de elementos, predição de valores, extração de regras, entre outros. Pela sua generalidade, podem ser usadas em várias áreas do conhecimento e conseguem ótimos resultados, muitas vezes melhores do que os obtidos com técnicas tradicionais, como pode ser observado em [da Silva Binoti, 2012; Wojciechowski, 2015; Sanquetta et al., 2015].

Na área florestal, em especial na mensuração florestal, os objetivos principais são fazer predições de informações que seriam muito custosas, ou inváveis, se o método tradicional (que muitas vezes envolve o abate de árvores) fosse aplicado à toda a população. As implicações deste custo vão desde o tempo gasto com a operação até o custo financeiro da derrubada das árvores, inviabilizando o negócio florestal ou gerando populações insustentáveis.

Como objeto desta tese, usa-se a relação hipsométrica, predição de volumes e estimativa de biomassa como estudo de caso para a experimentação das técnicas e comparações. Sendo assim, a seguir são apresentados exemplos de aplicações de técnicas de AM na área da mensuração florestal.

[1]Disponível em `http://www.cs.waikato.ac.nz/ml/weka/`

[2]Disponível em `http://www.r-project.org/`

As redes neurais artificiais são as ferramentas mais conhecidas e isso se deve à história da inteligência artificial, quando um dos objetivos era simular um cérebro humano no computador. Sendo assim, o potencial das RNAs já é explorado em várias áreas desde a sua criação, e não é diferente na área florestal.

A forte utilização das RNAs e consequente divulgação de resultados relevantes iniciou em 1991 com Guan e Gertner [Guan e Gertner, 1991], onde foram usadas redes neurais para estimar o crescimento de *Pinus resinosa*, usando como dado de entrada a medida de *dap* das árvores.

Em 1997, Schmoldt e Abbott [Schmoldt et al., 1997] apresentaram ótimos resultados no uso de redes neurais para classificação de defeitos internos da árvore (espaços vazios) através de imagens. Blackard e Dean [Blackard e Dean, 1999] obtiveram resultados comparáveis aos tradicionais na predição de tipos de cobertura de florestas, através de variáveis cartográficas.

Em 2000, Zhang, Hebda e Zhang [Zhang et al., 2000] modelaram a relação entre o crescimento de anéis das árvores através de dados climáticos. Leduc e outros [Leduc e Station, 2001] compararam o uso de RNAs e a função de distribuição de probabilidade de Weibull [Weibull, 1951] na aproximação de distribuição diamétrica de uma população de *Pinus palustris Mill.* ao sul dos Estados Unidos. Este estudo mostrou que as redes neurais foram superiores à distribuição Weibull, principalmente porque esta função não adere muito bem aos dados testados pelos autores.

Liu e outros [Liu et al., 2003] usaram RNAs, KNN (k-vizinhos mais próximos) e métodos estatísticos tradicionais para classificar dados florestais da FIA (*Forest Inventory and Analysis*). Neste estudo, as técnicas usando RNAs e KNN foram superiores aos tradicionais métodos estatísticos usados. Em 2004, Corne e outros [Corne et al., 2004] modelaram seis tipos de RNAs para predizer dados florestais de áreas não sobrevoadas no Alaska, usando dados topográficos como entrada. Os resultados foram similares aos obtidos através de análise de imagens por satélite, em uma determinada área de comparação.

Para estimativa de volumes, Diamantopoulou [Diamantopoulou, 2005] demonstrou a superioridade das RNAs na estimativa do volume da casca de *Pinus brutia*. Foram usadas 270 árvores de uma floresta de Thessaloniki, Grécia. Para treinamento da RNA, os dados foram particionados aleatoriamente em 70% para treinamento e 30% para teste. Como resultado, as RNAs tiveram um resultado superior aos modelos de regressão testados pela autora e melhoraram em 7,28% o erro na predição do volume da casca.

Em 2006, Avramidis e outros [Avramidis et al., 2006] usaram RNAs para predizer propriedades dielétricas de madeira, baseando-se nas condições elétricas e termais do ambinente e da estrutura química básica da madeira. Nestes experimentos, eles apresentaram RNAs com resultados de R^2 de $0,9945$ enquanto o processo de regressão resultou um R^2 de $0,8685$, indicando claramente que as RNAs têm potencial nesta área de pesquisa, mas necessitando maiores experimentos, conforme apontado pelo autores.

Ainda em 2006, Schoeninger em sua tese de doutorado [Schoeninger, 2006] utilizou imagens de satélite para obter mapas temáticos para estimativa de biomassa arbórea e quantidade de carbono orgânico armazenado em uma floresta ombrófila densa. Neste trabalho, as RNAs se mostraram mais exatas nas estimativas (com erros entre 3% e 4%), que os clássicos modelos de regressão (com erros entre 29% e 30%).

Schoemaker e Cropper [Shoemaker e Cropper Jr, 2008] compararam os resultados de predição do Índice de Área de Folhas (LAI - *Leaf Area Index*) usando métodos de regressão e redes neurais, usando informações de sensoriamento remoto. Muitas variáveis de entrada foram consideradas, tais como aplicação de herbicida, espécie da árvore, se a a área das folhas está em expansão, e três métodos foram testados: modelos lineares, modelos de regressão múltipla e

RNAs. Nos resultados, modelos lineares não tiveram R^2 maior que 0, 12, os modelos de regressão múltipla tiveram R^2 entre 0, 31 e 0, 70 o as RNAs valores entre 0, 40 e 0, 85, que foram o melhor método.

Também na área de sensoriamento remoto, Schoeninger e outros [Schoeninger et al., 2008a] mostraram que as RNAs são uma alternativa viável para o mapeamento de biomassa e carbono em grandes extensões florestais. Neste trabalho foi definido um método para mapeamento dos dados de entrada para a aplicação das RNAs. Os mesmos autores também fizeram este mapeamento para uma floresta ombrófila densa, sendo que as RNAs mostraram-se mais precisas na estimativa de biomassa e quantidade de carbono, do que os modelos de regressão linear ajustados [Schoeninger et al., 2008b].

Görgens e outros [Gorgens et al., 2009] usaram várias configurações de RNAs para estimar volumes de *Eucalyptos* sp., com dados de quatro empresas diferentes e dados de *Tectona grandis L.f.*, num total de cinco locais diferentes. Os resultados foram comparados com modelos de Schumacher e Hall [Schumacher e Hall, 1934] para cada local. O objetivo era usar somente uma rede para estimar o volume de todos os locais, o que se comprovou através dos resultados obtidos, não havendo diferença estatística entre os resultados das redes e dos modelos tradicionais.

Seguindo a linha de Görgens, Silva [da Silva et al., 2009] também avalaram modelos volumétricos de Schumacher e Hall de 21 cubagens de povoamentos de eucalipto (total de 862 árvores), usando como entradas o diâmetro na altura do peito e altura total. Foram treinadas várias RNAs em várias configurações, sendo indicadas duas arquiteturas que obtiveram melhores estimativas. Concluiu-se neste trabalho que as RNAs são recomendadas para a previsão de volumes, pois seu aspecto generalista permite o treinamento de uma rede para representar vários clones da mesma espécie, o que não ocorre com os modelos tradicionais, onde geralmente uma equação é ajustada para cada clone.

Com este histórico de resultados da abordagem baseada em RNAs, surgiram vários trabalhos na Universidade Federal de Viçosa, que culminaram em dissertações de mestrado e teses de doutorado, ratificando e consolidando os resultados obtidos. Em especial, a dissertação de mestrado de Binoti [da Silva Binoti, 2010] e sua tese de doutorado [da Silva Binoti, 2012], apresentaram vários experimentos de utilização de RNAs na mensuração e manejo florestal. Nestes trabalhos foram modelados os problemas de estimação de altura, redução no número de árvores a serem cubadas, projeção de parâmetros da função de distribuição de Weibull e validação de um software de simulação de redes neurais, o NeuroForest [3].

Não só no Brasil estas pesquisas estavam sendo feitas, Diamantopoulou [Diamantopoulou e Özçelik, 2012] fizeram experimentos de predição de altura total de árvores em florestas da Turquia. Neste trabalho, os modelos baseados em RNAs de regressão generalizadas [4] [Patterson, 1996] foram superiores aos modelos não lineares de regressão testados, para todas as espécies. As GRNNs não precisam de um procedimento interativo de treinamento, são baseadas em *kernels* e usam técnicas bayseanas para aproximar uma função entre um vetor de entrada e um vetor de saída.

Castro [Castro et al., 2013] usaram modelos de RNAs para estimar altura, diâmetro e mortalidade de eucaliptos no norte do Brasil. Várias configurações de redes foram testadas com diferentes parâmetros de entrada. Os modelos foram confrontados com as equações fornecidas pela empresa florestal local. Os níveis de erro em medições de árvores em pé foram de aproximadamente 0, 5% para as redes neurais e 6% para os modelos de regressão.

Binoti [Binoti et al., 2014b; da Silva Binoti et al., 2014] e Görgens [Gorgens et al., 2014] fizeram vários estudos sobre as melhores configurações de cama-

[3]Disponível em `http://neuroforest.ucoz.com`.

[4]Do inglês GRNN - *Generalized Regression Neural Networks*

das e números de neurônios para estimativa de volumes. Os resultados obtidos melhoram as metodologias de teste de RNAs, limitando o número de configurações que devem ser testadas na modelagem volumétrica.

Mais recentemente, [da Silva Binoti et al., 2015] modelaram várias RNAs para projeção de área basal e volumes, em clones de eucalipto no sul da Bahia. Os erros de estimativa foram da ordem de $12,5\%$, o que foi citado como muito satisfatório, pelos autores. Castro e outros [Castro et al., 2015] estudaram a mortalidade e probabilidade de mortalidade em um fragmento de floresta semidecidual, obtendo como resultado uma configuração de perceptron multi-camada com função de ativação exponencial como a de melhor estimativa.

Além das RNAs, outras técnicas de AM vêm sendo utilizadas com muito sucesso na mensuração florestal. Em especial, KNN foi utilizado para estimativa de estoque de biomassa de *Araucaria angustifolia* no Brasil [Sanquetta et al., 2013]. Os resultados obtidos com o uso do KNN não tiveram diferença estatística dos modelos tradicionais testados, o que mostrou que a área de estudo era promissora.

Wojciechowski [Wojciechowski, 2015] em sua tese de doutorado desenvolveu o JCarbon[5], que é um software para estimativa de volumes, biomassa e carbono usando modelos tradicionais e KNN. Seu principal objetivo é centralizar os dados de diversos locais, oferecendo ferramenta para a rápida execução de cálculos baseados em modelos alométricos e KNN. Na metodologia apresentada, uma base de dados é carregada e usada como referência (instâncias). Quando uma predição for efetuada, n-vizinhos das instâncias são visitados para determinar o dado a ser calculado (volume, biomassa ou carbono). Se mais de um vizinho for usado na predição, é feita a ponderação pelo inverso da distância.

Em [Sanquetta et al., 2015] o método proposto por Wojciechowski foi aplicado para predição de biomassa em 180 árvores de Floresta Atlântica, que demonstrou melhores resultados que os modelos alométricos. O melhor resultado foi através do uso de cinco vizinhos e distância Chebishev, que resultou em um ganho de $16,5\%$ na redução de erros de estimativa.

Schikowski [Schikowski, 2016] em sua dissertação de mestrado comparou técnicas de AM com modelos alométricos para predição de volumes e funções de afilamento para *Acacia mearnsii* De Wild. Foram empregados os algoritmos de K-NN, *Random Forest* e RNA, sendo que estes modelos foram mais acurados que o modelo volumétrico de Schumacher & Hall e o polinônio de Hradetzky.

Como pode-se observar através destas aplicações, cada vez mais as técnicas de AM têm se consolidado como uma ferramenta na área da Mensuração Florestal. Todos estes estudos apresentam alternativas ao tradicional método de modelagem através de modelos de regressão, apresentando aos engenheiros novas formas de extrair informações de suas bases de dados e com resultados, muitas vezes, melhores.

[5]Disponível em `www.jcarbon.ufpr.br`.

Capítulo 3

Materiais e Métodos

Neste capítulo são apresentados os dados analisados e os métodos de mensuração florestal que foram objeto deste estudo. Também são mostradas as técnicas de aprendizado de máquina a serem aplicadas e as estatísticas para medição da qualidade dos modelos obtidos e verificação da existência de diferença estatística entre os modelos.

3.1 Dados Analisados

Neste trabalho, três domínios de bases de dados foram usados para cinco estimativas de Relação Hipsométrica, Volumes, Biomassa, a saber: relação hipsométrica e predição de volumes de *Pinus taeda*, relação hipsométrica e avaliação de biomassa de acácia-negra e avaliação de biomassa seca acima do solo de árvores de florestas tropicais. As características das bases de dados utilizadas estão descritas a seguir.

3.1.1 *Pinus taeda*

O plantio de pinus se tornou viável no Brasil por conta da introdução de várias espécies e, sendo assim, se tornando uma importante fonte de madeira para vários segmentos, como energia, celulose, compensados, entre outros [Shimizu e Medrado, 2005]. O Brasil possui uma grande diversidade de condições ambientais, o que propiciou uma variabilidade genética alta e uma adaptação do gênero a várias condições ecológicas [Shimizu, 2006].

Dentre as espécies de pinus mais cultivadas no país está o *Pinus taeda*, que se destaca pelo elevado incremento volumétrico nas regiões mais frias do Brasil, além de possuir menor taxa de resina. Assim, esta é uma das espécies mais plantadas no sul do Brasil, através de programas de reflorestamento incentivados pelo governo.

O *Pinus taeda* é uma das principais espécies florestais cultivadas no sul dos Estados Unidos, cuja área de distribuição natural atinge uma grande extensão.

Segundo Hocker [Hocker, 1956], fatores climáticos como temperatura média, intensidade e frequência das precipitações, tornam a espécie propícia para o desenvolvimento natural. Onde estes requisitos não se apresentam, a espécie não ocorre naturalmente e nem se desenvolve satisfatoriamente quando plantada.

Como a área de distribuição natural é grande, com diversas condições climáticas, formam-se as raças geográficas. Vários ensaios foram feitos sobre locais de plantio ao sul e sudeste dos Estados Unidos, tanto em regiões montanhosas como em regiões costeiras [Wells e Wakeley, 1996; Kraus, 1967; La Farge, 1974; Grisby, 1977].

No Brasil foram observadas tendências volumétricas e de crescimentos semelhantes aos estudos citados, como apresentado nos ensaios Boletim de Pesquisa Florestal, Colombo, n. 2, p. 1-25, Jun. 1981. instalados em Rio Negro, PR [Baldanzi e Araujo, 1971], Telêmaco Borba, PR [Barrichelo et al., 1978], Lages, SC [da Fonseca et al., 1978], Capão Bonito, SP, Irati, PR, Três Barras, SC e Pelotas, RS [Araujo, 1980].

Segundo Marchiori [Marchiori, 1996] árvores do gênero Pinus podem alcançar 20 m de altura e 100 cm de diâmetro à altura do peito (*dap*), com copa densa, casca gretada e ramos acinzentados, sendo que a madeira é indicada para construções, móveis e caixotaria.

Na silvicultura preferem-se as espécies do gênero Pinus pelas seguintes razões [Lamprecht, 1990]:

- Possui várias espécies e, entre elas, sempre há uma que se adapta a um determinado sítio;
- Muitas adaptam-se em solos que naturalmente são pobres e secos e também em sítios degradados;
- Os incrementos em volume de algumas espécies, geralmente, são de alto a muito alto, mesmo em condições ambientais desfavoráveis;
- São muito apropriadas para reflorestamentos e para plantios com um manejo esquemático simples (monocultivo/corte raso);
- A madeira das coníferas é, por natureza, uma matéria-prima em escassez nos trópicos e os Pinus têm a capacidade de produzí-la em grande quantidade e com qualidade uniforme, a qual é necessária para a produção de polpa, papel, painéis, entre outros.

Para predição de volumes e relação hipsométrica foram utilizadas 302 amostras de *Pinus taeda* provenientes do acervo de uma empresa florestal da região central do Paraná. Para estas amostras foi feita uma cubagem rigorosa usando o método de Smalian [Azevedo Gomes, 1957]. Tanto para as técnicas de AM como para os modelos alométricos, optou-se pelo método de validação cruzada, por se tratar de uma base de dados com poucos indivíduos. Esta mesma base de dados foi usada para os testes com relações hipsométricas, desconsiderando-se o dado volume.

Os dados disponíveis nesta base são apresentados na Tabela 3.1.

Tabela 3.1: Descrição da Base de Dados de Pinus

Dado	Valor Mínimo	Valor Máximo	Média	Desvio Padrão
Idade ($anos$)	4,063	19,236	11,696	4,741
Diâmetro à altura do peito (cm)	5	45	24,953	11,436
Altura (m)	4,3	31	11,68	7,155
Volume (m^3)	0,006	2,057	0,586	0,562

3.1.2 Acácia-negra

A acácia-negra (*Acacia mearnsii*) é uma leguminosa arbórea, originária da Austrália, que foi introduzida no Rio Grande do Sul em 1918, por Alexandre Bleckmann, sendo estabelecida comercialmente em 1926 por Júlio C. Lohmann [Oliveira, 1968]. É usada para vários propósitos, tais como restauração de ambientes degradados, fixação de nitrogênio, produção de tanino e de energia, dentre outros.

Também é considerada uma espécie recuperadora de solos de baixa fertilidade, pois possui características específicas que melhoram as condições do solo [Carpanezzi, 1998]. Tem grande importância econômica, pois praticamente 60% das plantações são de pequenos proprietários, gerando uma grande quantidade de empregos diretos. A idade de corte no Brasil é em torno de 5,5 até 10 anos. Sua produtividade gira em torno de 10 a 25 $m^3.ha^{-1}.ano^{-1}$ e produção média de casca em torno de 15 $t.ano^{-1}$.

A rentabilidade do cultivo da acácia-negra é superior ao de muitas espécies, embora o rendimento quantitativo da madeira seja inferior. Assim, a grande rentabilidade deve-se à comercialização da casca, que representa o objetivo principal do cultivo desta espécie. Também utiliza-se a madeira para a fabricação de papel, chapas de aglomerados e lenha para a queima em fornos e fabricação de carvão.

Da casta é extraído o tanino, que é utilizado nas indústrias farmacêutica e coureira, entre outras. Até 1954 o Brasil importava grande quantidade de extratos vegetais curientes, então passou a ser auto-suficiente no produto e atualmente exporta seus excedentes, participando ativamente no mercado mundial.

A acacicultura no Rio Grande do Sul é uma atividade econômica sólida, que tem trazido vários benefícios e prosperidade para vários municípios gaúchos. Com a expansão dessa fonte de riqueza, permitiu-se aproveitar melhor áreas que antes eram pouco aproveitadas, justificando os vastos estudos feitos para se quantificar e projetar seus subprodutos.

Para os experimentos com avaliação de biomassa seca, foram utilizadas 544 amostras de acácia-negra de plantação comercial no estado do Rio Grande do Sul [Behling, 2014]. Para ajuste de equações alométricas utilizou-se validação cruzada, o mesmo usado para as técnicas de AM. Esta base de dados foi utilizada para obtenção de relação hipsométrica, desconsiderando-se o dado biomassa.

Os dados disponíveis nesta base são apresentados na Tabela 3.2 e na Tabela 3.3.

Tabela 3.2: Descrição da Base de Dados de Acácia-negra: dados categóricos

Dado Categórico	Quantidade
Local (categórico)	3
Fazenda (categórico)	9

Tabela 3.3: Descrição da Base de Dados de Acácia-negra: dados numéricos

Dado	Valor Mínimo	Valor Máximo	Média	Desvio Padrão
Idade ($anos$)	1	10	4,833	3,42
Diâmetro à altura do peito (cm)	0,637	23,555	9,525	4,66
Altura (m)	1,6	21,9	12,216	5,169
Altura da Copa (cm)	0	15,6	5,916	2,488
Biomassa (kg)	0,311	361,066	47,657	50,499

3.1.3 Biomassa de Florestas Tropicais

As florestas são consideradas importantes reservatórios globais de carbono, armazenando cerca de $296Gt$ de carbono. As concentrações de carbono encontram-se nas florestas tropicais da América do Sul e África Central, que estocam cerca de $120tC.ha^{-1}$, enquanto a média mundial é de $75tC.ha^{-1}$. Entretanto, as florestas tropicais têm sido as maiores vítimas do desflorestamento e da degradação [FAO – Food and Agriculture Organization, 2015]. Isso tem

ocasionado que as emissões acumuladas de Gases de Efeito Estufa (GEE) por atividades de uso do solo e florestas aumentassem de $490 \pm 180 GtCO_2$ em 1970 para $680 \pm 300 GtCO_2$ em 2010 [UNFCCC – United Nations Framework Convention on Climate Change, 2007].

A maior fração de carbono estocado nas florestas do mundo está na sua biomassa viva, com $250 GtC$ [FAO – Food and Agriculture Organization, 2015], e ainda é grande a incerteza com relação a esses estoques, principalmente devido à precariedade na estimação em grande escala. Um fator complicador é que qualquer modelagem aplicada em maior escala deve se embasar em determinações diretas, que são complexas, destrutivas e onerosas [Watzlawick et al., 2009].

É imprescindível desenvolver modelos precisos e acurados dos estoques de carbono em larga escala, mas isso não é uma tarefa simples. São diversas as variáveis que afetam os cálculos, como a composição e a estrutura da vegetação, informações específicas, como: a densidade ou massa específica dos tecidos, o teor de carbono nos tecidos, o método usado para o cálculo de áreas e a confiabilidade do inventário florestal, entre outras. Entre esses fatores, um dos mais importantes é a metodologia de modelagem empregada para estimar a biomassa ou o carbono individual a partir de variáveis dendrométricas, como o diâmetro e a altura das árvores.

Chave e outros [Chave et al., 2014] estimaram biomassa seca total acima do solo de uma grande quantidade de amostras de mais de 4.000 dados coletados em várias regiões dos trópicos, incluindo o Neotrópico, a África e o Sudeste Asiático. Os dados usados aqui foram cedidos por Jerome Chave, Diretor de Pesquisa no Centro Francês de Pesquisa Científica – CNRS, França. Correspondem a 4.004 observações de diâmetro à altura do peito (dap), altura total (ht), biomassa seca aérea total (b), densidade da madeira (massa específica básica, ρ), coletados nos seguintes países: Austrália, África do Sul, Brasil, Camboja, Camarões, República Centro-Africana, Colômbia, Costa Rica, Guiana Francesa, Gabão, Gana, Guadalupe, Índia, Indonésia, México, Madagáscar, Malásia, Moçambique, Nova Guiné, Peru, Porto Rico, Tanzânia, Venezuela e Zâmbia.

Os dados disponíveis nesta base são apresentados na Tabela 3.4 e na Tabela 3.5.

Tabela 3.4: Descrição da Base de Dados de Florestas Tropicais: dados categóricos

Dado Categórico	Quantidade
Local (categórico)	58

Tabela 3.5: Descrição da Base de Dados de Florestas Tropicais: dados numéricos

Dado	Valor Mínimo	Valor Máximo	Média	Desvio Padrão
Diâmetro à altura do peito (cm)	5	212	23,988	24,085
Altura (m)	1,2	70,7	16,038	10,772
Massa Específica (g/cm^3)	0,09	1,2	0,634	0,164
Biomassa (kg)	1,23	76063,52	1134,139	3917,972

3.2 Métodos de Mensuração Florestal

Para os experimentos realizados aqui usaram-se bases de dados para obtenção de relação hipsométrica, predição de volumes e avaliação de biomassa.

3.2.1 Relação Hipsométrica

Relação Hipsométrica é a relação que existe entre a altura e o diâmetro na altura do peito de uma árvore. Esta relação é importante pois, sendo o *dap* uma medida fácil de ser obtida e a altura uma medida difícil, demorada e mais onerosa, através da relação hipsométrica consegue-se estimar a altura de árvores através da simples observação do seu *dap*.

Vários modelos estão disponíveis [Caldeira et al., 2002] e na Tabela 3.6 são apontados alguns, dentre eles os que obtiveram melhores resultados em [Caldeira et al., 2002]. Pode-se observar modelos que usam somente o *dap* como variável de entrada e outros que, se estiver disponível, a Idade.

Tabela 3.6: Modelos Hipsométricos

Nome	Modelo	Coeficientes
Parabólico/Trorey	$h = \beta_0 + \beta_1 * dap + \beta_2 dap^2$	β_0, β_1, β_2
Stoffels	$\ln h = \beta_0 + \beta_1 \ln dap$	β_0, β_1
Curtis	$h = \beta_0 + \beta_1 \frac{1}{dap}$	β_0, β_1
Logarítmico	$\ln h = \beta_0 + \beta_1(\frac{1}{dap}) + \beta_2(\frac{1}{Idade}) + \beta_3(\frac{1}{dap \times Idade})$	β_0, β_1, β_2, β_3
Prodan	$h = \frac{dap^2}{\beta_0 + \beta_1 \ln dap + \beta_2 dap^2}$	β_0, β_1, β_2

Nos modelos onde as variáveis sofrem transformação logarítmica, deve-se fazer a correção das alturas pelo Fator de Correção de Meyer (Equação 3.1), para depois se fazer o recálculo do S_{yx}.

A análise da qualidade das equações ajustadas é feita calculando-se as medidas mostradas na Tabela 3.10.

3.2.2 Predição de Volumes

O volume da árvore é uma das informações mais importantes no inventário florestal, sendo o ponto de partida para a avaliação do conteúdo de madeira em povoamentos florestais.

Para obtenção do volume de madeira pode-se usar vários métodos, como através do xilômetro (deslocamento de água), métodos de cubagem rigorosa como o de Smalian, Huber e Newton [Azevedo Gomes, 1957], e métodos não destrutivos baseados em equações volumétricas de regressão. O método de cubagem pelo xilômetro é o único que fornece o volume verdadeiro, mas as equações volumétricas são as mais usadas por não exigirem o abate das árvores e produzirem resultados muito próximos do real.

As equações volumétricas são divididas em três grupos:

- Equações de simples entrada: o volume é estimado em função do *dap*;
- Equações de dupla entrada: o volume é estimado em função do *dap* e da altura total;
- Equações de tripla entrada: o volume é estimado em função do *dap*, altura total e de uma medida que expressa a forma da árvore.

Os modelos mais conhecidos são apresentados na Tabela 3.7 [Clutter, 1983], para os de entrada única e na Tabela 3.8 os de dupla entrada.
onde:

- v: volume estimado;

Tabela 3.7: Modelos de simples entrada

Autor	Modelo
Kopezky-Gehrhardt	$v = \beta_0 + \beta_1 dap^2 + \varepsilon$
Dissescu-Meyer	$v = \beta_1 dap + \beta_2 dap^2 + \varepsilon$
Hohenadl-Krenm	$v = \beta_0 + \beta_1 dap + \beta_2 dap^2 + \varepsilon$
Berkhout	$v = \beta_0 dap^{\beta_1} + \varepsilon$
Husch	$\ln v = \beta_0 + \beta_1 \ln dap + \varepsilon$
Brenac	$\ln v = \beta_0 + \beta_1 \ln dap + \beta_2 + \frac{1}{dap} + \varepsilon$

Tabela 3.8: Modelos de dupla entrada

Autor	Modelo
Spurr	$v = \beta_0 + \beta_1 dap^2 h_t + \varepsilon$
Spurr(log)	$\ln v = \beta_0 + \beta_1 \ln(dap^2 h_t) + \varepsilon$
Schumacher-Hall	$v = \beta_0 dap^{\beta_1} h_t^{\beta_2} + \varepsilon$
Schumacher-Hall(log)	$\ln v = \beta_0 + \beta_1 \ln dap + \beta_2 \ln h_t + \varepsilon$
Honner	$v = \frac{dap^2}{\beta_0 + \beta_1 \frac{1}{h_t}} + \varepsilon$
Ogaya	$v = dap^2(\beta_0 + \beta_1 h_t)$
Stoate	$v = \beta_0 + \beta_1 dap^2 + \beta_2 dap^2 h_t + \beta_3 h_t + \varepsilon$
Naslund	$v = \beta_1 dap^2 + \beta_2 dap^2 h_t + \beta_3 dap h_t^2 + \beta_4 h_t^2 + \varepsilon$
Takata	$v = \frac{dap^2 h_t}{\beta_0 + \beta_1 dap} + \varepsilon$
Meyer	$v = \beta_0 + \beta_1 dap + \beta_2 dap^2 + \beta_3 dap h_t + \beta_4 dap^2 h_t + \beta_5 h_t + \varepsilon$

- dap: diâmetro à altura do peito (1,30m);
- h_t: altura total, considerando a altura da ponta;
- $\beta_0, \beta_1, ..., \beta_n$: coeficientes a serem ajustados;
- ln: logaritmo neperiano;
- ε: erro da estimativa.

Nesta tese serão ajustados três modelos de regressão: Spurr, Husch e Schumacher & Hall. Estes modelos foram escolhidos por estarem entre os mais utilizados pela comunidade em modelagem volumétrica [Jorge, 1982; Fernandes et al., 1983; Silva e Carvalho, 1984; Queiroz, 1984; Higuchi e Ramm, 1985; Couto e Bastos, 1987; Souza e Jesus, 1991; Scolforo et al., 1994; Belchior, 1996; Baima et al., 2001; Chichorro et al., 2003; Schneider e Tonini, 2003; Batista et al., 2004], baseados em seus desempenhos observados.

Nos modelos onde as variáveis sofrem transformação logarítmica, deve-se fazer a correção dos volumes estimados pelo Fator de Correção de Meyer, FM, (Equação 3.1), para depois recalcular o S_{yx}.

$$FM = e^{0,5 S_{yx}^2} \tag{3.1}$$

onde:

- e: é o valor $2,718281828$;

- S_{yx}: é o erro padrão da estimativa.

A análise da qualidade das equações ajustadas é feito calculando-se as medidas mostradas na Tabela 3.10.

3.2.3 Predição de Biomassa

A biomassa é a quantidade de massa do material vegetal disponível em uma floresta [Martinelli et al., 1994], que pode ser expressa pela massa verde (material fresco amostrado contendo uma proporção variável de água) ou seca (obtida após secagem em estufa) [Caldeira, 2003].

Da mesma forma que o volume, a forma de medir a biomassa de forma confiável e com relativo baixo custo é por meio da construção de modelos empíricos que determinam a biomassa da floresta em pé a partir de seus atributos dendrométricos, com base numa amostra abatida composta apenas por um pequeno número de indivíduos [Maestri et al., 2005].

Os modelos mais usados podem ser vistos na Tabela 3.9.

Tabela 3.9: Modelos de Biomassa (b, em unidade de massa)

Nome	Modelo	Coeficientes
Berkhout	$b = \beta_0 + \beta_1 dap$	β_0, β_1
Kopezky & Gehrhardt	$b = \beta_0 + \beta_1 dap^2$	β_0, β_1
Spurr	$b = \beta_0 + \beta_1 dap^2 h_t$	β_0, β_1
Schumacher & Hall	$\ln b = \beta_0 + \beta_1 \ln dap + \beta_2 \ln h_t$	β_0, β_1, β_2
Husch	$\ln b = \beta_0 + \beta_1 \ln dap$	β_0, β_1
Stoate	$\ln b = \beta_0 + \beta_1 dap^2 h_t + \beta_2 dap^2 + \beta_3 h_t$	β_0, β_1, β_2, β_3

A análise da qualidade das equações ajustadas é feito calculando-se as mesmas medidas mostradas na Tabela 3.10.

Além dos modelos tradicionais, outros são apresentados por conta da particularidade dos dados levantados. Em especial, Chave e outros [Chave et al., 2014] estimaram biomassa seca total acima do solo de uma grande quantidade de amostras de mais de 4.000 dados coletados em várias regiões dos trópicos, incluindo o Neotrópico, a África e o Sudeste Asiático. Seu melhor modelo alométrico para a base de dados analisada, a saber:

$$AGB_{est} = \beta_0 (ME dap^2 h_t)^{\beta_1} \tag{3.2}$$

e um modelo alternativo, restringindo o expoente em um:

$$AGB_{est} = \beta_0 (ME dap^2 h_t) \tag{3.3}$$

onde:

- AGB_{est}: biomassa acima do solo estimada;
- dap: diâmetro à altura do peito (1,30m);
- h_t: altura total;
- ME: massa específica;
- β_0, β_1: coeficientes a serem ajustados.

Estes modelos obtiveram resultados muito satisfatórios para os dados especícificos e, portanto, são usados como modelos alométricos de referência neste trabalho.

3.2.4 Avaliação da Qualidade dos Modelos

Para avaliação da qualidade dos modelos pode-se calcular e comparar o Coeficiente de Determinação, Coeficiente de Determinação Ajustado, Erro Padrão da Estimativa, Erro Padrão da Estimativa em Porcentagem, Índice de Ajuste de Schlaegel, Critério de Informação de Akaike e Análise de Resíduos. Esta última é feita por meio de análise gráfica, a Tabela 3.10 apresenta como estas medidas são calculadas.

Tabela 3.10: Medidas calculadas para avaliação da qualidade dos modelos

Nome	Abreviatura	Fórmula
Soma de Quadrados dos Resíduos	$SQRes$	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Coeficiente de Correlação de Pearson	ρ	$\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}}_i)}}$
Coeficiente de Determinação	R^2	$1 - \frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
Coeficiente de Determinação Ajustado	R^2_{aj}	$1 - \frac{(n-1)}{n-(p+1)}(1-R^2)$
Erro Padrão da Estimativa	S_{yx}	$\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}}$
Erro Padrão da Estimativa em %	S_{yx}%	$\frac{S_{yx}}{\bar{y}}100$
Índice de Ajuste de Schlaegel	IA	$1 - \frac{n-1}{n-p}\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
Critério de Informação de Akaike [Akaike, 1973]	AIC	Se $\frac{n}{p} < 40$ Então: $-2\left(\frac{-n}{2}\ln\left(\frac{1}{n}SQRes\right)\right) + 2k\frac{n}{n-p}$ Senão: $-2\left(\frac{-n}{2}\ln\left(\frac{1}{n}SQRes\right)\right) + 2p$
Critério de Informação Bayesiano [Schwarz, 1978]	BIC	$-2\left(\frac{-n}{2}\ln\left(SQRes\right)\right) + \ln(n)p$

onde:

- e: é o valor $2,718281828$;
- y_i: é o valor observado;
- $\bar{y}$: é o valor médio observado;
- $\hat{y}_i$: é o valor estimado para y_i;
- $\bar{\hat{y}}_i$: é o valor médio estimado;
- n: é o número de observações;
- p: é o número de coeficientes no modelo de regressão.

O coeficiente de determinação (R^2) expressa a quantidade de variação da variável dependente que é explicada pelas variáveis independentes. Quanto mais próximo do valor um,

melhor é o ajuste. O valor do R^2 é usado para comparar modelos que possuem número de coeficientes diferentes.

O erro padrão da estimativa (S_{yx}) é uma estatística que mede a dispersão média entre os valores observados e estimados ao longo da linha de regressão, sendo que, quanto mais baixo for o valor do S_{yx}%, melhor é o ajuste. Para os modelos onde a variável dependente sofreu transformação, é necessário recalcular o erro padrão residual, para comparar estatisticamente as equações.

O Critério de Informação de Akaike [Akaike, 1973] é um critério utilizado para a seleção de modelos de regressão linear muito utilizado [McQuarrie e Tsai, 1998] e representa a distância relativa esperada entre dois modelos probabilísticos. O Critério de Informação Bayesiano [Schwarz, 1978] também se baseia na verossimilhança dos modelos, penalizando complexidades (número de parâmetros).

Um resíduo é a diferença entre o valor estimado e o valor observado. Análise de resíduos é um conjunto de técnicas usadas para avaliar a adequação de um modelo de regressão baseado nos resíduos calculados. A análise gráfica é útil para se avaliar tendências nos resíduos, valores discrepantes (*outliers*), etc e neste trabalho são mostrados os gráficos de resíduos para as estimativas apresentadas.

3.3 Métodos de Aprendizado de Máquina

Nesta seção apresentam-se as técnicas de aprendizado a serem usadas nos experimentos realizados, bem como uma breve justificativa de sua escolha.

As RNAs foram escolhidas neste trabalho por já serem difundidas na área florestal, embora não de forma bem difundida, e por serem utilizadas em vários trabalhos, como pode-se observar no Capítulo 2. São modelos inspirados em uma rede de neurônios biológicos, separados em camadas, com um algoritmo de aprendizado de retro-alimentação bem definido. No caso deste trabalho, os parâmetros a serem variados são: a taxa de aprendizado, o momentum e a quantidade de neurônios na camada oculta. Em especial, são utilizadas as Redes MLP.

As SVMs são sistemas baseados em otimização matemática, derivado da aprendizagem estatística. Foram escolhidas por apresentarem boa capacidade de generalização, robustez em bases com muitos dados e por possuírem uma base teórica bem definida. Outra característica interessante é o pequeno número de parâmetros a serem ajustados para se obter um modelo, no caso deste trabalho serão usados o custo (C) e *gamma*. Em especial, é utilizado o algoritmo SMO.

O *Random Forest* foi escolhido pela capacidade de representar regiões diversas dos dados, através de seu algoritmo de indução. Também pelo fato de seu treinamento ser extremamente rápido, espera-se que em bases de dados com alta variação dos dados, este modelo se comporte melhor. Este método também possui poucos parâmetros para ajuste, que no caso deste trabalho são a quantidade de árvores geradas e a quantidade de atributos a serem usados aleatoriamente em cada árvore.

3.4 Predição Volumétrica, Relação Hipsométrica e Estimativa de Biomassa como KDD

Esta seção mostra que os processos de predição de volumes, definição de relação hipsométrica e estimativa de biomassa podem ser mapeados como processos de KDD e, portanto, todo o avanço nesta área da Ciência da Computação é aplicável também na Mensuração Florestal.

A predição volumétrica na área de mensuração florestal possui um método bem definido e amplamente utilizado, a saber:

- Escolha das árvores amostras do povoamento que será medido;
- Cubagem das árvores da amostra;
- Medição das variáveis independentes das demais árvores, como altura e *dap*;
- Seleção dos modelos a serem ajustados, dentre todos os modelos disponíveis na literatura;
- Ajuste dos coeficientes dos modelos e obtenção das equações volumétricas;
- Validação dos modelos obtidos, aplicando-os ao conjunto de validação;
- Análise da qualidade da estimativa pelo cálculo das estatísticas apresentadas na Seção 3.2.4, e no caso deste estudo correlação, R^2, S_{yx} e análise de resíduos;
- Escolha da melhor equação a partir das estatísticas obtidas no passo anterior;
- Aplicação da equação em todo o povoamento.

Claramente, neste processo, as fases iniciais de medição e cubagem fazem parte da coleta dos dados e não são influenciadas pelas técnicas de AM. A Figura 3.1 ilustra o método tradicional de predição volumétrica.

A Figura 3.2 ilustra a proposta do novo método. As caixas assinaladas são as atividades que são substituídas por técnicas de AM. Pelas características do problema, qualquer técnica de regressão pode ser usada e aqui serão avaliadas as Redes Neurais Artificiais, Máquinas de Vetores de Suporte e *Random Forest*.

Para a obtenção da relação hipsométrica, o processo é muito parecido, o que muda são os modelos matemáticos a serem ajustados. Este procedimento é usado, pois a medição do *dap* é muito menos custosa que a da altura total. Dependendo da densidade do povoamento, condições, etc, pode-se medir a altura de forma imprecisa. Assim, treinam-se modelos de relação hipsométricas que podem ser aplicados ao povoamento (durante um período de tempo, por causa da dinamicidade do povoamento).

Como dados de treinamento tem-se um conjunto de informações contendo o *dap* e h_t. Com um modelo matemático ajustado, consegue-se tomar outras medidas de *dap* do povoamento e estimar suas alturas. Este resultado pode ser usado como entrada para os modelo volumétricos citados anteriormente.

Para estimativa de Biomassa de um indivíduo, o processo é similar, sendo que modelos de regressão diferenciados também foram propostos, como podem ser vistos no Capítulo 2.

Percebe-se um padrão nos métodos de estimativas usados na mensuração florestal. Este padrão envolve a coleta e pré-processamento dos dados, ajuste e escolha de uma equação e aplicação desta equação. Este processo se encaixa perfeitamente como KDD, pois somente a

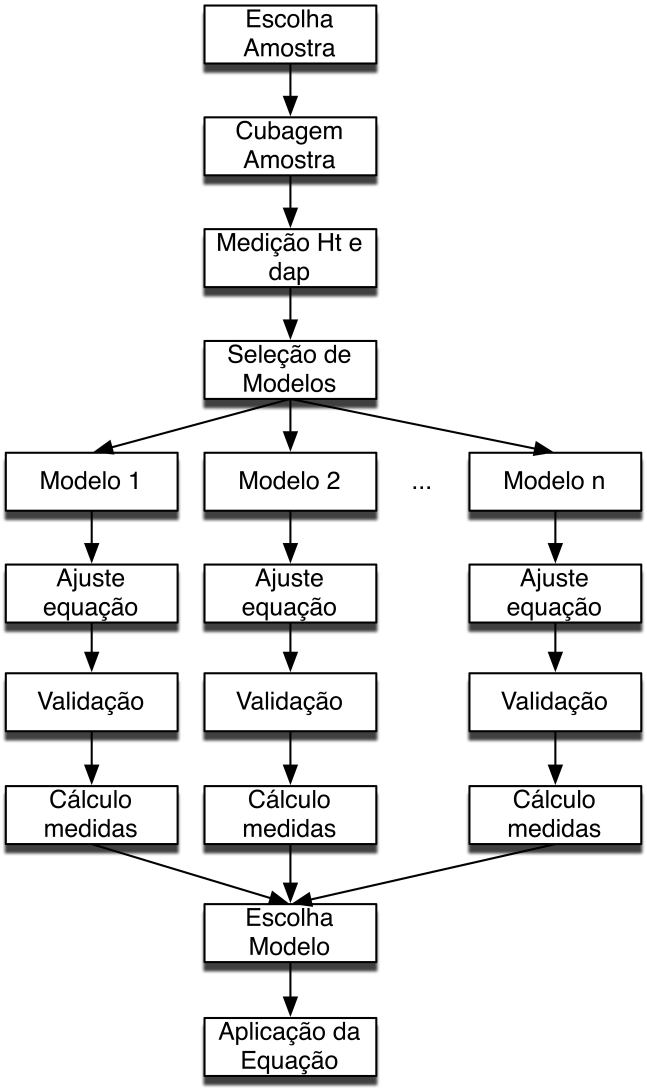


Figura 3.1: Método tradicional de predição volumétrica, onde vários modelos alométricos são ajustados e avaliados pela sua qualidade de estimativa.

Fonte: O autor.

etapa de ajuste da equação é convertida no treinamento de algum modelo de AM voltado para a regressão.

Assim, pode-se encarar a mensuração florestal como um processo de descoberta de conhecimento, onde o avanço na área da computação é aplicado também à área florestal.

3.5 Metodologia dos Experimentos

A área da Engenharia Florestal é rica em bases de dados rotuladas, isto é, com valores observados, para que as técnicas de AM possam ser aplicadas. Inicialmente, cinco bases foram testadas:

- Volume de *Pinus taeda*;
- Biomassa da Acácia-negra (*Acacia mearnsii*);

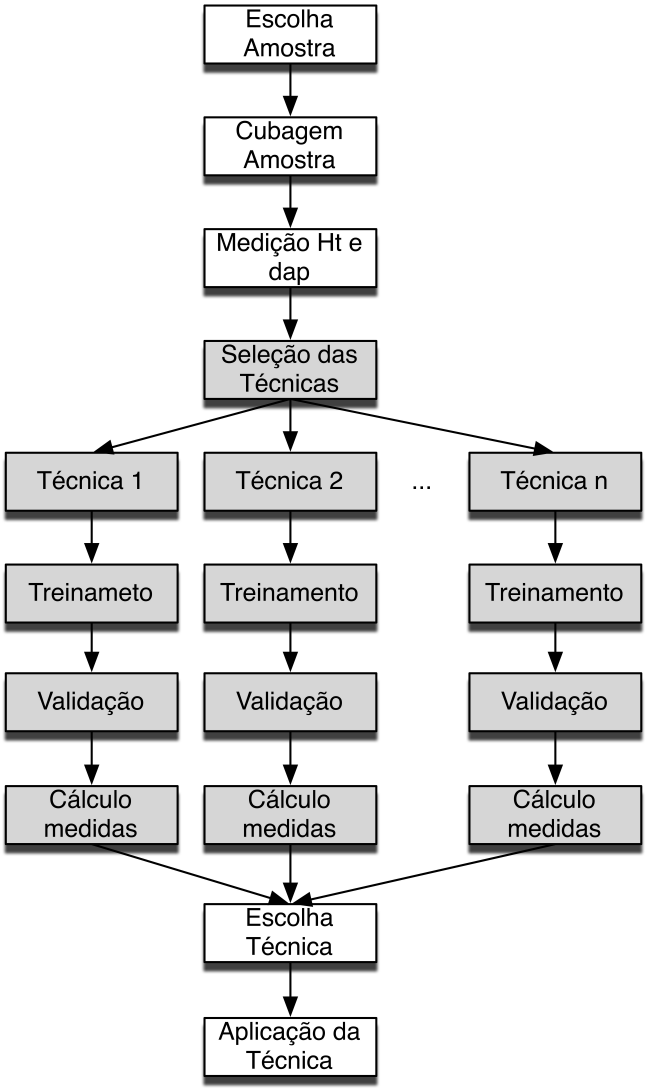


Figura 3.2: Predição volumétrica como KDD. Os modelos alométricos são substituídos por técnicas de AM, indicadas nas áreas sobreadas. Os modelos de AM obtidos também são avaliados conforme a sua qualidade de estimativa.

Fonte: O autor.

- Biomassa seca acima do solo de florestas tropicais;
- Relação hipsométrica de *Pinus taeda*;
- Relação hipsométrica da Acácia-negra (*Acacia mearnsii*).

As bases de pinus e acácia, são menores (302 e 544 amostras) e, portanto, foi utilizada validação cruzada, para diminuir a variância da estimativa em relação à amostra de treinamento. Foram divididos em 10 subconjuntos de tamanhos iguais.

A base de dados de florestas tropicais possui 4004 amostras e, portanto, foi usada a separação em dois subconjuntos, de treinamento e de teste. Os conjuntos são disjuntos para que não haja valores conhecidos no momento que os testes são efetuados. A intenção é se evitar o sobretreinamento (do inglês *overfitting*), treinando/ajustando modelos e testando com a mesma

base. Assim, neste trabalho tanto para modelos alométricos como para as técnicas de AM, as bases são separadas na seguinte proporção, objetivando justiça na comparação:

- Base de Treinamento: 70% da base total, escolhida de forma aleatória;
- Base de Teste: os demais 30% que foram descartados da base de treinamento e que não aparecem nesta.

Como as estatísticas de análise de regressão, em geral, levam em conta a quantidade de coeficientes ajustados, para os casos em que este fator seja necessário, nas técnicas de AM será usado o valor um. Mesmo assim, a principal medida de comparação da qualidade dos modelos sobre as bases testadas, nestes experimentos, é a Correlação, mas também são calculados o R^2, o S_{yx} e a Soma de Quadrados dos Resíduos. Uma breve análise de resíduos também é feita para que se possa analisar tendenciosidade nas estimativas.

As técnicas de AM são baseadas em ajustes de parâmetros de entrada e cada técnica possui suas particularidades. As redes neurais treinadas foram do tipo MLP, com uma camada oculta e quantidade de neurônios variando de 1 a 100. A taxa de aprendizado e o momentum foram variados entre 0,1 e 0,9, em saltos de 0,1. Foi usado um subconjunto do conjunto de treinamento, chamado de conjunto de validação, que totalizou 20% do conjunto de treinando. Foram efetuadas, no máximo, 2000 épocas.

As máquinas de vetores de suporte forma treinadas com o algoritmo SMO, variando-se o custo C e *gamma* até se encontrar os melhores valores de correlação. O parâmetro C foi variado até 10000 e, após, foi feita uma análise de sensibilidade para detectar regiões promissoras, com melhor correlação. Nestas regiões, os saltos foram diminuídos. O *kernel* utilizado foi do tipo RBF e o valor de *gamma* foi variado de 0,1 até 0,9 em saltos de 0,1, em cada teste efetuado. Em algumas faixas, *gamma* também foi variado de 0,01 até 0,09, em saltos de 0,01.

As florestas aleatórias foram foram treinadas variando-se a quantidade de árvores de 2 até 200. A quantidade de atributos, escolhidos de forma aleatória para geração das árvores, foram tomados de 2 até a quantidade total de atributos.

Para o treinamento e teste dos modelos alométricos foi utilizado o R. Já os modelos de AM foram treinados e testados usado o WEKA. Todos os experimentos foram realizados em uma máquina AMD Opteron(tm) Processor 6136, com 32 núcleos e 120Gb RAM.

3.5.1 Método de Análise dos Resultados e Escolha do Melhor Modelo

Para cada modelo testado, seja de alométrico ou AM, são calculadas as medidas estatísticas apresentadas no Capítulo 2. Apesar de se observar em trabalhos da área de regressão que o R^2 e S_{yx} são tradicionalmente usados para a escolha do modelo, estas medidas sofrem alteração pela quantidade de amostras e quantidade de coeficientes. Nos experimentos realizados, as medidas estatísticas apresentaram uniformidade sobre seu ranqueamento, podendo-se usar qualquer medida calculada para análise da qualidade dos modelso. Assim para a escolha do melhor modelo neste trabalho foi utilizado o Coeficiente de Correlação de Pearson.

Apesar da Soma de Quadrados de Resíduos também sofrer alteração pela quantidade de amostras, deve-se levar em consideração que este dá uma medida bruta dos resíduos produzidos pela estimativa que, quando se comparam técnicas diferentes com a mesma base de dados, dá um resultado preciso de ganho em termos de qualidade de predição.

São usadas também análises gráficas dos resíduos, para que se possa verificar visualmente a dispersão destes, bem como a possível identificação de padrões comuns à estimativas tendenciosas. O gráfico gerado para cada estimativa é o Gráfico de Resíduos, que no eixo x apresenta o valor observado e no eixo y o valor estimado.

3.5.2 Análise Estatística dos Resultados

Assim, para analisar os resultados de forma mais precisa, deve-se verificar se há diferença estatística entre os desempenhos de K modelos sobre N conjuntos de dados, que pode ser feito através do teste de Friedman [Demsar, 2006]. Este teste é uma versão não-paramétrica equivalente ao ANOVA de medidas repetidas, para análise de dados pareados. O teste fundamenta-se na comparação de posições de desempenhos (*rank*) e, portanto, para cada conjunto de dados, cada um dos modelos avaliados é associado a uma posição de acordo com seu desempenho, ordenados dos melhores para os piores. Em caso de empates, valores médios de posição são atribuídos. Após, é determinado o *rank* médio para cada um dos K modelos avaliados.

Assim, seja R_{ij} a posição de desempenho do modelo j, tal que $1 \leq j \leq K$ para um conjunto de dados i, tal que $1 \leq i \leq N$, o *rank* médio R_j é calculado conforme a seguinte equação.

$$R_j = \frac{\sum_{i=1}^{N} R_{ij}}{N} \tag{3.4}$$

onde:

- R_{ij} : posição de desempenho do modelo j para o conjunto de dados i;
- N : quantidade de bases de dados;
- R_j : *rank* médio do modelo j.

Sob a hipótese nula de que o desempenho de todos os modelos comparados são equivalentes, e que portanto seus valores de *rank* médio são iguais, a estatística de Friedman é expressa pela seguinte equação.

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{j=1}^{K} R_j^2 - \frac{K(K+1)^2}{4} \right] \tag{3.5}$$

onde:

- N : quantidade de bases de dados;
- K : quantidade de modelos;
- R_j : *rank* médio do modelo j;
- χ_F^2 : estatística de Friedman, de acordo com a distribuição χ^2.

Em [Demsar, 2006], entretanto, é recomendada a utilização de uma versão menos conservadora do teste de Friedman, proposta em [Iman e Davenport, 1980], a qual é distribuída de acordo com a distribuição F com $(K-1)$ e $(K-1)(N-1)$ graus de liberdade. O cálculo dessa estatística é apresentado na seguinte equação.

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1) - \chi_F^2} \tag{3.6}$$

onde:

- N : quantidade de bases de dados;

- K : quantidade de modelos;
- χ_F^2 : estatística de Friedman;
- F_F : estatística de Friedman relaxada.

Quando a hipótese nula é rejeitada pelo teste de Friedman, isso implica que existe diferença significativa de desempenhos, no entanto não permite discriminar quais modelos apresentam diferença. Nesse cenário, o pós-teste de Nemenyi [Nemenyi, 1963] pode ser utilizado para detectar quais diferenças entre os modelos são significativas [Demsar, 2006]. De acordo com esse teste, a eficácia de dois métodos é significativamente diferente sempre que seus correspondentes *ranks* médios diferem por pelo menos um determinado valor de diferença crítica (CD), dado pela seguinte equação.

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}} \tag{3.7}$$

onde:

- N : quantidade de bases de dados;
- K : quantidade de modelos;
- q_α : valor baseado na amplitude total estudentizada (*studentized range statistic*), dividida por $\sqrt{2}$;
- CD : diferença crítica calculada.

Quando a comparação de vários modelos é realizada em relação a um único modelo controle, recomenda-se a aplicação do pós-teste por meio da estatística de Bonferroni-Dunn [Dunn e Dunn, 1961], na qual q_α pode ser menos restritivo. Nesse caso, somente $K - 1$ comparações são realizadas, o que confere maior poder estatístico ao pós-teste.

Na Tabela 3.11 são apresentados os valores de q_α com $\alpha = 0,05$, para diferentes quantidades de métodos a serem comparados, utilizando os pós-teste de Nemenyi e de Bonferroni-Dunn [Nemenyi, 1963; Dunn e Dunn, 1961].

Tabela 3.11: Valores de q_α para $\alpha = 0,05$ para diferentes valores K no pós-teste de Nemenyi e de Bonferroni-Dunn.

K	Nemenyi	Bonferroni-Dunn
2	1,960	1,960
3	2,343	2,241
4	2,569	2,394
5	2,728	2,498
6	2,850	2,576
7	2,949	2,648
8	3,031	2,690
9	3,102	2,690
10	3,164	2,773

3.6 Considerações

Neste capítulo foram apresentados os dados usados nos experimentos e suas origens, os métodos de mensuração florestal abordados e as técnicas de aprendizado de máquina usados em substituição aos modelos alométricos. Os métodos de mensuração florestal em foco neste trabalho são a estimativa de volumes, avaliação de biomassa e relação hipsométrica. As técnicas de aprendizado de máquina abordadas aqui são as redes neurais artificiais, máquinas de vetores de suporte e *random forest*, conforme descritas no Capítulo 2.

Foi apresentado também o processo tradicional usado em mensuração florestal e sua estrutura quando da substituição dos modelos alométricos pelas técnicas de aprendizado de máquina. Para a decisão dos melhores modelos são usados a correlação, R^2, S_{yx}, soma de quadrados dos resíduos e análise gráfica dos resíduos. Após a obtenção destas métricas, faz-se o teste de Friedman para determinar se há diferença estatística e, em caso positivo, o pós-teste de Nemenyi para determinar quais modelos possuem esta diferença.

Capítulo 4

Resultados e Discussões

Neste capítulo são apresentados os resultados da pesquisa e as considerações relativas aos experimentos realizados sobre os dados apresentados na Seção 3.1. Foram treinados modelos de AM para RNAs, SVMs e RF, por meio de variação dos parâmetros conforme apresentados na Seção 3.5, e foram confrontados com os modelos alométricos tradicionalmente utilizados para relação hipsométrica, predição de volumes e avaliação de biomassa.

A avaliação da qualidade dos modelos obtidos foi efetuada através de análise estatística de indicadores, conforme apresentado na Seção 3.2.4 e sua comparação foi por meio do teste de Friedman e pós-tese de Nemenyi, conforme apresentado na Seção 3.5.2.

4.1 Predição de Volumes de *Pinus taeda*

Para os experimentos com a base da Acácia-negra foram usados os modelos de Spurr, Husch e Schumacher & Hall, que podem ser vistos na Tabela 4.1.

Tabela 4.1: Coeficientes ajustados dos modelos alométricos para a base de *Pinus taeda*

Nome	Modelo	β_0	β_1	β_2
Spurr	$v = \beta_0 + \beta_1 dap^2 H_t$	0,008841	0,00003412	
Husch	$\ln v = \beta_0 + \beta_1 \ln dap$	-9,696	2,715	
Schumacher & Hall	$\ln v = \beta_0 + \beta_1 \ln dap + \beta_2 \ln H_t$	-10,028	1,877	1,057

O treinamento das RNAs, SVMs e RF foram baseados em ajustes de parâmetros, que foram variados conforme apresentados na Tabela 4.2, Tabela 4.3 e Tabela 4.4.

Tabela 4.2: Parâmetros de Treinamento das RNAs para Volume de *Pinus taeda*

Parâmetro	Faixa	Saltos
Camadas Ocultas	1	
Neurônios na Camada Oculta	1 a 100	1
Taxa de Aprendizado	0,1 a 0,9	0,1
Momentum	0,1 a 0,9	0,1
Conjunto de Validação	20%	
Número de Épocas	2000	

Tabela 4.3: Parâmetros de Treinamento das SVMs para Volume de *Pinus taeda*

Parâmetro	Faixa	Saltos
Custo (C)	100 a 10000	100, análise de sensibilidade
	4000 a 6000	10
Gamma	0,1 a 0,9	0,1
	0,01 a 0,09	0,01

Tabela 4.4: Parâmetros de Treinamento das RFs para Volume de *Pinus taeda*

Parâmetro	Faixa	Saltos
Quantidade de Árvores	2 a 200	1
Quantidade de Atributos	2 a 3	1

A melhor RNA obtida foi com Taxa de Aprendizado de 0,1, Momentum de 0,1 e 2 neurônios na camada oculta. A SVM que obteve melhor correlação foi com Custo de 5000 e *gamma* de 0,01. A melhor RF foi obtida com 28 árvores e com 2 atributos tomados aleatoriamente.

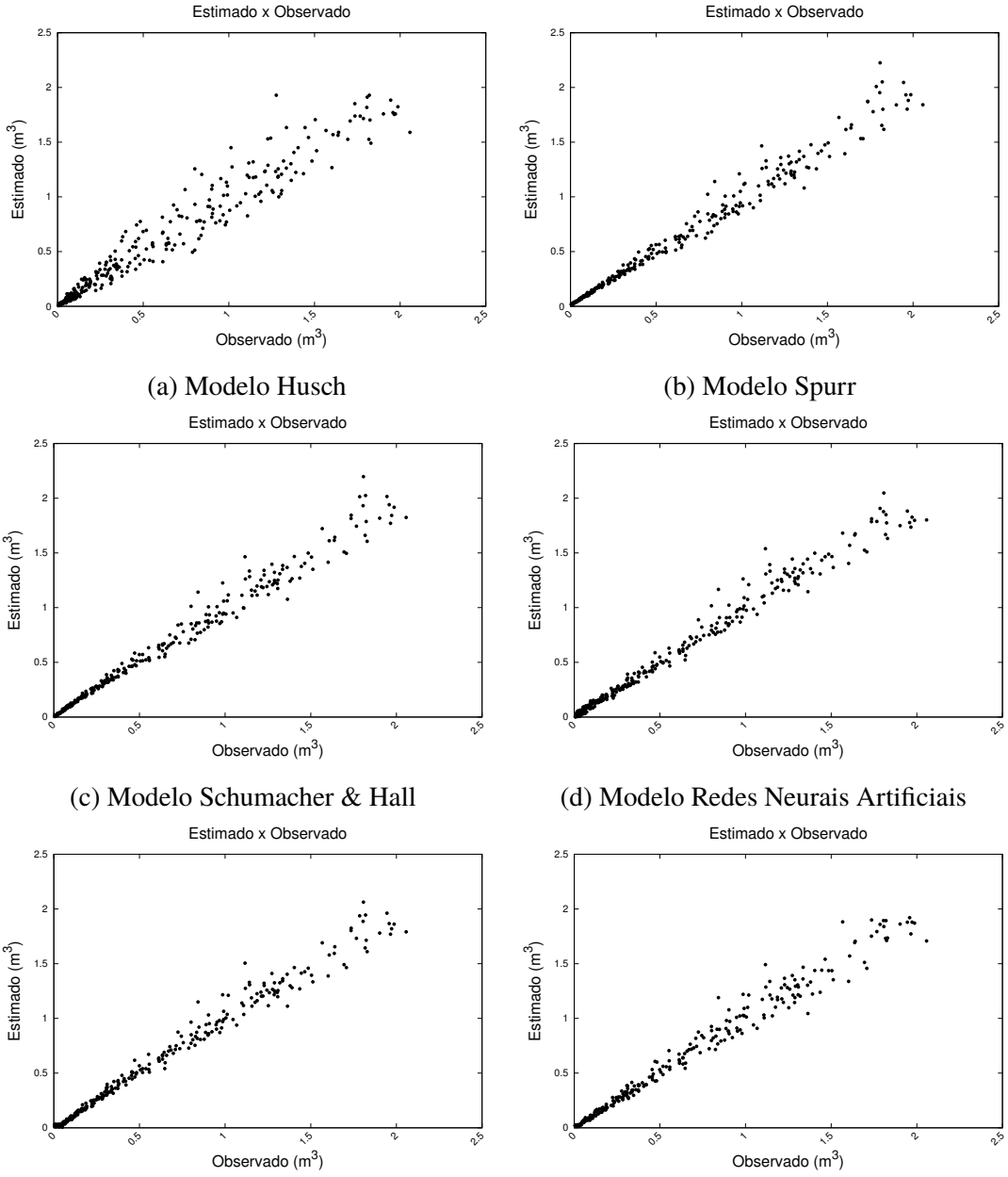
A Tabela 4.5 apresenta os resultados de Correlação, R^2, S_{yx}% e SQRes, para os dados de Pinus aplicados à base de dados usando-se validação cruzada em todos os métodos aplicados, alométricos ou de aprendizado de máquina.

Tabela 4.5: Resultados de Correlação, R^2, S_{yx} e Soma dos quadrados dos resíduos aplicados à base de Pinus usando-se validação cruzada. As células marcadas contém os melhores resultados.

Modelo	Correlação	R^2	S_{yx}%	SQRes
Husch	0,9733	0,9473	22,0572	5,0099E+00
Spurr	0,9906	0,9814	13,1086	1,7695E+00
Schumacher & Hall	0,9911	0,9823	12,7844	1,6830E+00
Redes Neurais Artificiais	0,9914	0,9828	12,5573	1,6292E+00
Máquinas de Vetores de Suporte	0,9919	0,9837	12,2242	1,5439E+00
Florestas Aleatórias	0,9904	0,9809	13,2522	1,8145E+00

Apesar da difença na correlação entre o melhor modelo alométrico e os dois melhores modelos de AM ser pequena, a superioridade dos modelos de AM é observada aqui levando-se em consideração que tanto RNA como SVM superaram as marcas do modelo Schumacher & Hall.

A Figura 4.1 apresenta os gráficos de valores observados versus estimados, gerados para todos os modelos testados. Percebe-se que todos os modelos apresentam pontos próximos à reta de regressão, sem a indicação de tendências nas estimativas. Como esperado, os três melhores modelos (Máquinas de Vetores de Suporte, Redes Neurais Artificiais e Schumacher & Hall) possuem os pontos mais concentrados na linha de regressão, indicando melhores estimativas.



(a) Modelo Husch

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

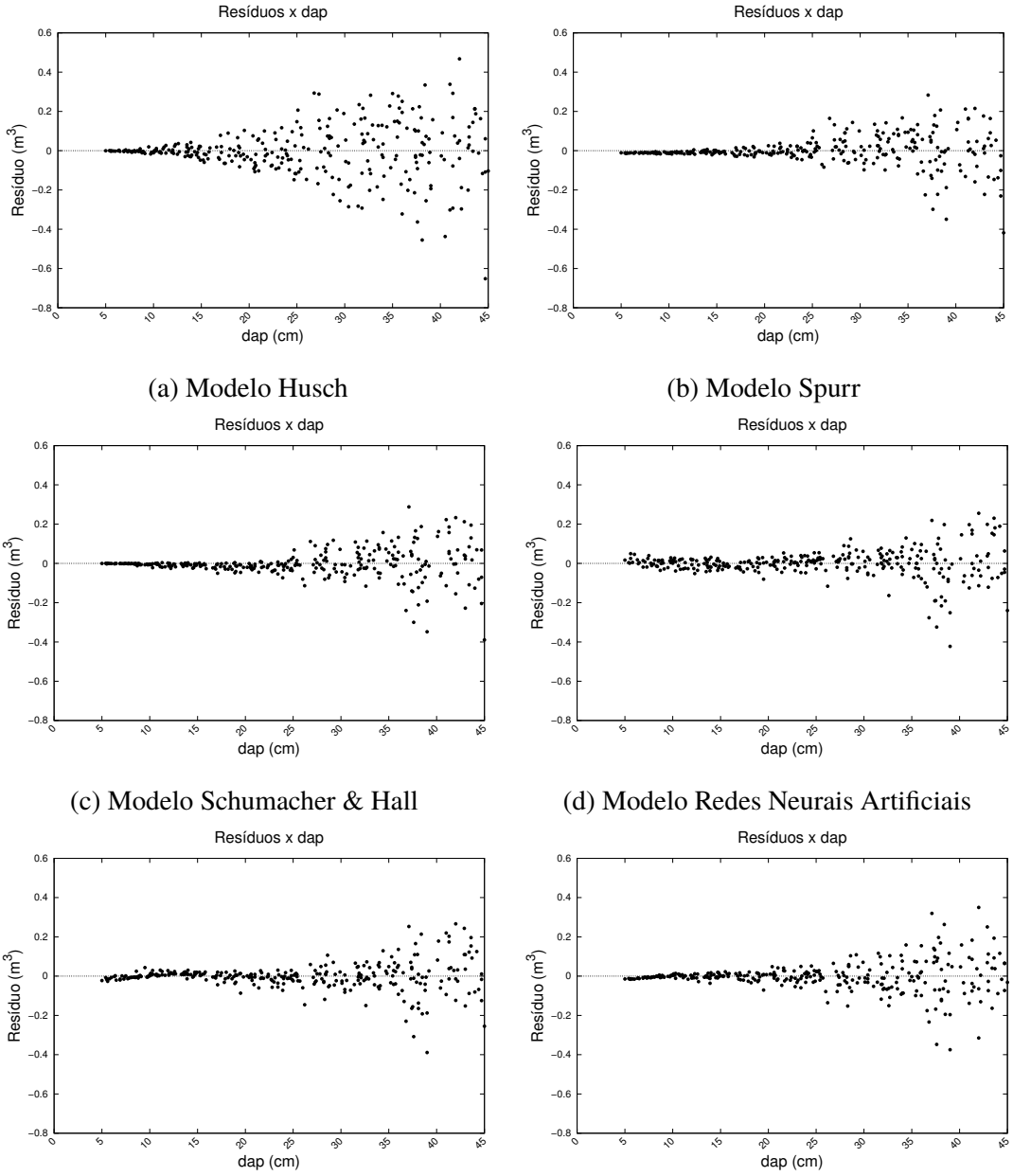
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.1: Gráfico de Resíduos, estimado versus observado, para estimativa de volumes de *Pinus taeda*, para os modelos alométricos de Husch, Spurr e Schumacher & Hall, e para os modelos de aprendizado de máquina RNA, SVM e RF.

A Figura 4.2 apresenta os gráficos de resíduos por *dap*, gerados para todos os modelos testados. Percebe-se por estes gráficos que os resíduos são muito maiores em árvores com *dap* maior, para todos os modelos testados.



(a) Modelo Husch

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

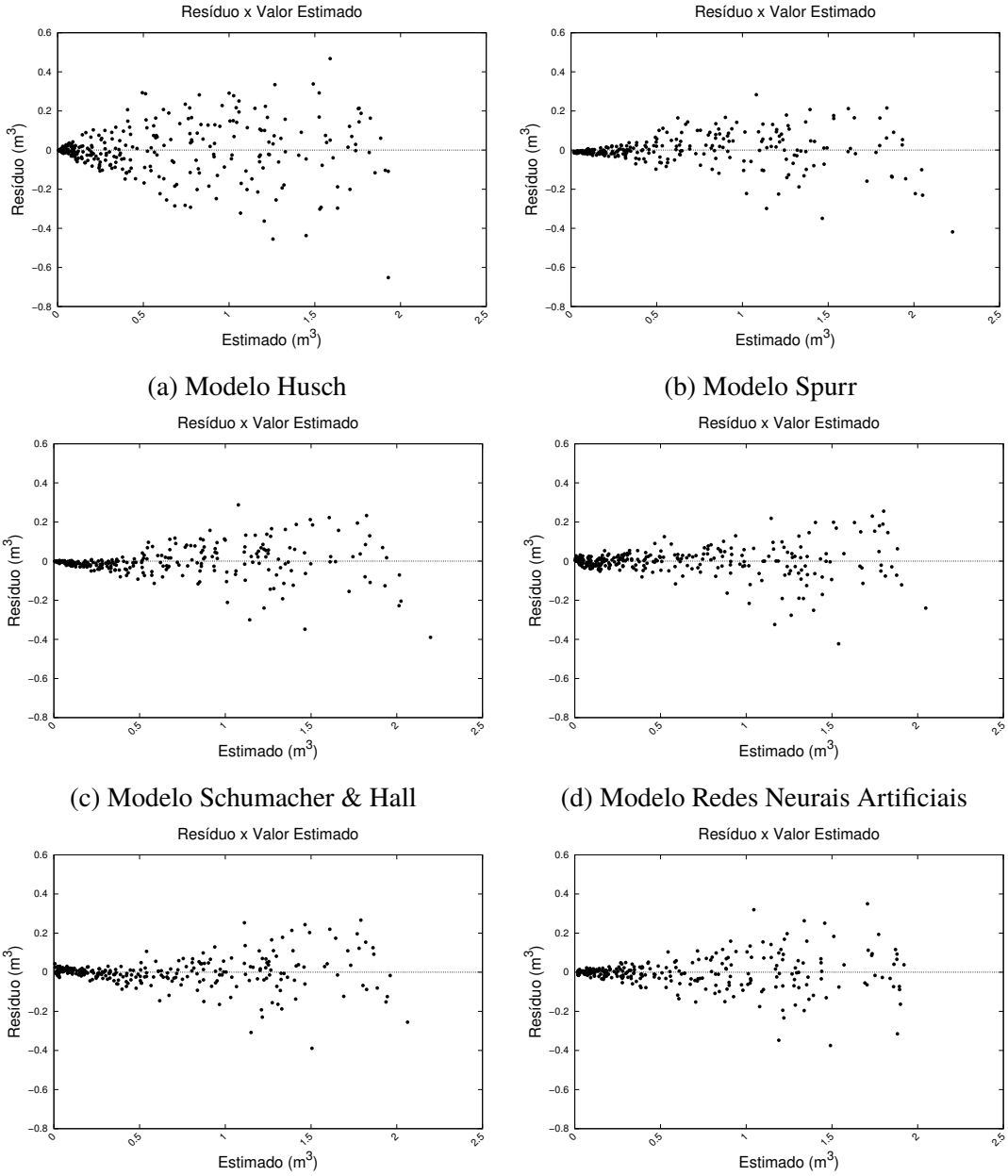
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.2: Gráfico de Resíduos por *dap* para estimativa de Volumes de *Pinus taeda*, para os modelos alométricos de Husch, Spurr e Schumacher & Hall, e para os modelos de aprendizado de máquina RNA, SVM e RF.

A Figura 4.3 apresenta os gráficos de resíduos por ajustes, gerados para todos os modelos testados. Estes gráficos são usados para analisar o pressuposto de homocedasticidade dos resíduos (mesma variância dos resíduos em relação ao valor estimado). Nos modelos testados, percebe-se um padrão em forma de funil, indicando que a variância dos resíduos aumenta conforme o valor estimado aumenta, indicando a heterocedasticidade dos modelos.



(a) Modelo Husch

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.3: Gráfico de Resíduos por Ajustes: Estimativa de Volumes de *Pinus taeda*

Para esta base de dados, o modelo treinado usando Máquinas de Vetores de Suporte obteve a maior correlação $(0,9919)$, maior R^2 $(0,9837)$, menor $S_{yx}\%$ $(12,2242)$ e menor soma dos quadrados dos resíduos $(1,5439E+00)$. Com os melhores resultados em todas as medidas efetuadas, conclui-se que para esta base de dados o modelo SVM é o melhor. Mesmo assim, os valores são muito próximos aos obtidos pelos modelos de Redes Neurias Artificiais e o modelo alométrico de Schumacher & Hall.

O pior modelo observado foi o de Husch, com correlação de $0,9733$. Apesar deste valor menor que os demais, o modelo de Husch usa somente o *dap* como valor de entrada para a estimativa de volume, apresentando uma grande simplicidade e evitando a medição das alturas das árvores da amostra.

4.2 Avaliação de Biomassa Seca da Acácia-negra

Para os experimentos com a base da Acácia-negra foram usados os modelos de Spurr, Kopezky & Gehrhardt e Schumacher & Hall, que podem ser vistos na Tabela 4.6.

Tabela 4.6: Coeficientes ajustados dos modelos alométricos para a base de Acácia-negra

Nome	Modelo	β_0	β_1	β_2
Spurr	$b = \beta_0 + \beta_1 dap^2 H_t$	1,17758	0,02596	
Kopezky & Gehrhardt	$\ln b = \beta_0 + \beta_1 \ln dap$	-1,364	2,158	
Schumacher & Hall	$\ln b = \beta_0 + \beta_1 \ln dap + \beta_2 \ln H_t$	-1,6857	1,6855	0,5509

O treinamento das RNAs, SVMs e RF foram baseados em ajustes de parâmetros, que foram variados conforme apresentados na Tabela 4.7, Tabela 4.8 e Tabela 4.9.

Tabela 4.7: Parâmetros de Treinamento das RNAs para Biomassa de Acácia-negra

Parâmetro	Faixa	Saltos
Camadas Ocultas	1	
Neurônios na Camada Oculta	1 a 100	1
Taxa de Aprendizado	0,1 a 0,9	0,1
Momentum	0,1 a 0,9	0,1
Conjunto de Validação	20%	
Número de Épocas	2000	

Tabela 4.8: Parâmetros de Treinamento das SVMs para Biomassa de Acácia-negra

Parâmetro	Faixa	Saltos
Custo (C)	100 a 10000	1000, análise de sensibilidade
Gamma	0,1 a 0,9	0,1
	0,01 a 0,09	0,01

Tabela 4.9: Parâmetros de Treinamento das RFs para Biomassa de Acácia-negra

Parâmetro	Faixa	Saltos
Quantidade de Árvores	2 a 200	1
Quantidade de Atributos	6	

A melhor RNA obtida foi com Taxa de Aprendizado de 0,5, Momentum de 0,2 e 18 neurônios na camada oculta. A SVM que obteve melhor correlação foi obtida com C de 9900 e *gamma* de 0,1. A melhor floresta foi obtida com 150 árvores.

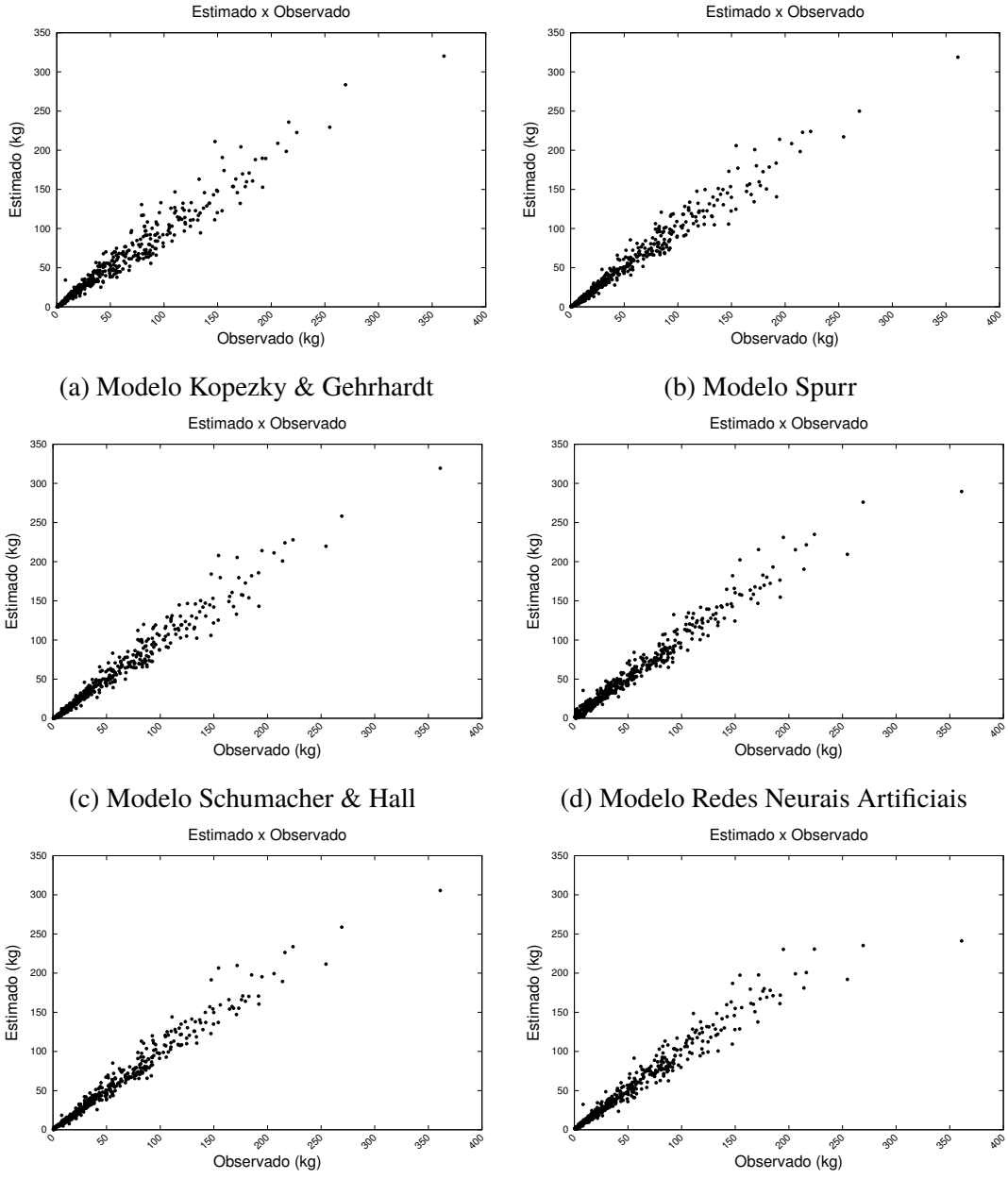
A Tabela 4.10 apresenta os resultados de Correlação, R^2, S_{yx}% e SQRes, para os dados de Acácia-negra aplicados à base de dados usando-se validação cruzada em todos os métodos aplicados, alométricos ou de aprendizado de máquina.

Tabela 4.10: Resultados de Correlação, R^2, S_{yx} e Soma dos quadrados dos resíduos aplicados à base de Acácia-negra usando-se validação cruzada. As células marcadas contém os melhores resultados.

Modelo	Correlação	R^2	S_{yx}%	SQRes
Kopezky & Gehrhardt	0,9768	0,9541	22,7161	6,1411E+04
Spurr	0,9830	0,9663	19,4647	4,5090E+04
Schumacher & Hall	0,9827	0,9656	19,6619	4,6008E+04
Redes Neurais Artificiais	0,9831	0,9651	19,7976	4,6734E+04
Máquinas de Vetores de Suporte	0,9860	0,9722	17,6807	3,7274E+04
Florestas Aleatórias	0,9782	0,9565	22,1013	5,8243E+04

Observa-se na Tabela 4.10 que os dois melhores modelos de AM, no caso RNA e SVM, foram superiores ao melhor modelo alométrico, nesta base sendo o Spurr. Apesar disso, mesmo a diferença em termos de correlação sendo pequena, ao se observar a soma dos quadrados dos resíduos percebe-se que as estimativas dos modelos de AM são melhores.

A Figura 4.4 apresenta os gráficos de valores observados versus estimados, gerados para todos os modelos testados. Percebe-se que todos os modelos apresentam pontos próximos à reta de regressão, sem a indicação de tendências nas estimativas. Como esperado, os três melhores modelos (Máquinas de Vetores de Suporte, Redes Neurais Artificiais e Spurr) possuem os pontos mais concentrados na linha de regressão, indicando melhores estimativas.



(a) Modelo Kopezky & Gehrhardt

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

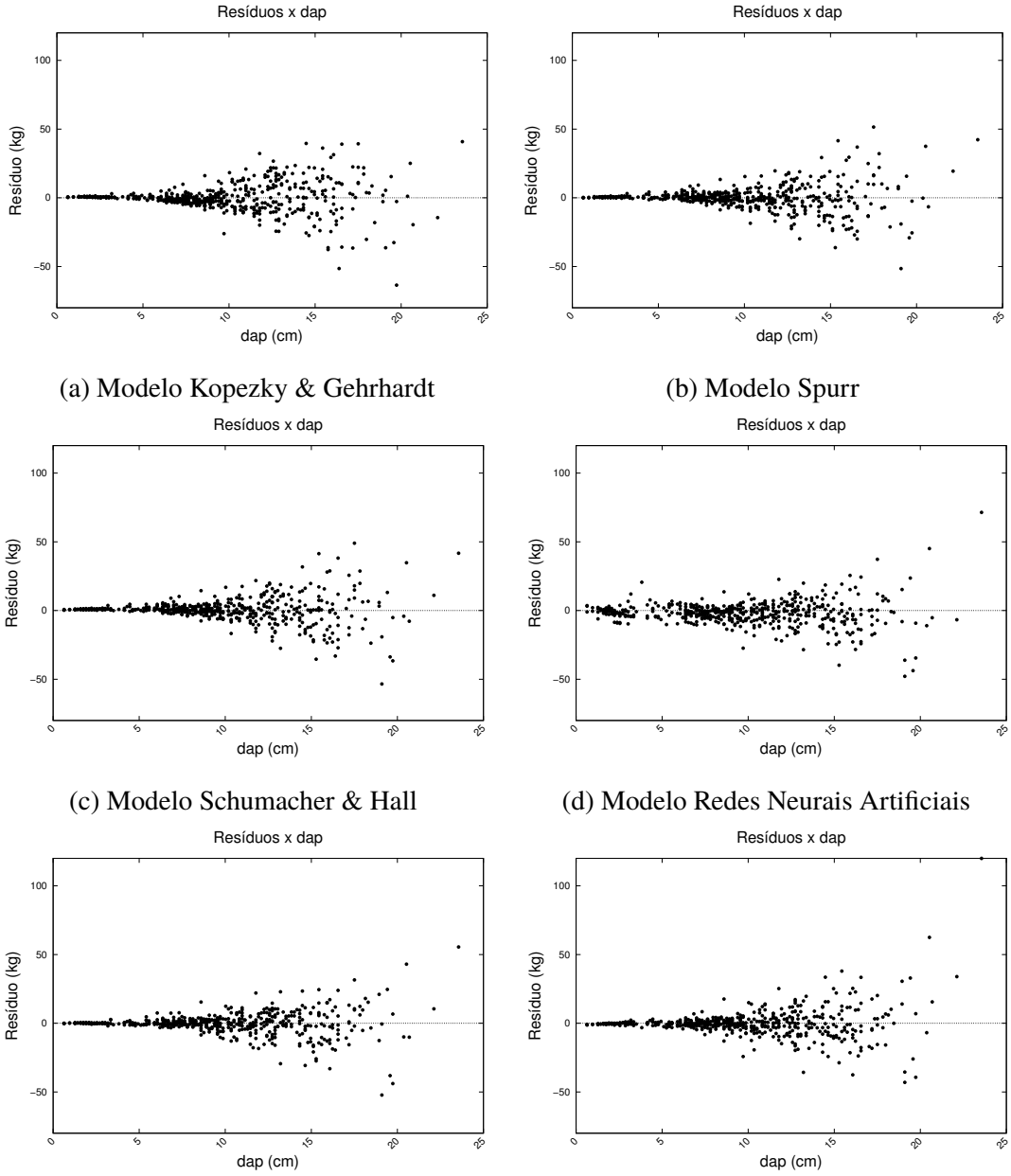
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.4: Gráfico de Resíduos: Estimativa de Biomassa da Acácia-negra

A Figura 4.5 apresenta os gráficos de resíduos por *dap*, gerados para todos os modelos testados. Percebe-se por estes gráficos que os resíduos são muito maiores em árvores com *dap* maior, para todos os modelos testados.



(a) Modelo Kopezky & Gehrhardt

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

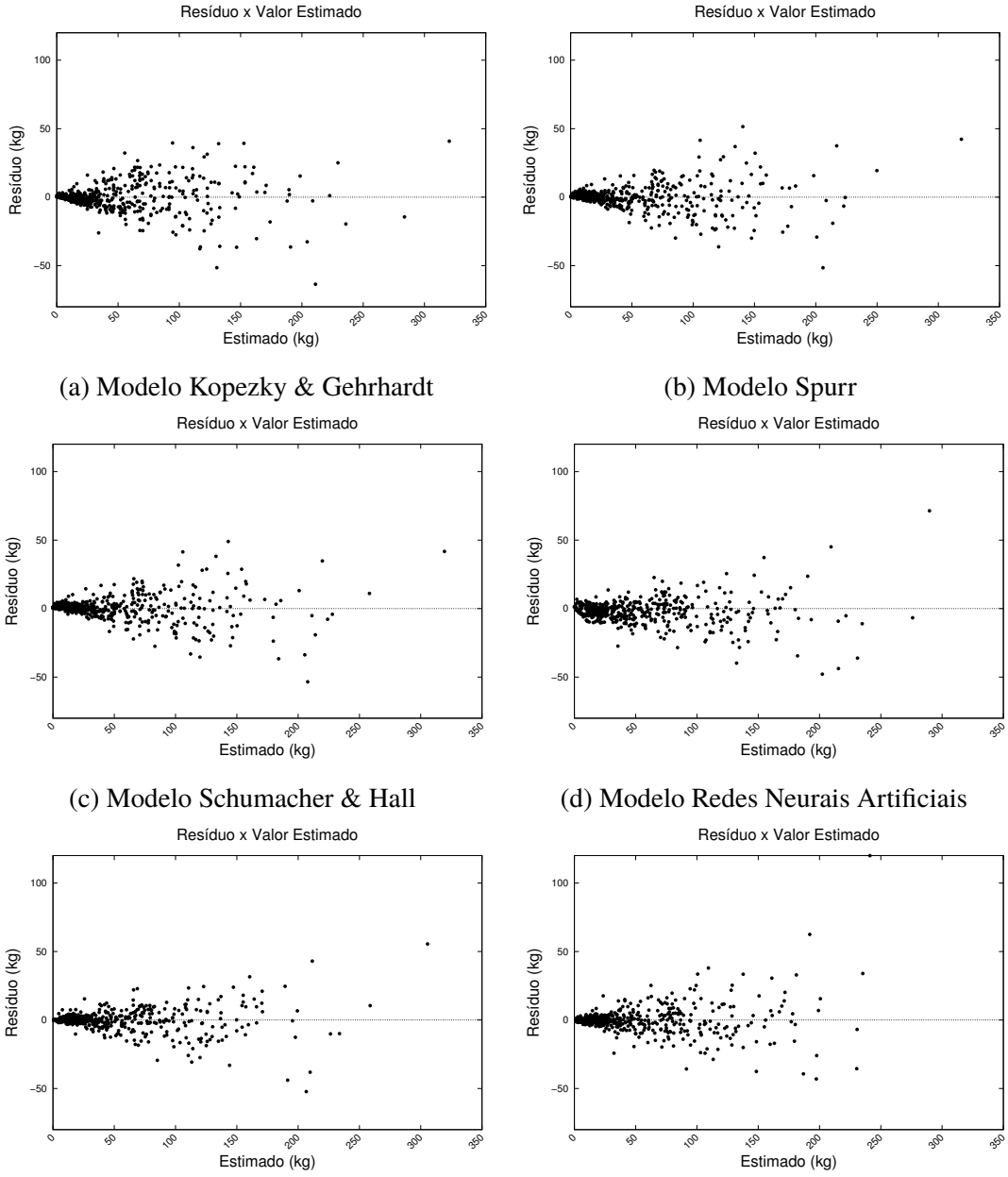
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.5: Gráfico de Resíduos por *dap*: Estimativa de Biomassa da Acácia-negra

A Figura 4.6 apresenta os gráficos de resíduos por ajustes, gerados para todos os modelos testados. Estes gráficos são usados para analisar o pressuposto de homocedasticidade dos resíduos (mesma variância dos resíduos em relação ao valor estimado). Nos modelos testados, percebe-se um padrão em forma de funil, indicando que a variância dos resíduos aumenta conforme o valor estimado aumenta, indicando a heterocedasticidade dos modelos.



(a) Modelo Kopezky & Gehrhardt

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.6: Gráfico de Resíduos por Ajustes: Estimativa de Biomassa da Acácia-negra

Para esta base de dados, o modelo treinado usando Máquinas de Vetores de Suporte obteve a maior correlação (0,9860), maior R^2 (0,9722), menor $S_{yx}\%$ (17,6807) e menor soma dos quadrados dos resíduos ($3,7274E+04$). Com os melhores resultados em todas as medidas efetuadas, conclui-se que para esta base de dados o modelo obtido de SVM é o melhor. Mesmo assim, os valores são muito próximos aos obtidos pelos modelos de Redes Neurais Artificiais, modelo alométrico de Spurr e o modelo alométrico de Schumacher & Hall.

4.3 Avaliação de Biomassa de Florestas Tropicais

Para os experimentos com a base de florestas tropicais apresentado por [Chave et al., 2014], os ajustes dos modelos alométricos sugeridos pelos autores foram re-

feitos usando-se a mesma base de treinamento aplicada aos modelos baseados em aprendizado de máquina. Foram usados 70% da base total, tomados ao acaso, como base de treinamento e os 30% restantes como base de teste.

A Tabela 4.11 apresenta os coeficientes das equações ajustadas para os modelos alométricos sugeridos, vem como para o modelo de Schumacher & Hall.

Tabela 4.11: Coeficientes ajustados dos modelos alométricos para a base de Florestas Tropicais

Modelo	Modelo	β_0	β_1	β_2
Alométrico 1 (3.2)	$AGB_{est} = \beta_0 \times (ME \times dap^2 \times H_t)^{\beta_1}$	0,05339	0,99389	
Alométrico 2 (3.3)	$AGB_{est} = \beta_0 \times (ME \times dap^2 \times H_t)$	0,0494		
Schumacher & Hall	$\ln AGB_{est} = \beta_0 + \beta_1 \ln dap + \beta_2 \ln H_t$	-3,0212	2,0077	0,8244

O treinamento das RNAs, SVMs e RF foram baseados em ajustes de parâmetros, que foram variados conforme apresentados na Tabela 4.12, Tabela 4.13 e Tabela 4.14.

Tabela 4.12: Parâmetros de Treinamento das RNAs para Biomassa de Florestas Tropicais

Parâmetro	Faixa	Saltos
Camadas Ocultas	1	
Neurônios na Camada Oculta	1 a 100	1
Taxa de Aprendizado	0,1 a 0,9	0,1
Momentum	0,1 a 0,9	0,1
Conjunto de Validação	20%	
Número de Épocas	2000	

Tabela 4.13: Parâmetros de Treinamento das SVMs para Biomassa de Florestas Tropicais

Parâmetro	Faixa	Saltos
Custo (C)	100 a 10000	1000, análise de sensibilidade
Gamma	0,1 a 0,9	0,1
	0,01 a 0,09	0,01

Tabela 4.14: Parâmetros de Treinamento das RFs para Biomassa de Florestas Tropicais

Parâmetro	Faixa	Saltos
Quantidade de Árvores	2 a 200	1
Quantidade de Atributos	3 a 4	1

A melhor RNA obtida foi com Taxa de Aprendizado de 0,4, Momentum de 0,3 e 23 neurônios na camada oculta. A SVM que apresentou melhor correlação foi obtida com Custo de 6500 e *gamma* de 0,5. A melhor RF foi obtida com 2 árvores e 3 atributos tomados aleatoriamente.

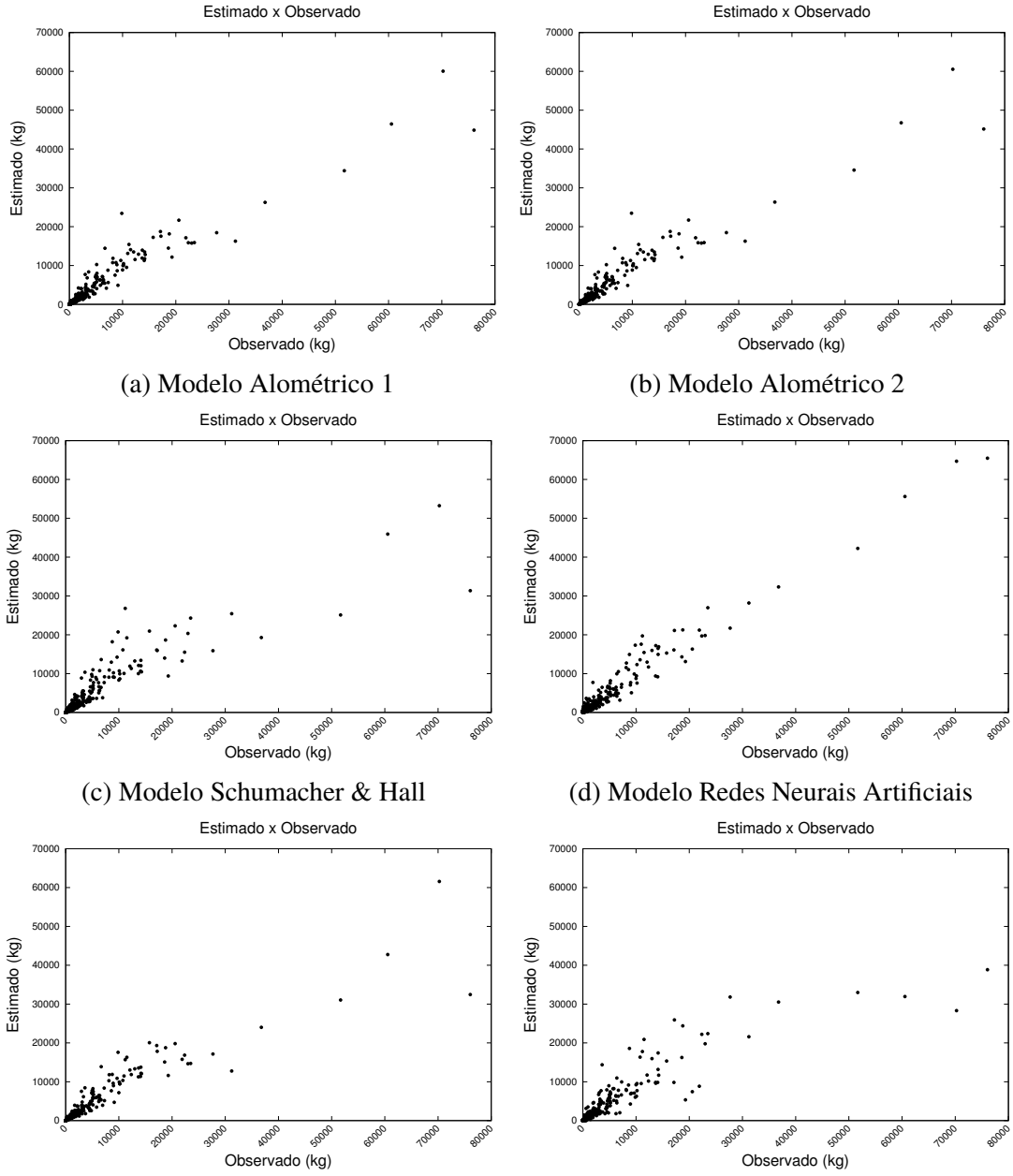
A Tabela 4.15 apresenta os resultados de Correlação, R^2, S_{yx}% e SQRes, para os dados de Florestas Tropicais usando-se a base de treinamento para ajuste dos modelos e a base de teste para levantamento da qualidade dos modelos ajustados.

Tabela 4.15: Resultados de Correlação, R^2, S_{yx} e Soma dos quadrados dos resíduos aplicados à base de Florestas Tropicais usando-se 70% para treinamento e 30% para teste. As células marcadas contém os melhores resultados.

Modelo	Correlação	R^2	S_{yx}%	SQRes
Alométrico 1	0,9614	0,8993	126,7880	2,7964E+09
Alométrico 2	0,9619	0,9008	125,8209	2,7539E+09
Schumacher & Hall	0,9145	0,8188	170,0594	5,0309E+09
Redes Neurais Artificiais	0,9806	0,9595	80,3866	1,1251E+09
Máquinas de Vetores de Suporte	0,9597	0,8775	139,7938	3,4024E+09
Florestas Aleatórias	0,9025	0,7832	185,9476	6,0199E+09

Pode-se observar a superioridade do modelo RNA obtido, que superou o melhor modelo alométrico em quase 2% de correlação e mais de 40% em termos de S_{yx}. Nesta base de dados, o melhor modelo SVM teve uma acurácia pior que os modelos alométricos, pois o tempo de treinamento, neste caso, se tornou inviável para algumas configurações onde o valor do parâmetro C era alto. Nesta base de dados, ao se observar a soma dos quadrados dos resíduos percebe-se que as estimativas do modelos de SVM são significativamente melhores, se comparados aos demais modelos.

A Figura 4.7 apresenta os gráficos de valores observados versus estimados, gerados para todos os modelos testados. Percebe-se que todos os modelos apresentam pontos próximos à reta de regressão, sem a indicação de tendências nas estimativas. Como esperado, os três melhores modelos (Redes Neurais Artificiais, Alométrico 2 e Alométrico 1) possuem os pontos mais concentrados na linha de regressão, indicando melhores estimativas.



(a) Modelo Alométrico 1

(b) Modelo Alométrico 2

(c) Modelo Schumacher & Hall

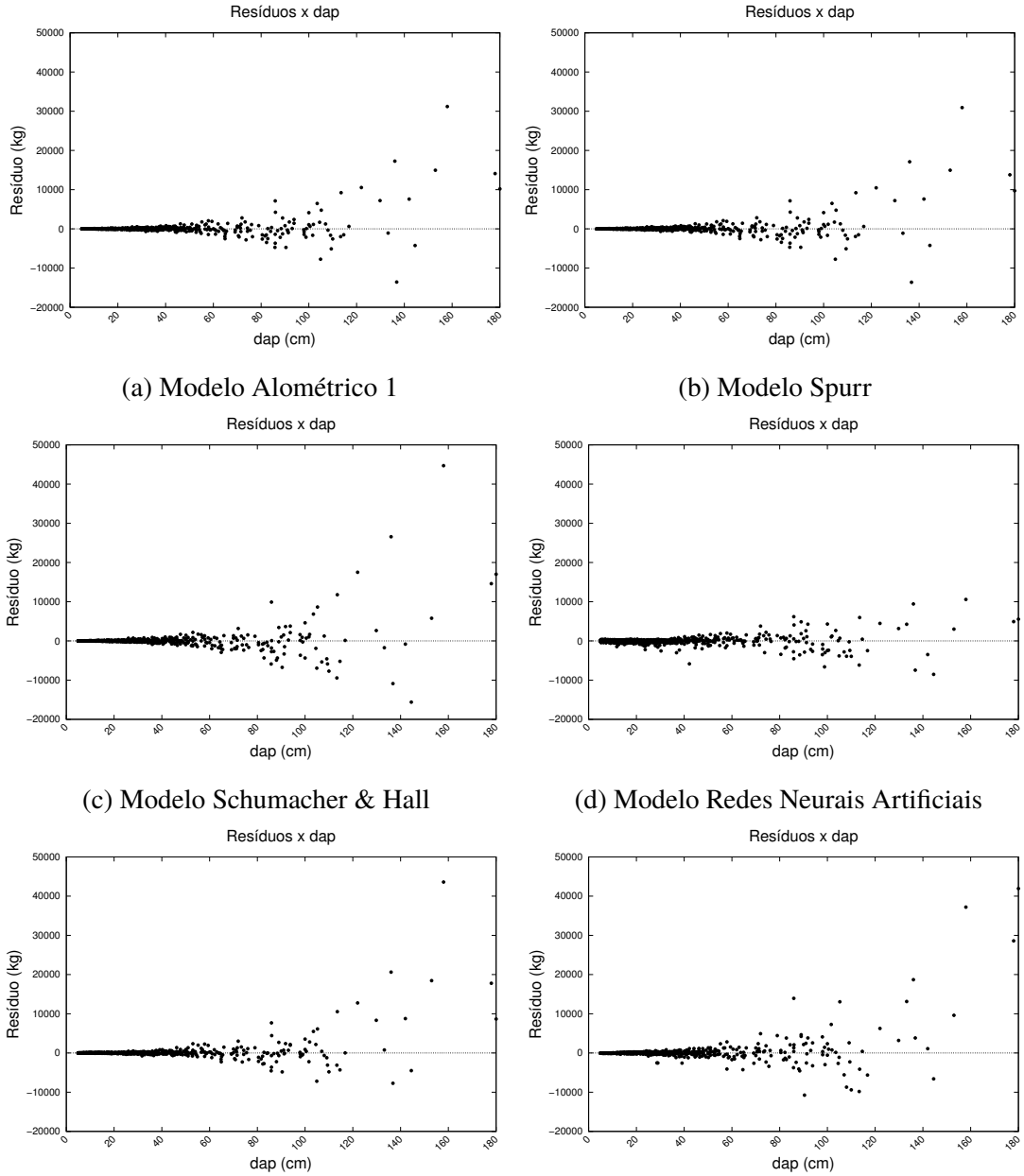
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.7: Gráfico de Resíduos: Estimativa de Biomassa de Florestas Tropicais

A Figura 4.8 apresenta os gráficos de resíduos por *dap*, gerados para todos os modelos testados. Percebe-se por estes gráficos que os resíduos são muito maiores em árvores com *dap* maior, para todos os modelos testados.



(a) Modelo Alométrico 1

(b) Modelo Spurr

(c) Modelo Schumacher & Hall

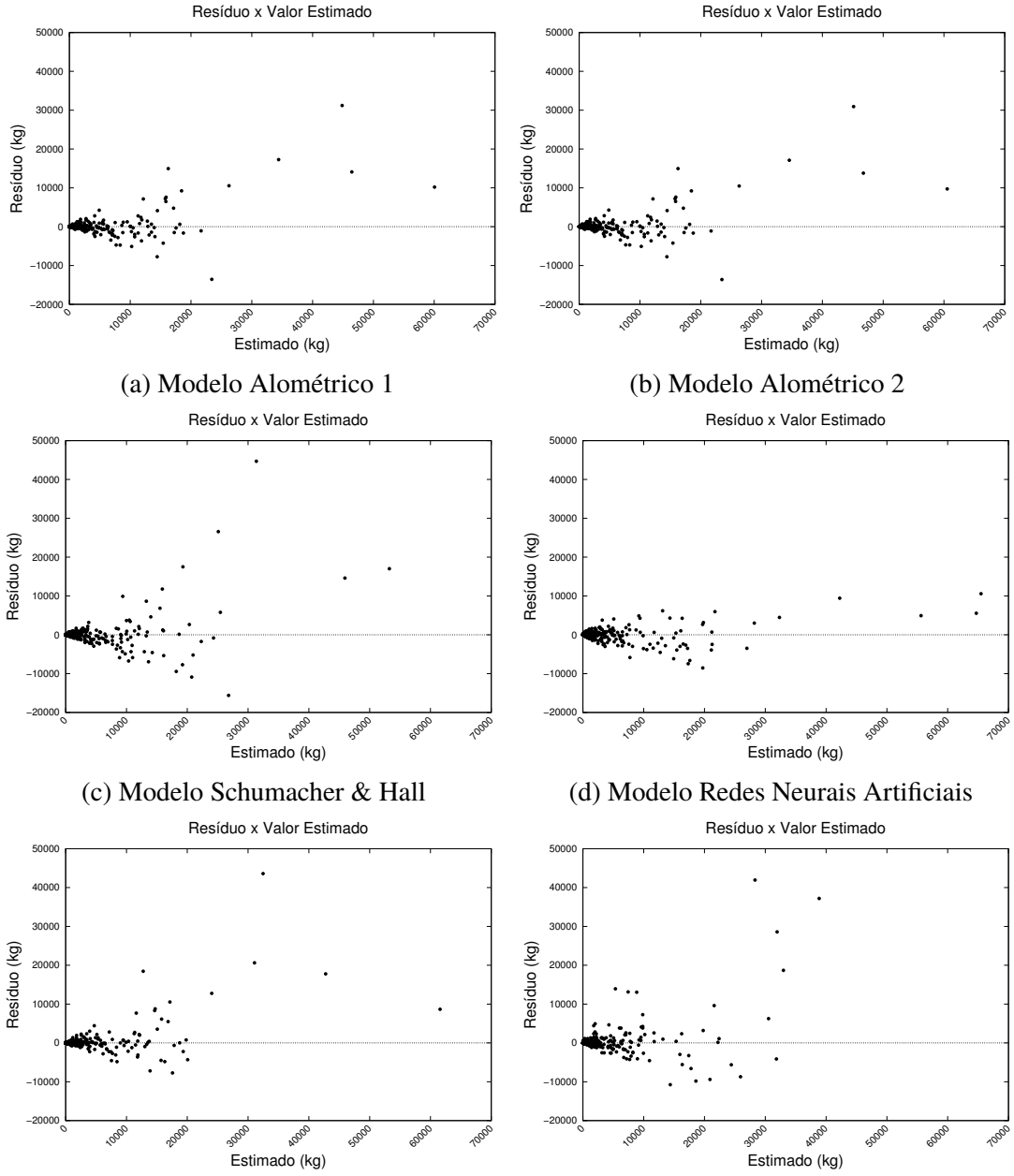
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.8: Gráfico de Resíduos por *dap*: Estimativa de Biomassa de Florestas Tropicais

A Figura 4.9 apresenta os gráficos de resíduos por ajustes, gerados para todos os modelos testados. Estes gráficos são usados para analisar o pressuposto de homocedasticidade dos resíduos (mesma variância dos resíduos em relação ao valor estimado). Nos modelos testados, percebe-se um padrão em forma de funil, indicando que a variância dos resíduos aumenta conforme o valor estimado aumenta, indicando a heterocedasticidade dos modelos.



(a) Modelo Alométrico 1

(b) Modelo Alométrico 2

(c) Modelo Schumacher & Hall

(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.9: Gráfico de Resíduos por Ajustes: Estimativa de Biomassa de Florestas Tropicias

A base usada neste caso de uso é complexa, pois a amplitude de valores de atributos e medições de biomassa é grande. Também, alguns indivíduos possuem biomassa muito fora dos padrões de seus atributos, o que caracteriza ruído na base, mas não pode ser desconsiderada.

Pelas estimativas apresentadas, o modelo usando redes neurais teve melhor acurácia nas estimativas. RNA apresentou correlação de 98, 06% entre a biomassa estimada e observada, na base de dados de teste. Já o modelo alométrico que conseguiu melhores resultados foi o Modelo 3.3, com 96, 19% de correlação.

4.4 Relação Hipsométrica da Acácia-negra

Para os experimentos de relação hipsométrica com a base da Acácia-negra foram usados os modelos Logarítmico, Parabólico e de Prodan, que podem ser vistos na Tabela 4.16. Estes modelos foram escolhidos pois em experimentos realizados em [Caldeira et al., 2002], foram os que obtiveram melhores estimativas dentre muitos outros, exatamente sobre uma base de dados da Acácia-negra.

Tabela 4.16: Coeficientes ajustados dos modelos alométricos para relação hipsométrica de Acácia-negra

Nome	Modelo	β_0	β_1	β_2	β_3
Logarítmico	$\ln h = \beta_0 + \beta_1(\frac{1}{dap}) + \beta_2(\frac{1}{Idade}) + \beta_3(\frac{1}{dap*Idade})$	3,3556	-5,7301	-0,92034	2,9994
Parabólico	$h = \beta_0 + \beta_1 dap + \beta_2 * dap^2$	-0,3392	1,7631	-0,03771	
Prodan	$h = \frac{dap^2}{\beta_0+\beta_1*\ln dap+\beta_2*dap^2}$	0,4064	0,4432	0,0258	

O treinamento das RNAs, SVMs e RF foram baseados em ajustes de parâmetros, que foram variados conforme apresentados na Tabela 4.17, Tabela 4.18 e Tabela 4.19.

Tabela 4.17: Parâmetros de Treinamento das RNAs para Relação Hipsométrica de Acácia-negra

Parâmetro	Faixa	Saltos
Camadas Ocultas	1	
Neurônios na Camada Oculta	1 a 100	1
Taxa de Aprendizado	0,1 a 0,9	0,1
Momentum	0,1 a 0,9	0,1
Conjunto de Validação	20%	
Número de Épocas	2000	

Tabela 4.18: Parâmetros de Treinamento das SVMs para Relação Hipsométrica de Acácia-negra

Parâmetro	Faixa	Saltos
Custo (C)	1 a 1000	100, análise de sensibilidade
Gamma	0,1 a 0,9	0,1
	0,01 a 0,09	0,01

Tabela 4.19: Parâmetros de Treinamento das RFs para Relação Hipsométrica de Acácia-negra

Parâmetro	Faixa	Saltos
Quantidade de Árvores	2 a 200	1
Quantidade de Atributos	2 a 4	1

A melhor RNA obtida foi com Taxa de Aprendizado de 0,1, Momentum de 0,1 e 38 neurônios na camada oculta. A SVM que obteve melhor correlação foi obtida com C de 300 e *gamma* de 0,4. A melhor RF foi obtida com 100 árvores e 2 atributos tomados de forma aleatória.

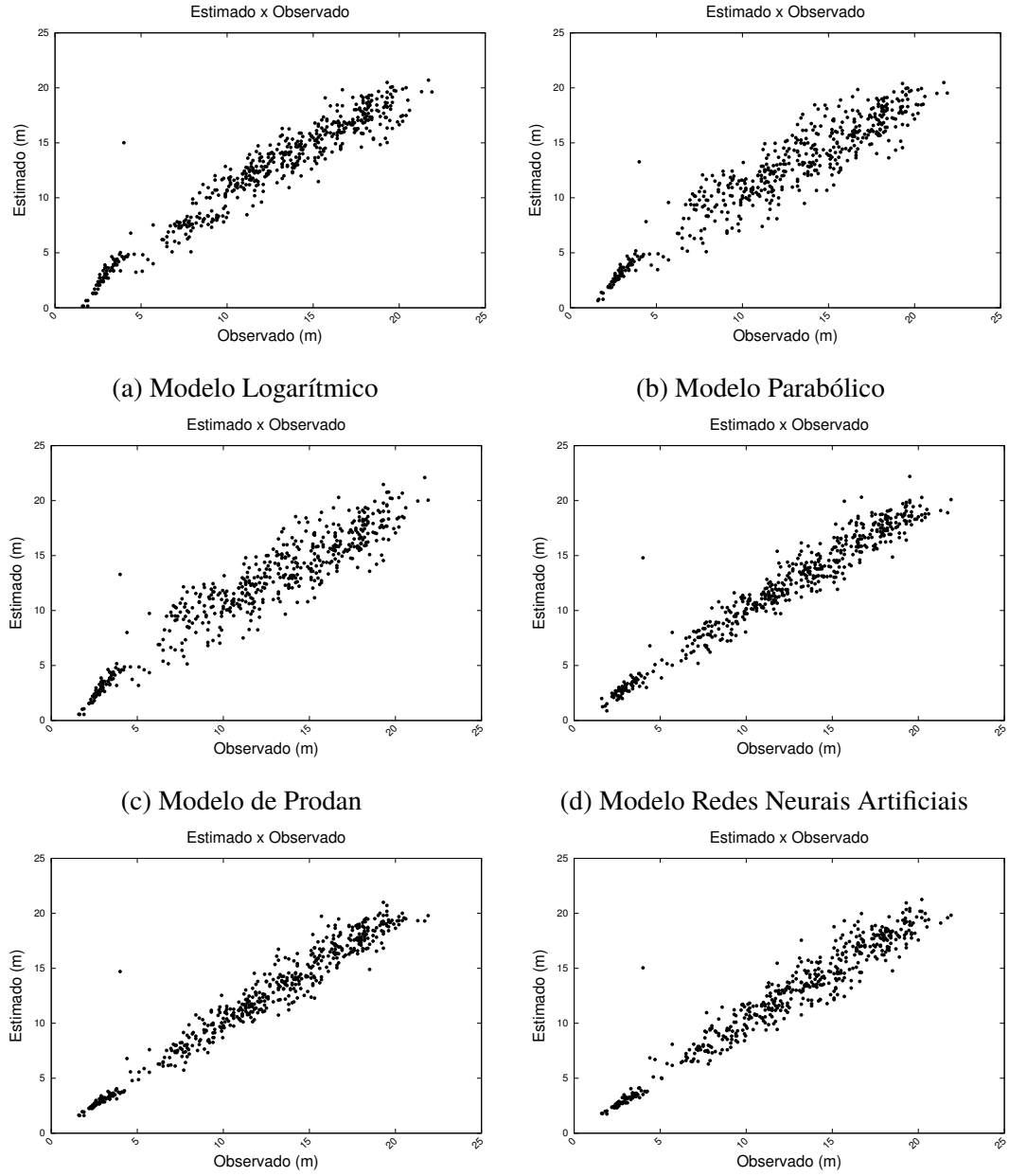
A Tabela 4.20 apresenta os resultados de Correlação, R^2, S_{yx}% e SQRes, para os dados de relação hipsométrica da Acácia-negra aplicados à base de dados usando-se validação cruzada em todos os métodos aplicados, alométricos ou de aprendizado de máquina.

Tabela 4.20: Resultados de Correlação, R^2, S_{yx} e Soma dos quadrados dos resíduos aplicados à base de relação hipsométrica da Acácia-negra usando-se validação cruzada. As células marcadas contém os melhores resultados.

Modelo	Correlação	R^2	S_{yx}%	SQRes
Logarítmico	0,9663	0,9336	10,9142	9,3142E+02
Parabólico	0,9444	0,8920	13,9225	1,5156E+03
Prodan	0,9432	0,8896	14,0759	1,5492E+03
Redes Neurais Artificiais	0,9742	0,9484	9,6111	7,2366E+02
Máquinas de Vetores de Suporte	0,9773	0,9547	9,0045	6,3520E+02
Florestas Aleatórias	0,9708	0,9422	10,1695	8,1019E+02

Observa-se na Tabela 4.20 que os três modelos de AM foram superiores ao melhor modelo alométrico, nesta base sendo o Logarítmico. Apesar disso, mesmo a diferença em termos de correlação sendo pequena, ao se observar a soma dos quadrados dos resíduos percebe-se que as estimativas dos modelos de AM são melhores.

A Figura 4.10 apresenta os gráficos de valores observados versus estimados, gerados para todos os modelos testados. Percebe-se que todos os modelos apresentam pontos próximos à reta de regressão, sem a indicação de tendências nas estimativas. Como esperado, os três melhores modelos (todos modelos de AM) possuem os pontos mais concentrados na linha de regressão, indicando melhores estimativas. O modelo alométrico logarítmico é o que mais se aproxima dos modelos de AM, em termos de resíduos.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

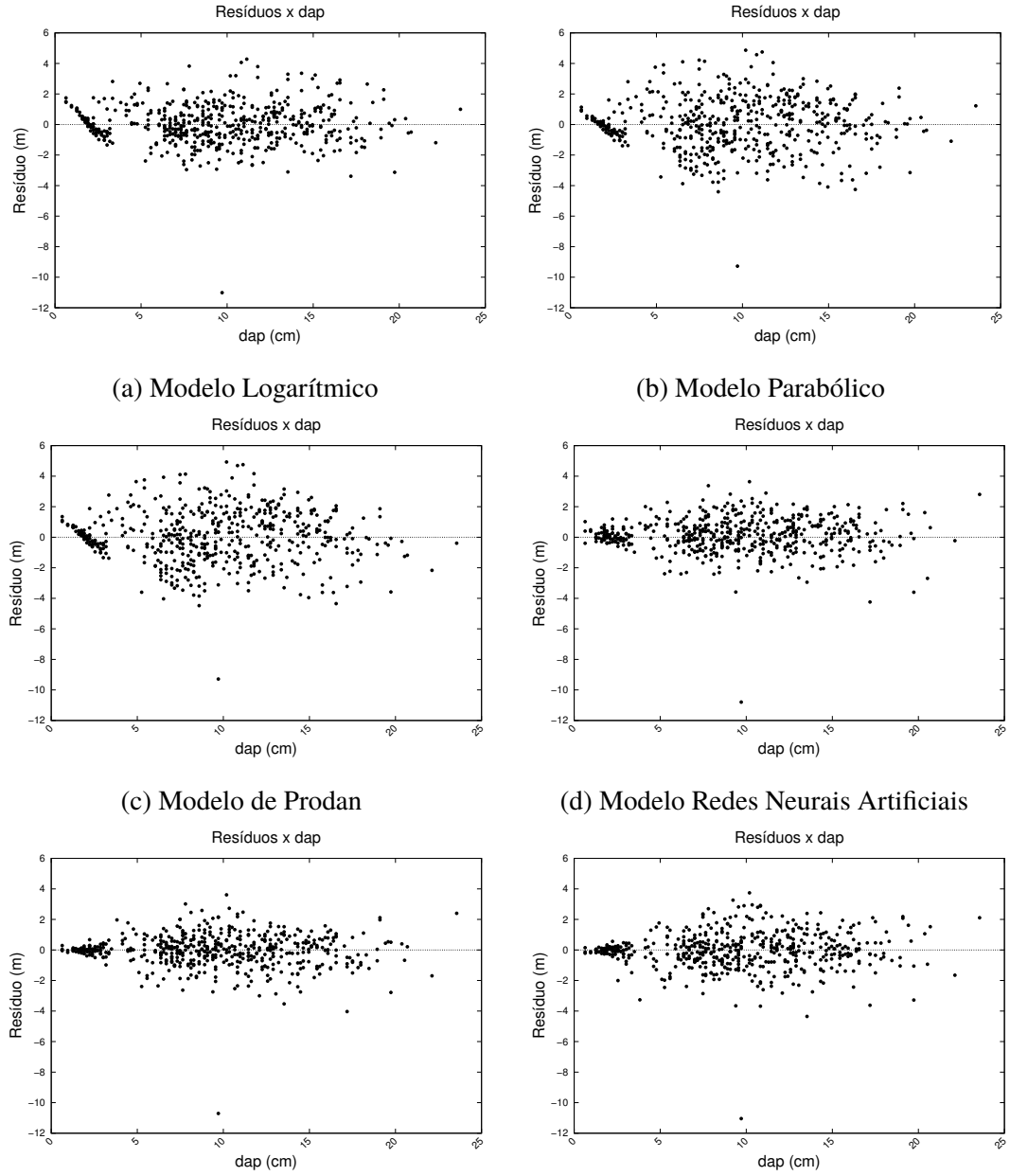
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.10: Gráfico de Resíduos: Relação Hipsométrica da Acácia-negra

A Figura 4.11 apresenta os gráficos de resíduos por *dap*, gerados para todos os modelos testados. Percebe-se por estes gráficos que, apesar dos resíduos serem maiores em árvores com *dap* maior, a dispersão dos valores em relação ao *dap* parece uniforme, indicando falta de tendenciosidade.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

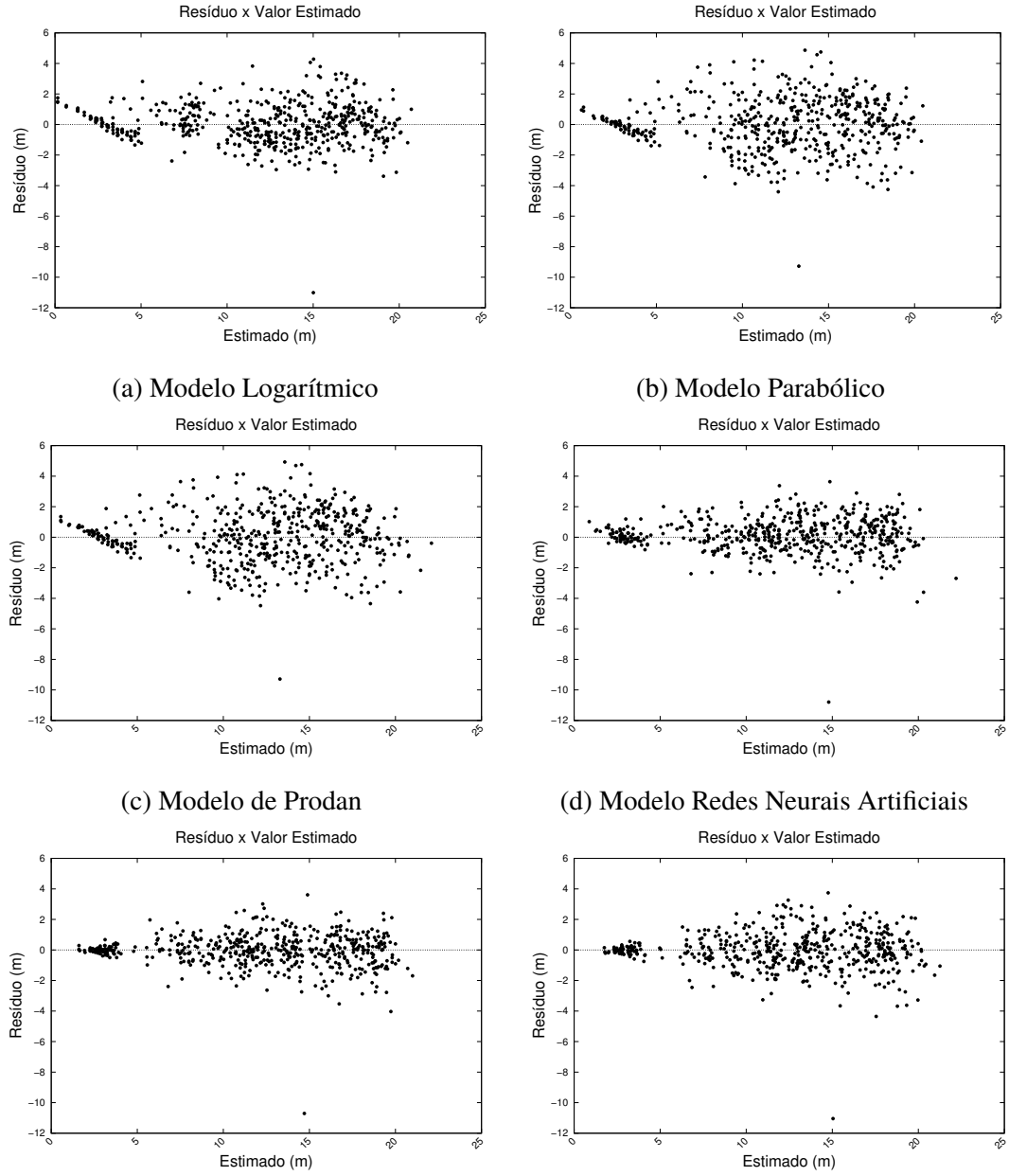
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.11: Gráfico de Resíduos por *dap*: Relação Hipsométrica da Acácia-negra

A Figura 4.12 apresenta os gráficos de resíduos por ajustes, gerados para todos os modelos testados. Estes gráficos são usados para analisar o pressuposto de homocedasticidade dos resíduos (mesma variância dos resíduos em relação ao valor estimado). Nos modelos testados, percebe-se um padrão em forma de funil, mais acentuado nos modelos alométricos parabólico e Prodan, indicando que a variância dos resíduos aumenta conforme o valor estimado aumenta (heterocedasticidade) destes modelos. Já nos demais modelos, este aspecto é menos acentuado, indicando uma variância mais uniforme.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.12: Gráfico de Resíduos por Ajuste: Relação Hipsométrica da Acácia-negra

Para esta base de dados, o modelo treinado usando Máquinas de Vetores de Suporte obteve a maior correlação (0, 9773), maior R^2 (0, 9547), menor $S_{yx}\%$ (9, 0045) e menor soma dos quadrados dos resíduos ($6,3520E+02$). Com os melhores resultados em todas as medidas efetuadas, conclui-se que para esta base de dados o modelo obtido de SVM é o melhor. Mesmo assim, os valores são muito próximos ao obtido pelo modelo de Redes Neurais Artificiais.

4.5 Relação Hipsométrica de Pinus

Para os experimentos de relação hipsométrica com a base de Pinus foram usados os modelos Logarítmico, Parabólico e de Prodan, que podem ser vistos na Tabela 4.21. Estes modelos foram usados por terem sido os mesmos nos experimentos com acácia-negra.

Tabela 4.21: Coeficientes ajustados dos modelos alométricos para relação hipsométrica de Pinus

Nome	Modelo	β_0	β_1	β_2	β_3
Logarítmico	$\ln h = \beta_0 + \beta_1(\frac{1}{dap}) + \beta_2(\frac{1}{Idade}) + \beta_3(\frac{1}{dap*Idade})$	3,9524	-10,4204	-8,2969	36,5726
Parabólico	$h = \beta_0 + \beta_1 dap + \beta_2 * dap^2$	2,0095	0,7269	-0,0033	
Prodan	$h = \frac{dap^2}{\beta_0+\beta_1*\ln dap+\beta_2*dap^2}$	-1,4853	1,1723	0.0104	

O treinamento das RNAs, SVMs e RF foram baseados em ajustes de parâmetros, que foram variados conforme apresentados na Tabela 4.22, Tabela 4.23 e Tabela 4.24.

Tabela 4.22: Parâmetros de Treinamento das RNAs para Relação Hipsométrica de *Pinus taeda*

Parâmetro	Faixa	Saltos
Camadas Ocultas	1	
Neurônios na Camada Oculta	1 a 100	1
Taxa de Aprendizado	0,1 a 0,9	0,1
Momentum	0,1 a 0,9	0,1
Conjunto de Validação	20%	
Número de Épocas	2000	

Tabela 4.23: Parâmetros de Treinamento das SVMs para Relação Hipsométrica de *Pinus taeda*

Parâmetro	Faixa	Saltos
Custo (C)	1 a 1000	100, análise de sensibilidade
Gamma	0,1 a 0,9	0,1

Tabela 4.24: Parâmetros de Treinamento das RFs para Relação Hipsométrica de *Pinus taeda*

Parâmetro	Faixa	Saltos
Quantidade de Árvores	2 a 200	1
Quantidade de Atributos	2	

A melhor RNA obtida foi com Taxa de Aprendizado de 0,1, Momentum de 0,5 e 4 neurônios na camada oculta. A SVM que obteve melhor correlação foi obtida com C de 200 e *gamma* de 0,1. A melhor RF foi obtida com 105 árvores.

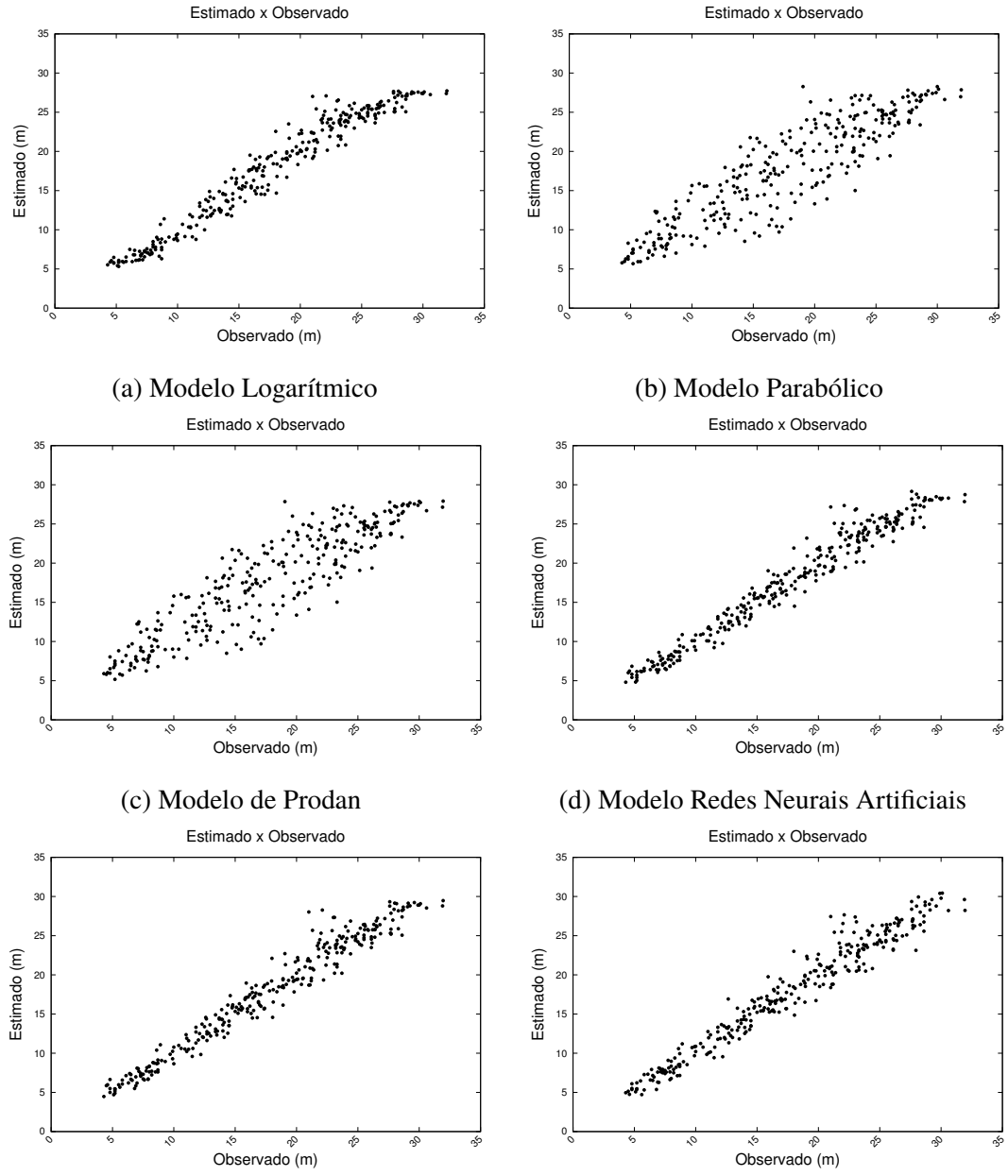
A Tabela 4.25 apresenta os resultados de Correlação, R^2, S_{yx}% e SQRes, para os dados de relação hipsométrica de Pinus aplicados à base de dados usando-se validação cruzada em todos os métodos aplicados, alométricos ou de aprendizado de máquina.

Tabela 4.25: Resultados de Correlação, R^2, S_{yx} e Soma dos quadrados dos resíduos aplicados à base de relação hipsométrica de Pinus usando-se validação cruzada. As células marcadas contém os melhores resultados.

Modelo	Correlação	R^2	S_{yx}%	SQRes
Logarítmico	0,9764	0,9533	8,7630	7,2008E+02
Parabólico	0,9010	0,8117	17,5887	2,9010E+03
Prodan	0,9014	0,8124	17,5554	2,8900E+03
Redes Neurais Artificiais	0,9802	0,9607	8,0227	6,0557E+02
Máquinas de Vetores de Suporte	0,9799	0,9598	8,1123	6,1916E+02
Florestas Aleatórias	0,9772	0,9547	8,6105	6,9755E+02

Observa-se na Tabela 4.25 que os três modelos de AM foram superiores ao melhor modelo alométrico, nesta base sendo o Logarítmico.

A Figura 4.13 apresenta os gráficos de valores observados versus estimados, gerados para todos os modelos testados. Percebe-se que todos os modelos apresentam pontos próximos à reta de regressão, sem a indicação de tendências nas estimativas. Como esperado, os três melhores modelos (todos modelos de AM) possuem os pontos mais concentrados na linha de regressão, indicando melhores estimativas.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

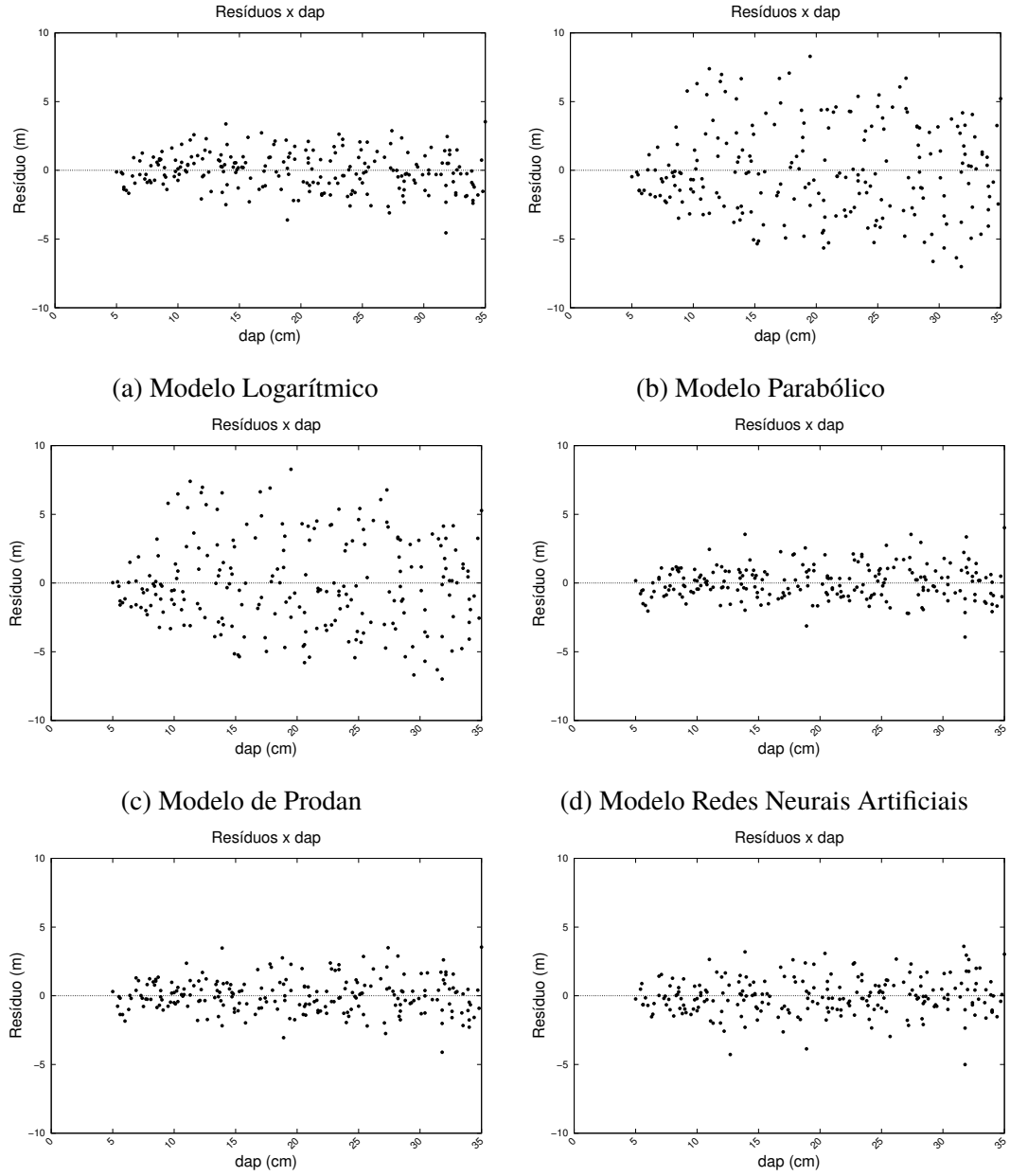
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.13: Gráfico de Resíduos: Relação Hipsométrica de Pinus

A Figura 4.14 apresenta os gráficos de resíduos por *dap*, gerados para todos os modelos testados. Em árvores com *dap* menor o resíduo é menor, aumentando para árvores maiores. A partir de um determinado *dap*, observando-se o gráfico, próximo a 5cm, o resíduo aparenta não apresentar tendenciosidade.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

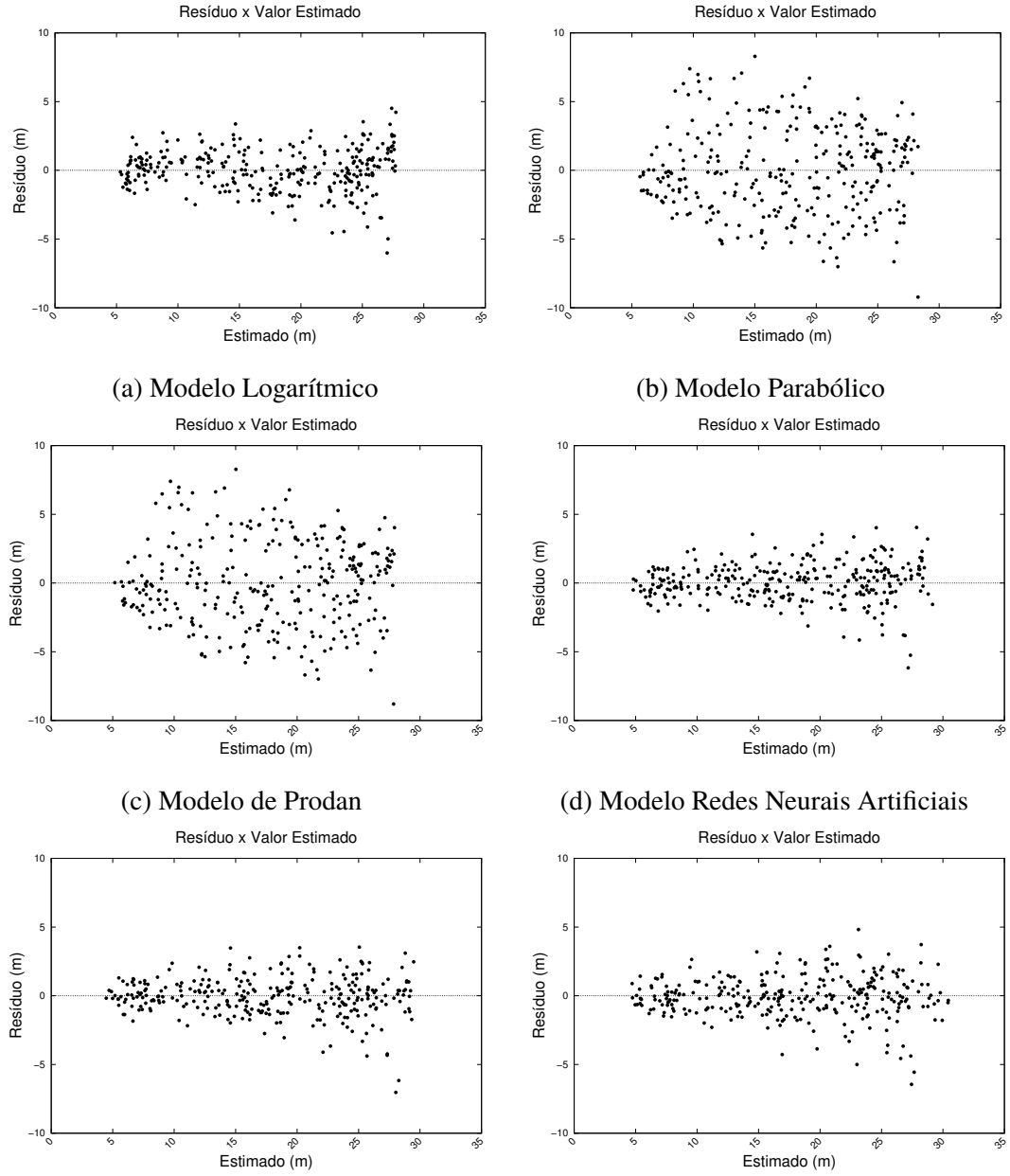
(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.14: Gráfico de Resíduos por *dap*: Relação Hipsométrica da Pinus

A Figura 4.15 apresenta os gráficos de resíduos por ajustes, gerados para todos os modelos testados. Estes gráficos são usados para analisar o pressuposto de homocedasticidade dos resíduos (mesma variância dos resíduos em relação ao valor estimado). Nos modelos testados, percebe-se um padrão em forma de funil em árvores com valor estimado baixo. Em valores maiores, o resíduo parece não apresentar qualquer padrão, indicando mesma variância.



(a) Modelo Logarítmico

(b) Modelo Parabólico

(c) Modelo de Prodan

(d) Modelo Redes Neurais Artificiais

(e) Modelo Máquinas de Vetores de Suporte

(f) Modelo Florestas Aleatórias

Figura 4.15: Gráfico de Resíduos por Ajuste: Relação Hipsométrica de Pinus

Para esta base de dados, o modelo treinado usando RNA obteve a maior correlação $(0,9802)$, maior R^2 $(0,9607)$, menor $S_{yx}\%$ $(8,0227)$ e menor soma dos quadrados dos resíduos $(6,0557E+02)$. Com os melhores resultados em todas as medidas efetuadas, conclui-se que para esta base de dados o modelo obtido com RNA é o melhor. Apesar dos valores serem muito próximos ao obtido pelo modelo de Redes Neurais Artificiais, no caso deste experimentos os três modelos de AM obtiveram resultados superiores aos alométricos, indicando a superioridade do método.

4.6 Análise Estatística dos Resultados

Um dos objetivos deste trabalho é apresentar a mensuração florestal como um processo de KDD. Assim sendo, somente analisar os resultados, isoladamente, não informa se as técnicas de AM resultam em valores estatisticamente próximos, melhores ou piores.

Nos experimentos realizados aqui, são feitas análises estatísticas por meio do teste de Friedman, com nível de significância de $\alpha = 5\%$, $K = 4$ (quantidade de algoritmos) e $N = 5$ (número de conjuntos de dados), para cada uma das medidas de qualidade utilizadas: Correlação, Coeficiente de Determinação, Erro Padrão da Estimativa e Soma dos Quadrados dos Resíduos. Quando são constatadas diferenças estatísticas, o pós-teste de Nemenyi é aplicado com $\alpha = 5\%$. O valor tabulado de q_α é $2,569$ e a diferença crítica observada é de $CD = 2,1$.

Neste trabalho, como foram observados ranqueamentos iguais para as quatro medidas de qualidade adotadas (correlação, R^2, S_{yx} e soma dos quadrados dos resíduos), e como o teste de Friedman é baseado neste ranqueamento, somente foi aplicado o teste sobre a correlação, pois a aplicação nas demais medidas chegariam ao mesmo resultado.

A Tabela 4.26 apresenta os resultados de correlação consolidados dos experimentos no formato *correl (rank)*, onde *correl* representa a correlação do melhor modelo encontrado para aquela base de dados, e *rank* é a posição relativa do algoritmo em relação aos demais, para uma determinada base de dados. As bases utilizadas para o cálculo da correlação são as bases de teste, em cada domínio. As últimas três linhas representam, para cada modelo, os valores de acurácia média, rank médio, e quantidade de vezes que apresentou a melhor acurácia em um conjunto de dados (1 vs. todos), respectivamente.

Tabela 4.26: Resultados dos experimentos realizados. Os valores estão no formato *correl (rank)*, onde *correl* é a correlação obtida pelo modelo e *rank* é a posição relativa do modelo em relação aos demais na base de dados testadas. As células destacadas indicam os melhores valores de correlação para cada base.

Base/Modelo	Alométrico	RNA	SVM	RF
Pinus	0,9911 (3)	0,9914 (2)	0,9919 (1)	0,9904 (4)
Acacia	0,983 (3)	0,9831 (2)	0,986 (1)	0,9782 (4)
Florestas Tropicais	0,9619 (2)	0,9806 (1)	0,9597 (3)	0,9025 (4)
Hipsométrica Acácia	0,9663 (4)	0,9742 (2)	0,9773 (1)	0,9708 (3)
Hipsométrica Pinus	0,9764 (4)	0,9802 (1)	0,9799 (2)	0,9772 (3)
Correlação Média	0,9757	0,9819	0,979	0,9638
Rank Médio	3,2	1,6	1,6	3,6
1 vs Todos	0	2	3	0

Por meio da aplicação do teste de Friedman sobre os valores de *rank* médio descritos na Tabela 4.26 a hipótese nula de que todos os modelos são equivalentes, em termos de correlação, não foi aceita, com valores de $F_F = 7,9048$ e $p - valor = 0,0036$, indicando que há diferença significativa.

Para tentar identificar essa diferença, aplicou-se o pós-teste de Nemenyi obtendo-se o diagrama de diferença crítica apresentado na Figura 4.16.

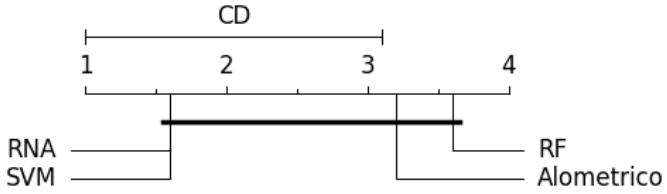


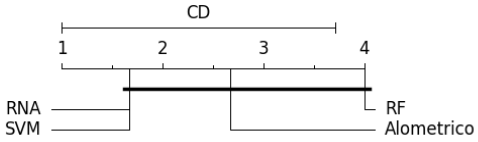
Figura 4.16: Diagrama de diferença crítica para o pós-teste de Nemenyi. O eixo x representa o *rank* médio de cada método, linhas abaixo, conectadas, represetam métodos sem diferença estatística com nível de significância de 95%.

Fonte: O autor.

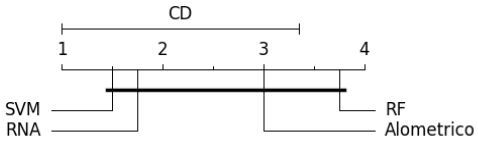
Neste diagrama, o eixo x representa o *rank* médio obtido por cada método. As linhas abaixo do eixo x conectam métodos que não possuem diferença estatisticamente significativa com nível de confiança de 95%. A diferença crítica (CD) é mostrada acima do eixo x.

A partir deste diagrama, percebe-se que o pós-teste de Nemenyi não foi capaz de apontar as diferenças entre os métodos, apesar do resultado do teste de Friedman evidenciar tais diferenças. Mesmo assim, percebe-se que os métodos nas extremidades são provavelmente os que apresentem estas diferenças, neste caso, RNA e SVM contra modelos alométricos e RF.

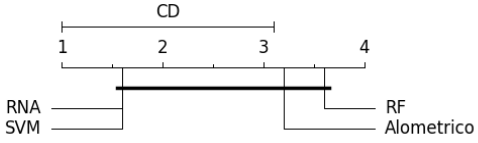
Outro aspecto a ser apresentado é que o teste de Friedman e pós-teste de Nemenyi são sensíveis à quantidade de bases de dados testadas. No caso deste trabalho somente cinco bases estavam disponíveis para treinamento, o que pode ter acobertado a diferença estatística no pós-teste aplicado. Então, analisando o gráfico de diferença crítica, percebe-se uma evolução no distanciamento entre os métodos quando se compara a análise para três, quatro e cinco bases de dados. A Figura 4.17 apresenta esta evolução, indicando a possibilidade de que o pós-teste de Nemenyi possa evidenciar a diferença estatística entre os métodos quando houver mais bases de dados disponíveis.



(a) Diagrama de diferença crítica para três bases de dados testadas



(b) Diagrama de diferença crítica para quatro bases de dados testadas



(c) Diagrama de diferença crítica para todas as bases de dados testadas

Figura 4.17: Evolução das Diferenças Críticas para três, quatro e cinco bases de dados. Percebe-se que o distanciamento entre os métodos de RNA e SVM aumenta, se comparados com RF e alométricos.

Fonte: O autor.

Isto posto, com os experimentos e bases de dados aqui disponíveis, não se pode evidenciar a diferença estatística através do pós-teste de Nemenyi, apesar do teste de Friedman tê-la apresentado. Mesmo assim há indícios de que experimentos mais extensos, com várias bases distintas, podem evidenciar esta diferença. Convém ressaltar que em todas as bases testadas os métodos AM obtiveram melhores resultados de estimação, o que os coloca em foco na hora da escolha.

Capítulo 5

Conclusões

A aplicação de técnicas de AM na área florestal é um tema que já foi explorado em vários artigos e trabalhos, conforme pode ser visto no Capítulo 2. Apesar disso, até o presente trabalho, este tema era visto como sendo simples aplicação de algoritmos em bases de dados.

Para permitir o amadurecimento desta nova metodologia, o processo de mensuração florestal foi modelado como um processo de KDD, como apresentado no Capítulo 3. As fases de escolha dos modelos e ajuste das equações, executadas quando do uso de regressão, foram substituídas pela escolha de técnicas de AM e treinamento dos modelos. As estatísticas para decisão de qual modelo deve ser escolhido foram mantidas as mesmas usadas pela área florestal, a saber: correlação, R^2, S_{yx}, soma dos quadrados dos resíduos e análise gráfica dos resíduos.

Isto posto, o presente trabalho consolida a interdisciplinaridade entre as áreas, indicando que todo avanço feito na área de aprendizado de máquina, voltado à regressão, pode e deve ser explorado em mensuração florestal, confrontando os resultados com os modelos alométricos comumente usados.

Como pode ser visto no Capítulo 2, as Redes Neurais Artificiais são os modelos mais difundidos, possuem baixo tempo de treinamento e entende-se que haveria mais facilidade por parte dos engenheiros em entender seus mecanismos.

Já as Florestas Aleatórias, apesar de possuirem um tempo de treinamento extremamente baixo, se comparados às RNAs e SVMs, e poucos parâmetros de ajuste, não são tão difundidos e o algoritmo de indução das árvores pode não ser de fácil compreensão para um público acostumado primariamente com o processo de regressão.

As Máquinas de Vetores de Suporte possuem uma teoria subjacente mais complexa e em bases de dados maiores gera um tempo de treinamento muito maior que as demais técnicas. A exemplo que foi observado neste trabalho, para a base de florestas tropicais, quando o parâmetro de custo (C) aumenta, o tempo de treinamento de um modelo torna o processo não instantâneo, chegando a demorar várias semanas. Levando-se em conta que vários modelos precisam ser treinados e comparados, este processo pode se tornar inviável para se obter resultados imediatos.

No Capítulo 4 foram mostrados os resultados dos experimentos efetuados, confrontando os modelos alométricos com os modelos de AM. Foram estimados o volume e relação hipsométrica de uma base de dados de *Pinus taeda*, biomassa seca e relação hipsométrica de acácia-negra e biomassa de florestas tropicais.

Pode-se observar que em todos os domínios testados os modelos gerados através de Redes Neurais e Máquinas de Vetores de Suporte tiveram os melhores resultados, se comparados aos melhores modelos alométricos. As condições de testes foram mantidas as mesmas, isto é, mesmas bases de treinamento e teste para os experimentos com florestas tropicais, e o uso de validação cruzada para os experimentos com pinus e acácia-negra.

O teste de Friedman apontou diferença estatística entre os métodos e foi discutido no Capítulo 4 que esta diferença ocorre entre os métodos RNA e SVM comparados com RF. Apesar disso, pelo fato do teste de Friedman e pós-teste de Nemenyi serem sensíveis à quantidade de bases de dados testadas, e pelo fato de estarem disponíveis aqui somente três bases de dados, estima-se que em testes com mais bases esta diferença fique mais evidenciada, indicando até diferença estatística entre os modelos de AM e alométricos.

Pode-se observar por este trabalho que as técnicas de aprendizado de máquina podem substituir os modelos alométricos que são classicamente utilizados na área de mensuração florestal e, através dos experimentos aqui realizados, percebe-se que para estas bases de dados, os resultados foram superiores, evidenciando que as técnicas de AM devem ser consideradas pelos engenheiros florestais de forma séria, como uma alternativa viável e segura à regressão.

Assim, além de consolidar as duas áreas como afins, este trabalho apresenta técnicas de AM com resultados melhores em seus experimentos, reforçando o fato de que a regressão, apesar de simples no seu manuseio e ajuste, pode não apresentar a melhor solução para mensuração florestal.

Assim, o objetivo principal desta tese foi atingido, apresentando a aplicação de técnicas de AM na área de Mensuração Florestal, abrindo uma vasta gama de possibilidades de pesquisa para os engenheiros florestais em termos de novas técnicas a serem treinadas. Para os cientistas da computação, as bases de dados com peculiaridades da área florestal poderão ser exploradas no melhoramento ou desenvolvimento de novas técnicas.

Também foi mostrado que a mensuração é um processo de KDD, onde a fase de extração de padrões pode ser substituída pelo treinamento de técnicas de AM, ao invés do uso de modelos alométricos. Neste contexto, as técnicas de AM foram confrontadas com modelos de regressão nos experimentos aqui realizados. Como resultado, observou-se que os modelos de AM apresentaram melhores estimativas, sendo que o teste de Friedman apontou diferença estatítsica entre eles, mesmo o pós-teste de Nemenyi não sendo capaz de evidenciar esta diferença.

Apesar dos benefícios aqui apontados, espera-se que os engenheiros florestais, por não serem familiarizados com ferramentas especializadas, como o WEKA, tenham dificuldade no treinamento dos modelos. A exemplo dos procedimentos aqui realizados, algumas SVMs treinadas para a base de florestas tropicais demoraram várias semanas para serem obtidas, o que, a princício, parece inviável para se colocar em prática no dia-a-dia da indústria florestal. Já os modelos de RNA e RF possuem treinamento mais rápido, o que contribui com o fato das RNAs serem mais difundidas.

Mesmo assim, para se obter RNAs realmente melhores foram necessários vários treinamentos, neste caso, com uma camada oculta com neurônios variando de 1 a 100, taxa de aprendizado variando de 0,1 a 0,9 e momentum variando de 0,1 a 0,9, resultou em 8100 redes treinadas a título de comparação. Apesar do tempo de treinamento ser menor, bastante poder computacional foi necessário para se treinar todas estas redes, várias em paralelo.

Fica claro, portanto, que as técnicas de AM não devem ser negligenciadas pelos profissionais e pesquisadores da área de mensuração florestal, por ter sido mostrado que são alternativa competitiva e promissora frente aos métodos clássicos. Também, a engenharia florestal possui bases de dados ricas que podem ser exploradas pelos cientistas da computação para o desenvolvimento de técnicas aprimoradas e direcionadas.

5.1 Recomendações

Como recomendações de trabalhos futuros, podemos destacar:

- Implementar um aplicativo para treinamento de modelos de AM, afim de facilitar o trabalho do engenheiro florestal que não possui familiaridade com as ferramentas especializadas, como o WEKA;
- Produzir um aplicativo para *smartphones* e *tablets* para agregar ao aplicativo a mobilidade e possibilidade de uso em campo;
- Um estudo mais aprofundado das técnicas baseadas em árvores (Árvore de Modelos e Árvore de Regressão), pois nos parece que estes modelos de aprendizado de máquina possuem um potencial maior para gerar estimativas melhores, mas suas implementações devem ser alteradas para se adequar ao problema em foco;
- Estudar variações das técnicas usadas aqui, como as redes neurais RBF;
- Um estudo de novas técnicas de regressão que podem resultar em melhores estimativas, não só fazendo o uso de um agrupamento de regressores (*Ensembles*), mas também com a produção de novas técnicas, baseadas nos estudos de caso fornecidos pelos problemas da área;
- Verificar se o pressuposto da homoscedasticidade dos resíduos pode ser relaxado para modelos baseados em aprendizado de máquina.

Referências Bibliográficas

[Aha et al., 1991] Aha, D. W., Kibler, D. e Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning 6*, 6:37–66.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Em Petrov, B. N. e Csaki, F., editores, *Second International Symposium on Information Theory*, páginas 267–281, Budapest. Akadémiai Kiado.

[Araujo, 1980] Araujo, A. J. d. (1980). *Early results of provenance studies of loblolly and slash pines in Brazil*. Tese de doutorado, Michigan State University.

[Avramidis et al., 2006] Avramidis, S., Iliadis, L. e Mansfield, S. D. (2006). Wood dielectric loss factor prediction with artificial neural networks. *Wood Science and Technology*, 40(7):563–574.

[Azevedo Gomes, 1957] Azevedo Gomes, A. M. d. (1957). *Medição dos arvoredos*. A Terra e o homen; colecção de livros agricolas, 30. 3a. Secção: A exploração e a cultura des plantas: b) Culturas florestais. Livraria Sá da Costa.

[Back, 1996] Back, T. (1996). *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, Oxford, UK.

[Baima et al., 2001] Baima, A., Silva, S. e Silva, J. (2001). Equações de volume para floresta tropical de terra firme em Moju, PA. páginas 367–392.

[Baldanzi e Araujo, 1971] Baldanzi, G. e Araujo, A. (1971). Ensaio comparativo de espécies e procedências de Pinus, na estação de pesquisas florestais de Rio Negro, Paraná. *Floresta*, 3(2).

[Barrichelo et al., 1978] Barrichelo, L., Kageyama, P., Speltz, R., Bonish, H., Brito, J. e Ferreira, M. (1978). Estudos de procedências de *Pinus taeda* visando seu aproveitamento industrial.

[Batista et al., 2004] Batista, J., Marchesini, M. e Viana, V. (2004). Equações de volume para árvores de caixeta (*Tabebuia cassinoides*) no estado de São Paulo e sul do estado do Rio de Janeiro. *Scientia Forestalis, Piracicaba*, páginas 162–175.

[Behling, 2014] Behling, A. (2014). A produção de biomassa e o acúmulo de carbono em povoamentos de acácia negra em função de variáveis bioclimáticas. Dissertação de Mestrado, Universidade Federal do Paraná.

[Belchior, 1996] Belchior, P. R. M. (1996). Estimação de volume total, de fuste e de galhos em mata secundária no município de Rio Vermelho, MG. Dissertação de Mestrado, Universidade Federal de Viçosa, MG.

[Binoti et al., 2014a] Binoti, D. H. B., da Silva Binoti, M. L. M. e Leite, H. G. (2014a). Configuração de redes neurais artificiais para estimação do volume de árvores. *Ciência da Madeira*, 5:58–67.

[Binoti et al., 2014b] Binoti, D. H. B., da Silva Binoti, M. L. M. e Leite, H. G. (2014b). Configuração de redes neurais artificiais para estimação do volume de árvores. *Brazilian Journal of Wood Science*, 5(1):283–288.

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA.

[Blackard e Dean, 1999] Blackard, J. A. e Dean, D. J. (1999). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131 – 151.

[Braga et al., 2007] Braga, A., Carvalho, A. C. e Ludermir, T. B. (2007). *Redes Neurais Artificiais: Teoria e aplicações*, volume 2. LTC Editora.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

[Caldeira, 2003] Caldeira, M. (2003). *Determinação de biomassa e nutrientes em uma Floresta Ombrófila Mista Montana em General Carneiro, Paraná*.

[Caldeira et al., 2002] Caldeira, M. V. W., Schumacher, M. V., Scheeren, L. W., Barichello, L. R. e Watzlawick, L. F. (2002). Relação hipsométrica para acacia mearnsii com diferentes idades. *Boletim de Pesquisa Florestal*, 45:57–68.

[Carpanezzi, 1998] Carpanezzi, A. A. (1998). Espécies para recuperação ambiental. Em *GALVÃO, APM*, páginas 43–53, Colombo, PR. Embrapa Florestas.

[Castellanos et al., 2007] Castellanos, A., Martinez Blanco, A. e Palencia, V. (2007). Applications of radial basis neural networks for area forest. *International Journal ITA*.

[Castro et al., 2015] Castro, R. V. O., Soares, C. P. B., Leite, H., Souza, A. L., Martins, F. B., Nogueira, G. S. e Romarco, M. L. O. (2015). Artificial neural networks effectiveness to estimate mortality in a Semi-deciduous Seasonal Forest. *Australian Journal of Basic and Applied Sciences*, 9(5).

[Castro et al., 2013] Castro, R. V. O., Soares, C. P. B., Leite, H. G., de Souza, A. L., Nogueira, G. S. e Martins, F. B. (2013). Individual growth model for eucalyptus stands in Brazil using artificial neural network. *ISRN Forestry*, 2013:1–12.

[Cestnik et al., 1987] Cestnik, B., Kononenko, I. e Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. Em *EWSL*, páginas 31–45.

[Chave et al., 2014] Chave, J., Rejou-Mechain, M., Burquez, A., Chidumayo, E., Colgan, M., Delitti, W., Duque, A., Eid, T., Fearnside, P., Goodman, R., Henry, M., Martinez-Yrizar, A., Mugasha, W., Muller-Landau, H., Mencuccini, M., Nelson, B., Ngomanda, A., Nogueira, E., Ortiz-Malavassi, E., Pelissier, R., Ploton, P., Ryan, C., Saldarriaga, J. e Vieilledent, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology*, 20(10):3177–3190.

[Chichorro et al., 2003] Chichorro, J., Resende, J. e Leite, H. (2003). Equações de volume e de taper para quantificar multiprodutos da madeira em Floresta Atlântica. páginas 799–809.

[Clutter, 1983] Clutter, J. (1983). *Timber management: a quantitative approach*. Wiley.

[Corne et al., 2004] Corne, S. A., Carver, S. J., Kunin, W. E., Lennon, J. J. e van Hees, W. W. S. (2004). Predicting forest attributes in Southeast Alaska using artificial neural networks. *Forest Science*, 50(2):259–276.

[Cortes e Vapnik, 1995] Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Couto e Bastos, 1987] Couto, H. e Bastos, N. (1987). Modelos de equações de volume e relações hipsométricas para plantações de eucalyptus no estado de São Paulo. *Revista IPEF, Piracicaba*, páginas 33–44.

[Cover, 1965] Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on Electronic Computers*, EC-14:326–334.

[Cristiani e Shawe-Taylor, 2000] Cristiani, N. e Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314.

[da Fonseca et al., 1978] da Fonseca, S. M., Kageyama, P. Y., Ferreira, M. e Jacob, W. S. (1978). Síntese do programa de melhoramento genético de *Pinus* spp que vem sendo conduzido, sob a coordenação do IPEF, na região Sul do Brasil.

[da Silva et al., 2009] da Silva, M. L. M., Binoti, D. H. B., Gleriani, J. M. e Leite, H. G. (2009). Ajuste do modelo de Schumacher e Hall e aplicação de redes neurais artificiais para estimar volume de Árvores de eucalipto. *Revista Árvore*, 33:1133 – 1139.

[da Silva Binoti, 2010] da Silva Binoti, M. L. M. (2010). Redes neurais artificiais para prognose da produção de povoamentos não desbastados de eucalipto. Dissertação de Mestrado, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil.

[da Silva Binoti, 2012] da Silva Binoti, M. L. M. (2012). *Emprego de Redes Neurais Artificiais em Mensuração Florestal e Manejo Florestal*. Tese de doutorado, Universidade Federal de Viçosa.

[da Silva Binoti et al., 2014] da Silva Binoti, M. L. M., Binoti, D. H. B., Leite, H. G., Garcia, S. L. R., Ferreira, M. Z., Rode, R. e da Silva, A. A. L. (2014). Redes neurais artificiais para estimação do volume de Árvores. *Revista Árvore*, 38:283 – 288.

[da Silva Binoti et al., 2015] da Silva Binoti, M. L. M., Leite, H. G., Binoti, D. H. B. e Gleriani, J. M. (2015). Prognose em nível de povoamento de clones de eucalipto empregando redes neurais artificiais. *CERNE*, 21:97 – 105.

[Demsar, 2006] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[Diamantopoulou e Özçelik, 2012] Diamantopoulou, M. e Özçelik, R. (2012). Evaluation of different modeling approaches for total tree-height estimation in Mediterranean Region of Turkey. *Forest Systems*, 21(3):383–397.

[Diamantopoulou, 2005] Diamantopoulou, M. J. (2005). Artificial neural networks as an alternative tool in pine bark volume estimation. *Computers and electronics in agriculture*, 48(3):235–244.

[Dietterich, 1998] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923.

[Dillewijn, 1968] Dillewijn, F. J. V. (1968). Curso de dendrometria.

[Duda et al., 2000] Duda, R. O., Hart, P. E. e Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.

[Dunn e Dunn, 1961] Dunn, J. e Dunn, O. J. (1961). Multiple comparisons among means. *American Statistical Association*, páginas 52–64.

[Faceli et al., 2011] Faceli, R., Lorena, A. C., Gama, J. e de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC, Rio de Janeiro, Brasil.

[FAO – Food and Agriculture Organization, 2015] FAO – Food and Agriculture Organization (2015). Global forest resources assessments 2015. Relatório técnico, Roma, Itália.

[Fayyad et al., 1996a] Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. (1996a). Knowledge discovery and data mining: Towards a unifying framework. Em *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, páginas 82–88, Menlo Park, USA.

[Fayyad et al., 1996b] Fayyad, U. M., Piatetsky-Shapiro, G. e Smyth, P. (1996b). Advances in knowledge discovery and data mining. capítulo: From Data Mining to Knowledge Discovery: An Overview, páginas 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Fernandes et al., 1983] Fernandes, N., Jardim, F. e Higuchi, N. (1983). Tabelas de volume para a Floresta de Terra-firme da Estação Experimental de Silvicultura Tropical. *Acta Amazonica*, 13:537–545.

[Gorgens et al., 2014] Gorgens, E. B., Leite, H. G., Gleriani, J. M., Soares, C. P. B. e Ceolin, A. (2014). Influência da arquitetura na estimativa de volume de Árvores individuais por meio de redes neurais artificiais. *Revista Árvore*, 38:289 – 295.

[Gorgens et al., 2009] Gorgens, E. B., Leite, H. G., Santos, H. d. N. e Gleriani, J. M. (2009). Estimação do volume de árvores utilizando redes neurais artificiais. *Revista Árvore*, 33:1141 – 1147.

[Grisby, 1977] Grisby, H. (1977). A 16-year provenance test of loblolly pine in southern arkansas. Em *Southern Forest Tree Improvement Conference*.

[Guan e Gertner, 1991] Guan, B. T. e Gertner, G. (1991). Modeling red pine tree survival with an artificial neural network. *Forest Science*, 37(5):1429–1440.

[Hastie et al., 2003] Hastie, T., Tibshirani, R. e Friedman, J. (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, corrected edition.

[Haykin, 1999] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, USA.

[Haykin, 2001] Haykin, S. S. (2001). *Redes Neurais: Princípios e Prática*. Bookman.

[Hearst et al., 1998] Hearst, M. A., Schölkopf, B., Dumais, S., Osuna, E. e Platt, J. (1998). Trends & controversies: Innovations in electronic academic publishing. *IEEE Intelligent Systems*, 13(1):6–13.

[Herbrich, 2001] Herbrich, R. (2001). *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, MA, USA.

[Higuchi e Ramm, 1985] Higuchi, N. e Ramm, W. (1985). Developing bole wood volume equations for a group of tree species of central amazon (Brazil). *The Commonwealth Forestry Review*, 64(1 (198)):33–41.

[Hocker, 1956] Hocker, H. W. (1956). Certain aspects of climate as related to the distribution of loblolly pine. *Ecology*, 37(4):824–834.

[Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79(8):2554–2558.

[Hush et al., 2002] Hush, B., Beers, T. W. e Kershaw Jr, J. A. (2002). *Forest Mensuration*. Wiley.

[Iman e Davenport, 1980] Iman, R. L. e Davenport, J. M. (1980). Approximations of the critical region of the fbietkan statistic. *Communications in Statistics - Theory and Methods*, 9(6):571–595.

[Jorge, 1982] Jorge, L. (1982). Equações de volume comercial com casca em floresta tropical pluvial no norte do Espírito Santo. *Silvicultura*, páginas 456–467.

[Karalic, 1992] Karalic, A. (1992). Linear regression in regression tree leaves. Em *Proceedings of ECAI-92*, páginas 440–441. John Wiley & Sons.

[Karalič e Cestnik, 1991] Karalič, A. e Cestnik, B. (1991). The bayesian approach to tree-structured regression. Em *In Proceedings of ITI-91*. Universisty of Zagreb, Croatia.

[Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, páginas 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Kraus, 1967] Kraus, J. F. (1967). A study of racial variation in loblolly pine in georgia – tenth-year results. *Ninth South. Conf. For. Tree Improv. Proc.*, páginas 78–85.

[La Farge, 1974] La Farge, T. (1974). Genetic variation among and within three loblolly pine stands in Georgia. *Forest Science*, 20(3):272–275.

[Lamprecht, 1990] Lamprecht, H. (1990). *Silvicultura en los trópicos: los ecosistemas forestales en los bosques tropicales y sus especies arbóreas ; posibilidades y métodos para un aprovechamiento sostenido*. TZ-Verlag-Ges.

[Leduc e Station, 2001] Leduc, D. J. e Station, U. S. F. S. S. R. (2001). *Predicting diameter distributions of longleaf pine plantations a comparison between artificial neural networks and other accepted methodologies*. Asheville, NC : U.S. Dept. of Agriculture, Forest Service, Southern Research Station. Title from title screen (viewed May 24, 2007).

[Lippmann, 1989] Lippmann, R. P. (1989). Pattern classification using neural networks. *Communications Magazine, IEEE*, 27(11):47–50.

[Liu et al., 2003] Liu, C., Zhang, L., Davis, C., Solomon, D., Brann, T. e Caldwell, L. (2003). Comparison of neural networks and statistical methods in classification of ecological habitats using fia data. *Forest Science*, 49(4):619–631.

[Maestri et al., 2005] Maestri, R., Sanquetta, C. R., Machado, S., Scolforo, J. R. e Côrte, A. P. D. (2005). Viabilidade de um projeto florestal de *Eucalyptus grandis* considerando o sequestro de carbono. *Floresta*, 34(3).

[Marchiori, 1996] Marchiori, J. (1996). *Dendrologia das gimnospermas*. UFSM.

[Martinelli et al., 1994] Martinelli, L. A., Moreira, M. Z., Brown, I. F. e Victoria, R. L. (1994). Incertezas associadas às estimativas de biomassa em florestas tropicais: o exemplo de uma floresta 55 situada no estado de rondônia. Em *Seminário Emissão x Seqüestro de CO_2*, página 221. Companhia Vale do Rio Doce.

[May et al., 2003] May, P., Lustosa, M. e Da Vinha, V. (2003). *Economia do meio ambiente: teoria e prática*. CAMPUS - RJ.

[McCulloch e Pitts, 1943] McCulloch, W. S. e Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

[McQuarrie e Tsai, 1998] McQuarrie, A. e Tsai, C. (1998). *Regression and Time Series Model Selection*. World Scientific.

[Mercer, 1909] Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446.

[Minsky e Papert, 1969] Minsky, M. e Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Boston, USA.

[Morgan e Sonquist, 1963] Morgan, J. N. e Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58:415–435.

[Nemenyi, 1963] Nemenyi, P. (1963). *Distribution-free Multiple Comparisons*. Princeton University.

[Netto e Brena, 1997] Netto, S. e Brena, D. (1997). *Inventário florestal*. UFPR.

[Oliveira, 1968] Oliveira, H. A. d. (1968). *Acácia-negra e Tanino no Rio Grande do Sul*. La Salle, Canoas.

[Patterson, 1996] Patterson, D. W. (1996). *Artificial Neural Networks: Theory and Applications*. Prentice-Hall Series in Advanced Communications. Prentice Hall.

[Platt, 1999] Platt, J. C. (1999). Advances in kernel methods. capítulo: Fast Training of Support Vector Machines Using Sequential Minimal Optimization, páginas 185–208. MIT Press, Cambridge, MA, USA.

[Prodan, 1997] Prodan, M. (1997). *Mensura forestal*. Número 1. Agroamerica.

[Queiroz, 1984] Queiroz, W. (1984). *Análise de fatores pelo método da máxima verossimilhança: aplicação ao estudo da estrutura de florestas tropicais*. Tese de doutorado, Universidade de São Paulo, Escola Superior de Agricultura Luíz de Queiroz.

[Quinlan, 1979] Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. Em Michie, D., editor, *Expert Systems in the Micro-Electronic Age*, páginas 168–201. Edinburgh University Press, Edinburgh.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[Quinlan, 1992] Quinlan, R. J. (1992). Learning with continuous classes. Em *5th Australian Joint Conference on Artificial Intelligence*, páginas 343–348, Singapore. World Scientific.

[Rezende, 2003] Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Manole, Barueri, Brasil.

[Rivest, 1987] Rivest, R. L. (1987). Learning decision lists. *Mach. Learn.*, 2(3):229–246.

[Rondeux, 1999] Rondeux, J. (1999). Forest inventories and biodiversity. *Unasylva*, (196):35–41.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E. e Williams, R. J. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. capítulo: Learning Internal Representations by Error Propagation, páginas 318–362. MIT Press, Cambridge, MA, USA.

[Russell e Norvig, 2010] Russell, S. e Norvig, P. (2010). *The Artificial Intelligence*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition.

[Sanquetta et al., 2004] Sanquetta, C. R., Balbinot, R. e Zilliotto, M. A. (2004). Metodologias para determinação de biomassa florestal. *Simpósio Latino Americano sobre Fixação de Carbono*, (5):77–93.

[Sanquetta et al., 2009] Sanquetta, C. R., Watzlawick, L. F., Côrte, A. P. D., Fernandes, L. A. V. e Siqueira, J. D. P. (2009). *Inventários Florestais: Planejamento e Execução*. Multi-graphic.

[Sanquetta et al., 2013] Sanquetta, C. R., Wojciechowski, J., Corte, A. P. D., Rodrigues, A. L. e Maas, G. C. B. (2013). On the use of data mining for estimating carbon storage in the trees. *Carbon Balance Manag*, 8:6–6. 1750-0680-8-6[PII].

[Sanquetta et al., 2015] Sanquetta, C. R., Wojciechowski, J., Dalla Corte, A. P., Behling, A., Péllico Netto, S., Rodrigues, A. L. e Sanquetta, M. N. I. (2015). Comparison of data mining and allometric model in estimation of tree biomass. *BMC Bioinformatics*, 16(1):1–9.

[Schikowski, 2016] Schikowski, A. B. (2016). Estimativa do volume e da forma do fuste utilizando técnicas de aprendizado de máquina. Dissertação de Mestrado, Universidade Federal do Paraná.

[Schmoldt et al., 1997] Schmoldt, D. L., Li, P. e Abbott, A. (1997). Machine vision using artificial neural networks with local 3d neighborhoods. *Computers and Electronics in Agriculture*, 16(3):255 – 271.

[Schneider e Tonini, 2003] Schneider, P. e Tonini, H. (2003). Utilização de variáveis dummy em equações de volume para *Acacia mearnsii* De Wild. *Ciência Florestal*, páginas 121–129.

[Schoeninger, 2006] Schoeninger, E. R. (2006). *Uso de redes neurais artificiais para mapeamento de biomassa e carbono orgânico no componente arbóreo de uma floresta ombrófila densa*. Tese de doutorado, Federal University of Paraná, Curitiba, Paraná, Brazil.

[Schoeninger et al., 2008a] Schoeninger, E. R., Koehler, H. S., Watzlawick, L. F. e de Oliveira Filho, P. C. (2008a). Uso de redes neurais artificiais como uma alternativa para mapeamento de biomassa e carbono orgânico no componente arbóreo de florestas naturais. *Ambiência*, 4(3):529–549.

[Schoeninger et al., 2008b] Schoeninger, E. R., Koehler, H. S., Watzlawick, L. F. e de Oliveira Filho, P. C. (2008b). Uso de redes neurais artificiais para mapeamento de biomassa e carbono orgânico no componente arbóreo de uma Floresta Ombrófila Densa. *Ambiência*, 4(2):179–195.

[Schumacher e Hall, 1934] Schumacher, F. e Hall, D. (1934). *Logarithmic Expression of Timber-tree Volume*. U.S. Government Printing Office.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

[Scolforo, 1998] Scolforo, J. (1998). *Biometria Florestal: modelagem do crescimento e da produção de florestas plantadas e nativas*. UFLA/FAEPE.

[Scolforo, 2011] Scolforo, J. (2011). *Manejo Florestal*. UFLA Textos Acadêmicos.

[Scolforo et al., 1994] Scolforo, J., Mello, J. e Lima, C. (1994). Obtenção de relações quantitativas para estimativa de volume do fuste em Floresta Estacional Semidecídua Montana. *Cerne*, 1(1):123–134.

[Shimizu, 2006] Shimizu, J. Y. (2006). Pinus na silvicultura brasileira. *Revista da Madeira*, 16(99):4–14.

[Shimizu e Medrado, 2005] Shimizu, J. Y. e Medrado, M. J. S. (2005). *Cultivo do Pinus*. Embrapa Florestas.

[Shoemaker e Cropper Jr, 2008] Shoemaker, D. A. e Cropper Jr, W. P. (2008). Prediction of leaf area index for southern pine plantations from satellite imagery using regression and artificial neural networks. *Proceedings of the 6th Southern Forestry and Natural Resources GIS Conference (2008)*, páginas 139–160.

[Silva e Carvalho, 1984] Silva, J. e Carvalho, M. (1984). Equações de volume uma floresta secundária no planalto do Tapajós, Belterra -– Pará.

[Smola e Schölkopf, 2004] Smola, A. J. e Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

[Souza e Jesus, 1991] Souza, A. e Jesus, R. (1991). Equações de volume comercial e fator de forma para espécies da mata atlântica ocorrentes na reserva florestal da companhia vale do Rio Doce, Linhares – ES. *Revista Árvore*, páginas 257–273.

[Theodoridis e Koutroumbas, 2008] Theodoridis, S. e Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. Academic Press, 4th edition.

[UNFCCC – United Nations Framework Convention on Climate Change, 2007] UNFCCC – United Nations Framework Convention on Climate Change (2007). Fourth assessment report: Summary for policymakers. Relatório técnico, Geneva, Switzerland.

[van Laar e Akça, 2007] van Laar, A. e Akça, A. (2007). *Forest Mensuration*.

[Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

[Vapnik e Chervonenkis, 1971] Vapnik, V. N. e Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

[Watzlawick et al., 2009] Watzlawick, L. F., Kirchner, F. F. e Sanquetta, C. R. (2009). Estimativa de biomassa e carbono em floresta com araucária utilizando imagens do satélite Ikonos II. *Ciência Florestal*, 19:169–181.

[Weibull, 1951] Weibull, W. (1951). A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics*, 18:293–305.

[Wells e Wakeley, 1996] Wells, O. O. e Wakeley, P. C. (1996). Geographic variation in survival, growth, and fusiform-rust infection of planted loblolly pine. *Forest Science*, 12(3):a0001–z0001.

[Witten e Frank, 2005] Witten, I. H. e Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Elsevier, San Francisco, California, USA, 2 edition.

[Wojciechowski, 2015] Wojciechowski, J. (2015). *JCarbon – Software na Web com Data Mining para Estimativas de Volume, Biomassa e Carbono em Florestas*. Tese de doutorado, UFPR–Federal University of Paraná, Curitiba, Paraná.

[Wolpert, 1992] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

[Zhang et al., 2000] Zhang, Hebda, J., Zhang e Alfaro, I. (2000). Modeling tree-ring growth responses to climatic variables using artificial neural networks. *Forest Science*, 46(2):229–239.

[Zhang, 1992] Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, 87(419):732–737.

Apêndice A

Arquivos ARFF de dados

Os dados apresentados nos apêndices A.1, A.1.1 e A.1.2 apresentam, de forma parcial, os dados utilizados nos experimentos deste trabalho.

O formato dos arquivos é o padrão do WEKA, chamado ARFF, contendo um cabeçalho que descreve todos os atributos (@ATTRIBUTE) que podem ser numéricos (NUMERIC) ou categóricos (como Local, Fazenda, entre outros) e abaixo os dados (@DATA). Cada informação de um indivíduo é uma linha de dado.

A.1 Dados de *Pinus taeda*

```
@RELATION VolumesPinus

@ATTRIBUTE Idade      NUMERIC
@ATTRIBUTE Dap        NUMERIC
@ATTRIBUTE Altura     NUMERIC
@ATTRIBUTE Volume     NUMERIC

@DATA
10.01369863,8.6,11.05,0.0325727
17.35342466,43,28.6,1.823623045
11.04109589,18.4,15.34,0.185971559
11.04657534,33.4,17.97,0.697658061
11.02191781,32.7,19.74,0.709558326
```

A.1.1 Dados de Acácia-negra

```
@relation BiomassaAcaciaNegra

@attribute Local {Cristal,Encruzilhada,Piratini}
@attribute Fazenda {Camboata,'Cerro Partido',Coronilha,'Da Armada',
                    'Deny de Oliveria','Duas Figueiras','Nova Era',
                    'Ouro Verde',Timbaúva}
@attribute Idade numeric
@attribute DAP numeric
@attribute Altura numeric
@attribute Copa numeric
```

```
@attribute Biomassa numeric

@data
Cristal,'Ouro Verde',10,16.074649,19.2,7.1,129.352185
Piratini,'Nova Era',5,12.732395,16.4,6.4,80.045309
Encruzilhada,'Cerro Partido',10,7.321127,11.5,1.4,17.450814
Encruzilhada,Coronilha,5,8.276057,13.2,7.1,26.023056
```

A.1.2 Dados de Florestas Tropicais

```
@relation BiomassaTropicais

@attribute Local {Australia,BraMan2,BraPara1,BraPara3,BraRond,
                  Cambodia,Cameroon,Cameroon3,CentralAfric,
                  ColombiaC1,ColombiaG1,ColombiaG2,ColombiaM1,
                  ColombiaM2,CostaRic,FrenchGu,Gabon,Ghana,
                  IndiaCha,Jalisco,Kaliman1,Kaliman2,Kaliman4,
                  Kaliman6,Karnataka,Llanosec,Llanosol,Madagascar1,
                  Madagascar2,Madagascar3,Madagascar4,Madagascar5,
                  Malaysia,Malaysia2,MFrenchG,MGuadel,Moluccas,
                  Mozambique,NewGuinea,Peru,PuertoRi,PuertoRi2,
                  SaoPaulo3,Sarawak,SouthAfrica,SouthBrazil1,
                  SouthBrazil2,SouthBrazil3,Sumatra,Sumatra2,
                  Tanzania1,Tanzania2,Tanzania3,Tanzania4,Venezuela2,
                  WestJava,Yucatan,Zambia}

@attribute DAP numeric
@attribute HT numeric
@attribute ME numeric
@attribute AGB numeric

@data
Madagascar2,18.5,9,0.53,128.39
FrenchGu,11.1,12.9,0.67,28.75
SouthAfrica,22,6.4,0.84,80.57
Madagascar4,28,15,0.63,369.44
```