UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE ENGENHARIA FLORESTAL

MARCO ANTÔNIO DIAS MACHADO

DETECÇÃO E INFLUÊNCIA DE OUTLIERS NA QUALIDADE DE MODELOS DE RELAÇÃO HIPSOMÉTRICA SOB O PONTO DE VISTA PREDITIVO

TRABALHO DE CONCLUSÃO DE CURSO II

DOIS VIZINHOS
2016

MARCO ANTÔNIO DIAS MACHADO

DETECÇÃO E INFLUÊNCIA DE OUTLIERS NA QUALIDADE DE MODELOS DE RELAÇÃO HIPSOMÉTRICA SOB O PONTO DE VISTA PREDITIVO

Trabalho de Conclusão de Curso apresentado à disciplina de Trabalho de Conclusão de Curso II, do Curso Superior de Engenharia Florestal da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para a obtenção do título de Bacharel em Engenharia Florestal.

Orientador: Prof. Dr. Edgar de Souza Vismara.

DOIS VIZINHOS

2016

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Dois Vizinhos
Curso de Engenharia Florestal

TERMO DE APROVAÇÃO

DETECÇÃO E INFLUÊNCIA DE OUTLIERS NA QUALIDADE DE MODELOS DE RELAÇÃO HIPSOMÉTRICA SOB O PONTO DE VISTA PREDITIVO

por

Marco Antônio Dias Machado

Este Trabalho de Conclusão de Curso foi apresentado em 05 de Dezembro de 2016 como requisito parcial para a obtenção do título de Bacharel em Engenharia Florestal. O(a) candidato(a) foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Dr. Edgar de Souza Vismara
Orientador(a)

_____
Prof. Dr. Claudio Thomas
Membro titular (UTFPR)

_____
Prof. Dr. Maurício Romero Gorenstein
Membro titular (UTFPR)

_____
Prof. Me. Lilian de Souza Vismara
Membro titular (UTFPR)

- O Termo de Aprovação assinado encontra-se na Coordenação do Curso -

M149d Machado, Marco Antônio Dias.
Detecção e influência de outliers na qualidade de modelos de relação hipsométrica sob o ponto de vista preditivo – Dois Vizinhos: [s.n], 2016.
53f.:il.

Orientador: Edgar de Souza Vismara
Trabalho de Conclusão de Curso (graduação) - Universidade Tecnológica Federal do Paraná. Curso de Engenharia Florestal, Dois Vizinhos, 2016.
Bibliografia p.42-45

1. Florestas - Medição 2. Levantamentos florestais 3. Modelos matemáticos I. Vismara, Edgar de Souza, orient. II. Universidade Tecnológica Federal do Paraná – Dois Vizinhos. III. Título
CDD: 634.9

Ficha catalográfica elaborada por Rosana Oliveira da Silva CRB: 9/1745
Biblioteca da UTFPR-Dois Vizinhos

Entia non sunt multiplicanda praeter necessitatem
(CORK, 1639)

AGRADECIMENTOS

Agradeço a Deus por ter me dado saúde e força.

Agradeço a minha família, principalmente a mãe, que mesmo longe sempre me apoiou de todas as formas possíveis.

Aos meus colegas, amigos e demais pessoas que fizeram parte de toda a minha jornada universitária.

Agradeço a todos os professores que me proporcionaram um caminho mais prazeroso ao conhecimento, aos meus orientadores de monitorias e grupos de pesquisas, em especial ao Prof. Dr. Edgar Vismara, que desde o quarto semestre me orientou em diversos projetos, sendo um dos principais responsáveis pelas oportunidades que eu tive durante a graduação e poderei ter como engenheiro florestal.

RESUMO

Modelos que expressam a relação da altura da árvore com o diâmetro da mesma são denominados modelos de relação hipsométrica, que tem o objetivo de estimar a altura das árvores não mensuradas, reduzindo gastos em inventários florestais. Em meio à amostra utilizada para ajustar o modelo de relação hipsométrica é comum observar a presença de *outliers*, que podem proporcionar estimativas instáveis nos parâmetros do modelo ajustado, bem como proporcionar um pior desempenho na predição das alturas. O objetivo deste trabalho foi avaliar a influência de *outliers* em modelos de relação hipsométrica e comparar diferentes abordagens como alternativas na melhoria destes modelos frente à presença de valores atípicos. Os dados de *Tectona grandis* são provenientes de plantios comerciais da empresa Floresteca, que foram ajustados modelos de relação hipsométrica por meio dos MQA, MQO com a acomodação dos valores atípicos e o reajuste destes modelos com posterior identificação e tratamento dos *outliers*. O modelo selecionado pelo AIC, RMSE relativo e distribuição residual foi o modelo Embrapa, no qual este foi utilizado para avaliar o desempenho das abordagens propostas neste trabalho, cujo desempenho entre estas teve uma diferença negligenciável em relação ao ajuste do modelo por meio dos MQO com acomodação dos valores atípicos.

Palavras-chave: *Outliers*, Relação Hipsométrica, Ajuste de Modelos.

ABSTRACT

Models that express the relation between tree height and tree diameter are called hypsometric relationship models, where the objective is to estimate the height of unmeasured trees, reducing expenses in forest inventories. Among the sample used to adjust the model of hypsometric relation is common to observe the presence of outliers, where these can provide unstable estimates in the parameters of the adjusted model, as well as provide a worse performance in predicting heights. The objective of this study was to evaluate the influence of outliers on hypsometric models and to compare different approaches as alternatives in the improvement of these models against the presence of atypical values. *Tectona grandis* data were obtained from commercial plantations of Floresteca, where from these, hypsometric relation models were adjusted through the MQA, MQO with the accommodation of the atypical values and the readjustment of these models with subsequent identification and treatment of the outliers. The model selected by the AIC, relative RMSE and residual distribution was the Embrapa model, in which it was used to evaluate the performance of the approaches proposed in this work, where the performance between them had a negligible difference in relation to the adjustment of the model through the OLS With accommodation of the atypical values.

Keywords: Outliers, Hypsometric relationship, Model’s fit.

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

AIC	Critério de Informação de Akaike
BIC	Critério de Informação Bayesiano
DAP	Diâmetro à Altura do Peito
IFC	Inventário Florestal Contínuo
IMA	Incremento Médio Anual
MQA	Mínimos Quadrados Aparados
MQO	Mínimos Quadrados Ordinários
EPE	Erro Padrão de Estimativa
TRMV	Teste da Razão Máxima Verossimilhança

SUMÁRIO

1 INTRODUÇÃO

Inventários florestais, independentemente de sua escala, são imprescindíveis para se obter estimativas referentes à floresta em questão, permitindo alimentar um banco de dados que será de suma importância para tomadas de decisões referentes ao povoamento florestal.

As principais variáveis quantitativas mensuradas durante a realização de um inventário florestal são o DAP, altura total e volume de determinadas árvores do povoamento, geralmente selecionadas a partir de um sistema de amostragem que se adeque a floresta em questão, possibilitando a estimativa das variáveis de interesse para todo o povoamento.

Para redução de custos em inventários florestais, emprega-se a relação hipsométrica que consiste na construção de um modelo preditivo construído a partir da medição do diâmetro de todas árvores da parcela e da altura de um subconjunto dessas árvores. A partir destes pares de dados de diâmetro com altura, mais a adição de parâmetros do povoamento, pode-se ajustar modelos de relação hipsométrica que expressem com certa precisão a curva altura-diâmetro com o objetivo de predizer a altura individual das árvores, por meio dos diâmetros mensurados.

Em meio ao inventário florestal, é comum haver dois tipos de erros, o erro amostral que é decorrente ao inventário florestal a ser realizado por meio de amostragem, isto é, o erro atribuído pela área do povoamento que deixou de ser amostrada. O segundo tipo de erro é o erro não amostral, este podendo ser oriundo de inúmeras causas como erros medição, falta de precisão dos aparelhos utilizados durante as mensurações, entre outras causas.

Erros significativos na mensuração de árvores, erros de digitação no transporte de dados ou outra causa que resulte em uma diferença do valor real do objeto mensurado e o valor obtido pode atribuir uma significativa perda na precisão de inventários e ser uma das causas para o aparecimento de observações discrepantes no conjunto de dados.

Os *outliers* não possuem definição rigorosa, sendo geralmente denominados como valores que se distanciam muito da maioria das observações, sendo um problema na análise de dados por mascarar a amostra. Estes valores podem estar presentes em amostras de grande ou pequena dimensão, interferindo no contexto

da estimação de parâmetros de regressão de modelos de hipsometria, pois os estimadores dos mínimos quadrados ordinários, comumente utilizados para a estimação dos parâmetros são muito sensíveis a presença de valores extremos, podendo gerar estimativas enviesadas.

Os procedimentos para lidar com os *outliers*, geralmente compreendem a detecção dessas observações atípicas por meio de uma grande variedade de métodos como: algoritmos baseados em distância, análise de resíduos, testes de significância, entre outros métodos. Após a detecção, uma técnica comum é a eliminação destas observações. Por outro lado, a eliminação dos possíveis *outliers* pode culminar na perda de informações importantes do conjunto de dados e segundo Barbieri (2012, p. 12), poderá limitar a capacidade de generalização do modelo e diminuição do poder estatístico ou estabilidade das estimativas.

Surgindo como alternativa a metodologia simplista de detecção e eliminação de *outliers*, a estimação dos parâmetros por meio de modelos robustos como o estimador dos mínimos quadrados aparados, que não são tão sensíveis à presença de observações extremas no conjunto de dados parece ser uma alternativa interessante para produzir estimativas consistentes.

Acredita-se que a utilização destes métodos, seja a partir das técnicas de detecção e tratamento dos *outliers* para posterior ajuste dos parâmetros de relação hipsométrica ou ajuste inicial via de regressão robusta podem culminar em predições mais acuradas para o povoamento em questão.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Avaliar a capacidade de diferentes técnicas na detecção de *outliers*, sua influência na qualidade das predições de modelos de relação hipsométrica e o ajuste destes modelos por meio de regressão robusta.

2.2 OBJETIVOS ESPECÍFICOS

a) Avaliar a capacidade da distância de *Cook* e da *DFBETA* na detecção de ″outliers″ nos modelos ajustados pelo método dos mínimos quadrados;

b) Ajustar modelos hipsométricos pelo método dos mínimos quadrados ordinários e mínimos quadrados aparados;

c) Comparar o ajuste dos modelos e a capacidade preditiva dos mesmos;

d) Avaliar a influência dos *outliers* sob os modelos hipsométricos ajustados sob o ponto de vista preditivo.

3 JUSTIFICATIVA

A consistência dos dados é indispensável em qualquer tipo de banco de dados, sendo de suma importância em inventários florestais antes da realização de qualquer procedimento com a informação obtida.

Uma das etapas na análise de dados é a detecção de observações que provavelmente não pertençam ao conjunto de dados, tais observações podem ser denominadas de *outliers*.

As técnicas para detecção destas observações vêm sendo discutidas amplamente durante as últimas décadas, sendo desenvolvidas diversas técnicas para detecção destas observações, por meio de algoritmos, testes de significância, análise residual, entre outros métodos, sendo necessário algum tratamento após a identificação destas observações. Um dos procedimentos simplistas utilizados é a eliminação destes valores, culminando na perda de informações que podem ser importantes para explicar determinado fenômeno. A identificação e posterior tratamento destas observações se torna um processo imprescindível na análise de dados, pois os *outliers* podem mascarar medidas de posição e dispersão da amostra, podendo interferir em testes de comparação de médias ou de análise de variância, bem como proporcionar viés nas estimativas dos parâmetros quando se utiliza o método dos mínimos quadrados ordinários, decorrente da sensibilidade deste estimador à presença de valores extremos.

Como alternativa ao procedimento de detecção e remoção dos *outliers* para o ajuste de modelos de regressão via mínimos quadrados ordinários, surge como opção para o ajuste de relação hipsométrica a estimação dos parâmetros por meio dos mínimos quadrados aparados que tem como característica a robustez, sendo resistente a presença de observações extremas, gerando estimativas consistentes.

4 REFERÊNCIAL TEÓRICO

4.1 CARACTERIZAÇÃO DA ESPÉCIE

Pertencente à família Lamiaceae, gênero *Tectona* e espécie *Tectona grandis* L. F., a teca é nativa das florestas tropicais situadas entre 10° e 25°N no subcontinente índico e no sudeste asiático, principalmente na Índia, Burma, Tailândia, Laos, Camboja, Vietnã e Java (ANGELI, 2003, p. 1).

A Teca, assim como o Cedro do Líbano, é uma das mais antigas madeiras comercializadas no mundo. Especula-se que, desde 4.000 a.C., essas espécies já eram utilizadas no comércio mundial (DRESCHER, 2004 , p. 24).

No Brasil, onde as primeiras experiências com o plantio iniciaram na década de 1960 no Estado do Mato Grosso (ANGELI, 2003, p. 1), tendo ultimamente conquistado espaço entre as principais culturas florestais do país (KUBOYMA, 2012, p. 14).

Além de Mato Grosso, hoje, há plantios de teca no Paraná, São Paulo, Tocantins, Pará, Acre, Rondônia e Amapá (DRESCHER, 2004, p. 27) (Figura 1).

Figura 1 – Distribuição da Teca no Brasil
Fonte: Drescher (2004).

A *Tectona grandis* L. F. , é uma árvore de grande porte, caducifólia, copa arredondada, com fuste cilíndrico revestido de uma casca grossa, apresenta alargamentos na base da árvore, produzidas por inchaço exagerado das raízes e folhas com 30 a 60 cm de comprimento (FLÓREZ, 2012, p. 18).

A madeira é valorizada no mercado internacional, apresentando preços mais elevados do que a madeira de mogno (*Swietenia macrophylla* King) (ROCHA, 2012, p. 1). É de boa trabalhabilidade com ferramentas manuais e elétricas, mas contém sílica que tende a diminuir a afiação dos instrumentos. A madeira é fácil de colar, de fácil acabamento e utilização de pregos e parafusos (FONSECA, 2004, p. 13).

É largamente utilizada na construção naval, pois suporta o contato permanente com a água do mar durante dezenas de anos, sem sofrer deteriorações por brocas marinhas. Apresenta também como característica uma combinação de estabilidade, durabilidade, resistência, beleza e facilidade de ser trabalhada. É usada na fabricação de móveis para ambientes externos, como para jardins, onde são mantidos sem aplicação de tintas ou vernizes. Em ambientes internos sua madeira é utilizada para fabricação de pisos, portas, batentes, janelas e móveis em geral. (RONDON NETO et al., 1998, p. 9-10).

Em relação à produtividade, a espécie apresenta um IMA de 10 a 15 m^3/ha/ano (LADRACH, 2009, p. 6). Segundo o mesmo autor, a partir de estudos desenvolvidos em Trinidad com teca, foi possível determinar que o padrão de incremento assemelha-se a todas as espécies de rápido crescimento, onde o valor máximo de IMA ocorre a uma idade entre os 7 e 12 anos, portanto é possível maneja-la em ciclos curtos de 15 a 25 anos.

As condições climáticas adequadas para o pleno desenvolvimento da teca no Brasil proporcionam taxas de crescimento superiores às dos plantios da maioria dos países produtores dessa madeira, obtendo madeira de dimensões comerciais em ciclos de 20 a 25 anos, os quais estimulam a implantação de plantios comerciais da espécie no país (FLÓREZ, 2012, p. 21).

4.2 RELAÇÃO HIPSOMÉTRICA

Em inventários florestais, a mensuração da altura é de suma importância na quantificação do volume de madeira do povoamento florestal, bem como na classificação produtiva do mesmo. A altura pode ser definida, segundo Finger (2006, p. 27), como a distância linear entre o nível do solo e o ápice da árvore (altura total), sendo uma variável de difícil mensuração quando comparada ao DAP, pois a obtenção da altura é realizada geralmente a partir de métodos indiretos que

demandam tempo e aparelhos, resultando em um procedimento oneroso e sujeito a erros de medição.

Uma alternativa para a redução de gastos em inventários florestais é o emprego da relação hipsométrica, que são modelos que descrevem a relação da altura de uma árvore com o diâmetro da mesma por meio de um modelo matemático, possibilitando uma redução de gastos em inventários florestais.

É importante frisar que a relação hipsométrica possui comportamento distinto para diferentes espécies, sítios e idades. Segundo Loetsch (1973, p. 469), as diferenças de níveis entre as curvas das relações hipsométricas diminuem com o aumento da idade, refletindo na redução do crescimento em altura.

Sendo assim, para uma maior precisão dos modelos de hipsometria, estes devem ser utilizados para locais com características muito próximas da amostra utilizada para o ajuste da relação hipsométrica.

A construção do modelo de relação hipsométrica começa na seleção das árvores que serão coletadas o DAP e a altura, que segundo Schneider (1986), apud Finger (2006, p. 27), 30 a 40 pares de dados por hectare são suficientes se distribuídas em todas as classes diamétricas do povoamento. A importância de que ocorra representatividade em todas as classes diamétricas do povoamento se deve principalmente para evitar a extrapolação das alturas na realização da predição.

Após a amostragem das árvores, parte-se para o ajuste dos modelos hipsométricos. Na literatura é possível encontrar diversos modelos que foram ajustados nas últimas décadas, alguns deles foram ajustados e avaliados quanto a sua estabilidade por Batista (2001, p. 153) e são apresentados na Quadro 1.

Quadro 1 – Forma de ajuste e forma funcional de diversos modelos de relação hipsométrica.

Modelo		Forma de Ajuste	Forma Funcional
Modelos Lineares			
Polinômios	1	$h = \beta_0 + \beta_1 d + \varepsilon$	-
	2	$h = \beta_0 + \beta_1 d + \beta_2 d^2 + \varepsilon$	-
	3	$h = \beta_0 + \beta_1 d + \beta_2 d^2 + \beta_3 d^3 + \varepsilon$	-
Hiperbólicos	4	$h = \beta_0 + \beta_1 (1/d^2) + \varepsilon$	$h = [(\beta_0 d^2 + \beta_1)/(d^2)]$
	5	$1/\sqrt{h} = \beta_0 + \beta_1 (1/d^2) + \varepsilon$	$h = [[(d^2)/(\beta_0 d^2 + \beta_1)]]^2$
	6	$1/h = \beta_0 + \beta_1 (1/d) + \beta_2 (1/d^2) + \varepsilon$	$h = [(d^2)/(\beta_0 + \beta_1 d + \beta_2 d^2)]$
	7	$d^2/h = \beta_0 + \beta_1 d + \beta_2 d^2 + \varepsilon$	$h = [(d^2)/(\beta_0 + \beta_1 d + \beta_2 d^2)]$
	8	$d/\sqrt{h} = \beta_0 + \beta_1 d + \beta_2 d^2 + \varepsilon$	$h = [(d^2)/((\beta_0 + \beta_1 d + \beta_2 d^2)^2)]$
Potência	9	$ln(h) = \beta_0 + \beta_1 ln(d) + \varepsilon$	$h = \beta_0' d^{\beta_1}$
	10	$ln(1/h) = \beta_0 + \beta_1 ln(d) + \beta_2 ln^2(d) + \varepsilon$	$h = [1/(\beta_0')]\, d^{[\beta_1 + \beta_2 ln(d)]}$
	11	$ln(h) = \beta_0 + \beta_1 ln(d/(1+d)) + \varepsilon$	$h = \beta_0'([d/(1+d)])^{\beta_1}$
Exponencial	12	$ln(h) = \beta_0 + \beta_1 (1/d) + \varepsilon$	$h = \beta_0' exp(\beta_1 d^{-1})$
Semilogarítmico	13	$h = \beta_0 + \beta_1 ln(d) + \varepsilon$	$h = \beta_0 + \beta_1 ln(d)$ $\Leftrightarrow \quad d = exp([(h - \beta_0)/(\beta_1)])$
Chapman-Richards	14	$h = \beta_0 [1 - exp(-\beta_1 d)]^{\beta_2} + \varepsilon$	-
Weibull	15	$h = \beta_0 [1 - exp(-\beta_1 d^{\beta_2})] + \varepsilon$	-
Monomolecular	16	$h = \beta_0 [1 - \beta_1 exp(-\beta_2 d)] + \varepsilon$	-
Gompertz	17	$h = \beta_0 exp[-\beta_1 exp(-\beta_2 d)] + \varepsilon$	-
Logístico	18	$h = \beta_0 / [1 + \beta_1 exp(-\beta_2 d)] + \varepsilon$	-

h - altura total das árvores individuais (metros);

d - diâmetro à altura do peito (DAP em centímetros);

$\beta_0, \beta_1, \beta_2$ - parâmetros a serem estimados por quadrados mínimos, $\beta'_0 = exp(\beta_0)$;

ε - erro estatístico com distribuição Normal, média zero e variância constante;

ln - logaritmo neperiano

Fonte: Batista (2001).

Além dos modelos listados no Quadro 1, que têm como variável resposta somente a altura, existem diversos outros modelos em que a variável resposta é a razão da altura com a altura dominante do povoamento, ou modelos que incluem parâmetros do povoamento como diâmetro médio quadrático.

Diversas são as estatísticas utilizadas para seleção de modelos, entre elas estão o AIC, BIC e o TRMV (NETO, 2012, p. 16). Entretanto, a maioria dos trabalhos presentes na literatura relacionados à avaliação de modelos de relação hipsométrica

faz uso principalmente da análise gráfica e do coeficiente de determinação (R^2) que segundo Batista (2004, p. 55) é a proporção da variável resposta (Y) que é explicada pelo modelo e pela variável preditora (X).

O R^2 pode variar de 0 a 1. Valores próximos a 0 (zero) indicam baixa ou nenhuma relação das variáveis preditoras com a variável resposta, em contrapartida modelos que apresentam R^2 próximos a 1 (um) indicam uma boa relação da variável resposta com as variáveis explicativas, indicando um bom ajuste do modelo. Vale frisar que segundo Batista (2004, p. 55) os limites para quais se considera ajuste "bom" depende muito do tipo de conjunto de dados e a situação em que se aplica a regressão.

Após o ajuste e a seleção do modelo que melhor represente a relação hipsométrica existente no povoamento, é possível por meio desta realizar predições das alturas individuais mensurando apenas o DAP das árvores da parcela.

4.3 OUTLIERS

Os *outliers* não possuem definição rigorosa, sendo geralmente denominados como valores discrepantes, estranhos, aberrantes, anormais ou de acordo com Barnett e Lewis (1978, p. 4), observações inconsistentes com os dados, sendo uma das possíveis causas do surgimento destes valores a variabilidade natural (BECKMAN e COOK, 1983 apud MACHADO, 1997, p. 8).

Mesmo com todas essas definições, não existe uma definição matemática rígida do que constitui um *outlier* (BARROS, 2013, p. 21).

Wisnowski et al. (2002, p. 1) enfatizam que os *outliers* têm se tornado comuns em experimentos com grande volume de dados.

A contaminação de uma amostra pela presença de *outliers* pode resultar em vários efeitos de acordo com a proporção da amostra contaminada, entre eles alteração das estimativas da amostra como média e superestimação de variância. Podem influenciar, segundo Chicareli et al. (2009, p. 145), nos resultados da estatística F para tratamentos e testes de comparações de médias, ocultar correlações reais ou forjar associações inexistentes (FILHO et al., 2014, p. 66). Segundo Barbieri (2012, p. 12), a presença de valores extremos pode gerar estimativas instáveis de parâmetros e estimativas de erros padrão inflacionadas.

Além dos efeitos citados acima, os *outliers* podem impactar tanto o coeficiente

de regressão quanto o intercepto, pois na regressão clássica, o desvio da linha de melhor ajuste é calculado pela soma do quadrado dos resíduos, que na presença de valores extremos apresentará maior magnitude (CHERNICK; FRIIS, 2003, p. 265).

Na tentativa de tentar solucionar os problemas resultantes da contaminação da amostra por *outliers*, existem segundo (MACHADO, 1997, p. 8) dois métodos para se lidar com observações discrepantes: o da Acomodação e o da Identificação. O primeiro consiste em acomodar os valores atípicos mediante a alteração do método de trabalho de modo que não seja necessário o tratamento dos *outliers*, utilizando por exemplo métodos robustos. O segundo método para lidar com os conjuntos de dados contaminados consiste na identificação e posterior tratamento dos valores atípicos, uma das metodologias simplistas utilizadas nesta segunda abordagem é a remoção dos valores atípicos identificados para trabalhar com o restante dos dados remanescentes.

Para a identificação dos *outliers* foram desenvolvidos diversos métodos, em amostras univariadas, podemos citar o teste de *Grubbs* que consiste em um teste de significância onde se o valor calculado for maior que o valor tabelado (GRUBBS, 1950, p. 28) a um nível de significância definido, conclui-se que o valor é um *outlier*.

A análise de resíduos a partir de um modelo ajustado também pode ser útil na detecção de informações atípicas. Entre as opções disponíveis estão: a distância de *Cook* que mede a influência de determinada observação sobre todos os valores ajustados (PORTAL ACTION, 2015, p. 1). Segundo Barbieri (2012, p. 13), é comum analisar as alterações nos coeficientes individuais de regressão devido a casos específicos identificados como influentes, quando esta medida de influência é maior que 0,5.

Outra medida de influência é o *DFBETA* que segundo (PORTAL ACTION, 2015, p. 1), representa a mudança padronizada no parâmetro quando a i-ésima observação é excluída da amostra de ajuste, quantificando a influência de uma dada observação na estimação de determinado parâmetro.

4.4 REGRESSÃO ROBUSTA

É senso comum que a análise de regressão é a ferramenta estatística mais utilizada em aplicações para estimar os parâmetros de um modelo linear de regressão, onde comumente se aplica o método dos MQO (2008, BULHÕES & LIMA, p. 1).

Segundo Doornik (2001, p. 1) as estimativas geradas pelos MQO são muito sensíveis até mesmo uma pequena quantidade de contaminação dos dados, pois segundo Rousseew e Leroy (1987, p. 10) o estimador MQO possui um ponto de ruptura igual o inverso do tamanho da amostra, o que significa que a existência de um único ponto amostral contaminando a amostra pode comprometer a qualidade das estimativas, sendo assim o ponto de ruptura do estimador MQO é de 0 % (CHAGAS ,2001,.p 33).

Gnanadesikan (1997, p. 145) definiu ponto de ruptura como a maior fração das observações em uma amostra, que podem ser valores extremos sem distorcer o valor do estimador.

Uma das técnicas a serem utilizadas na presença de *outliers* é a detecção e remoção dos mesmos, porém segundo Farcomeni & Ventura (2010, p.2) o rastreio dos *outliers* para posterior exclusão com intuito de trabalhar com os dados remanescentes pode ser falho pela dificuldade na detecção destes valores.

Uma alternativa a remoção dos possíveis *outliers* da amostra, é a utilização de regressão robusta, que consiste em um método onde seu ponto de ruptura é superior a 0 %. De acordo com Barbieri (2012, p. 15), estes métodos começaram a surgir na década de 60, com o objetivo de minimizar o impacto de valores extremos nas estimativas dos parâmetros, com o objetivo de produzir os mesmos resultados na presença ou ausência de *outliers*, devido ao ajuste do modelo a maioria dos dados.

Tais métodos além de robustos, também são chamados de resistentes e têm como objetivo uma estimativa confiável do vetor de coeficientes apesar da presença de observações atípicas no conjunto de dados (MACHADO, 1997, p. 64).

Um dos estimadores robustos presentes na literatura é método dos mínimos quadrados aparados (ROUSSEEUW, 1987, p. 132), que segundo Machado (1997, p. 72) consiste em achar o subconjunto de observações cuja retirada do conjunto de dados resulta na regressão com a menor soma de quadrados dos resíduos.

Segundo o mesmo autor, o método dos mínimos quadrados aparados tem alto ponto de ruptura, tendendo a 50 % quando a amostra tende ao infinito e não sofre o efeito de mascaramento e/ou "swamping" como o método dos mínimos quadrados. (MACHADO, 1997, p. 75).

Na literatura florestal, o método dos mínimos quadrados aparados já foi utilizado por (CUNHA et al., 2002, p. 401) para determinação de incremento periódico de diâmetro e área basal, cujos estimadores foram mais eficientes que pelo o método dos mínimos quadrados ordinários.

5 METODOLOGIA

5.1 CONJUNTO DE DADOS

Dispõe-se de um conjunto de dados composto pelas variáveis DAP, altura total e variáveis a nível de povoamento (altura dominante e diâmetro médio quadrático) de árvores da espécie *Tectona grandis*, coletados a partir IFC de plantios comerciais da empresa Floresteca, nos estados do Mato Grosso e Pará.

A amostra é composta por coletas compreendidas entre o período de 2002-2012, consistindo em mais de 10.000 pares de dados de diversas fazendas.

5.2 MODELOS CONCORRENTES

5.2.1 Ajuste dos modelos

Foi ajustado e avaliado o desempenho de cada modelo de relação hipsométrica da Tabela 1 em nível de projeto, onde para todos os modelos lineares, foi testado o efeito da inclusão da variável h_{dom} (média das alturas dominantes). Esta variável foi incluída no modelo, testando-se os níveis hierárquicos de talhão e parcela. Apenas a variante do modelo com o menor valor de AIC (AKAIKE, 1974, p. 716-718) concorreu com os outros modelos a fim de selecionar o modelo utilizado para alcançar os objetivos propostos para este trabalho.

O AIC é computado da seguinte forma:

$$AIC = -2\log\left(L\left(\left(\widehat{\sigma^2}\right)\right)\right) + 2p \qquad (1)$$

Onde:

$\widehat{\sigma^2}$ = estimativa da verossimilhança da variância e p = número de parâmetros do modelo.

Para o modelo de Pienaar, as variáveis Dg (diâmetro médio quadrático) e h_{dom}, foram testadas a partir do mesmo procedimento utilizado na inclusão da variável h_{dom} nos modelos lineares, porém o modelo selecionado, será escolhido a partir do RMSE relativo, devido a variável resposta dos dois modelos serem distintas.

Tabela 1: Modelos hipsométricos concorrentes.

Modelo	Equação
Curtis	$\ln(ht) = \beta_0 + \beta_1 DAP^{-1} + \varepsilon$
Parábola	$ht = \beta_0 + \beta_1 DAP + \beta_2 DAP^2 + \varepsilon$
Exponencial	$ht = e^{\beta_0 + \frac{\beta_1}{DAP}} + \varepsilon$
Potência	$\ln(ht) = \beta_0 + \beta_1 \ln(DAP) + \varepsilon$
Embrapa	$ht = \beta_0 + \beta_1 DAP + \beta_2 DAP^2 + \beta_3 (DAP.h_{dom}) + \varepsilon$
Pienaar	$\frac{ht}{hdom} = \beta_1 \left(1 - \beta_2 e^{-\beta_3 \frac{DAP}{Dg}}\right) + \varepsilon$

Fontes: Curtis (1967); Schumacher (1939) e Stoffels e Soest (1953).

Onde: DAP = é o diâmetro à altura do peito (cm), ht = é a altura total da árvore (m), β_i = são os parâmetros da regressão, hdom = média das alturas das árvores dominantes, Dg = diâmetro médio quadrático e ε = é o erro de estimativa.

A inclusão da variável altura dominante nos modelos, de acordo com Cardoso (1989) deve-se ao fato de ser uma informação disponível durante todo o ciclo de vida de um povoamento e por se tratar de uma variável que exprime a qualidade do sítio, sendo pouco afetada por variações de densidade, auxiliando o modelo na explicação da altura total.

5.2.2 Seleção dos modelos

Entre os modelos concorrentes, foi selecionado apenas o que teve o melhor desempenho de acordo com a distribuição residual e menor EPE relativo (na escala original da altura), onde este último, é computado pela seguinte equação:

$$\text{Erro padrão de estimativa relativo } (\%) = \frac{\left[\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{(n-p)}}\right]}{\bar{Y}}.100 \qquad (2)$$

Onde:

$\hat{Y}$ = é o valor estimado da variável resposta (m), Y = é o valor observado da altura (m), n = é o tamanho da amostra utilizada no ajuste, p = é o número de coeficientes do modelo e $\bar{Y}$ = é o valor médio das alturas (m).

O erro padrão de estimativa relativo final, foi computado a partir de uma média ponderada, onde os pesos foram a quantidade de pares de dados presentes para cada projeto.

5.3 DETECÇÃO DOS OUTLIERS

A detecção dos *outliers* ocorreu a partir de quatro metodologias:

A primeira e segunda metodologia de detecção de *outliers* foi realizada a partir do resíduo normalizado, metodologia bastante utilizada devido sua obtenção ser bastante simples, podendo ser calculada da seguinte maneira:

$$r = \frac{\varepsilon}{\sqrt{QM_{erro}}} \qquad (3)$$

Onde:

ε = resíduo e $\sqrt{\text{QMerro}}$ = é o desvio padrão estimado dos resíduos.

Se os erros têm distribuição normal, então 95 % dos resíduos normalizados devem estar no intervalo (-2,2). Desta forma, serão testados para detecção de *outliers* resíduos normalizados com valores acima de $|2|$ e acima de $|3|$, cujo último, é representado por resíduos normalizados fora do intervalo de 99 % dos resíduos.

A terceira metodologia foi a medida de influência Distância de *Cook* (D_i) que pode ser definida da seguinte maneira:

$$Di = \frac{\sum_{j=1}^{n}\left(\hat{y_j} - \hat{y_{j(i)}}\right)^2}{pxQM_{erro}} \qquad (4)$$

Onde:

$\hat{y_j}$ = é a previsão do modelo de regressão para a observação *j*, $\hat{y}_{j(i)}$ = é a previsão da observação *j* de um modelo que foi reajustado com a ausência da observação *i*, *p*= é o número de parâmetros do modelo ajustado e QM_{erro} = é o quadrado médio do modelo de regressão.

Sendo que se $D_i > \frac{4}{tamanho\ da\ amostra}$, o valor será considerado um *outlier*.

A quarta metodologia utilizada na detecção dos *outliers* foi a medida de influência DFBETA que pode ser computada da seguinte forma:

$$DFBETA = \frac{b_{(-i)j} - b_j}{\hat{\sigma}(b_j)} \qquad (5)$$

Onde:

$\hat{\sigma}(b_j)$ = é o erro estimado de b_j e b = é o valor dos parâmetros estimados.

Sendo que $\text{DFBETA} > 2\sqrt{n}$ corresponde aos valores de alta influência no

modelo, sendo considerados como *outliers*.

5.4 AJUSTE POR MEIO DOS MÍNIMOS QUADRADOS APARADOS

Esta abordagem tem como finalidade a acomodação dos valores atípicos, ou seja, será realizado o ajuste dos modelos descritos no capítulo 3.2 por meio do estimador dos mínimos quadrados aparados por meio do programa R, utilizando o pacote *robustbase* sem a necessidade da remoção das observações atípicas.

O modelo que irá representar esta abordagem será o mesmo selecionado para o ajuste por meio dos mínimos quadrados ordinários.

5.5 AVALIAÇÃO DO PERFORMANCE DAS ABORDAGENS

A avaliação de todas as abordagens foram realizadas utilizando a metodologia validação cruzada, que consiste na utilização de amostras diferentes para ajustar e realizar a predição dos dados. Onde o procedimento foi repetido 5 vezes, alternando os dados entre amostra de ajuste e amostra de predição, gerando por fim uma média.

As etapas presentes nesse processo, são as seguintes: o conjunto de dados completo (em nível de projeto) gerou três amostras, onde a amostra “A” foi composta 20 % das árvores do projeto em questão, contemplando o mesmo número de pares de dados para as classes diamétrica do projeto (três classes diamétricas), a fim de evitar extrapolação de valores no momento da predição. A amostra “B” foi dividida em 4 subamostras, todas oriundas da amostra “A”, porém sem a presença dos *outliers* identificados para cada método de identificação descrito no capítulo 3.3. Por fim, a amostra “C” (amostra de predição) foi composta pelas observações remanescentes não inclusas na amostra “A”.

A partir do modelo ajustado por meio dos mínimos quadrados ordinários selecionado com base nos critérios descritos no capítulo 3.2, o mesmo foi ajustado a partir da amostra “A” e “B” com o objetivo de determinar se a remoção dos *outliers* identificados a partir de diferentes métodos de detecção proporcionaria um melhor desempenho na predição das alturas. Este desempenho foi avaliado a partir da predição das alturas da amostra “C”, onde os valores preditos foram comparados com os valores reais, por meio do RMSE relativo.

Foi ajustado também o modelo com estimador robusto para a amostra “A”,

onde a performance do mesmo foi comparada as demais abordagens.

Os procedimentos descritos acima podem ser visualizados recorrendo ao fluxograma disposto na Figura 3.

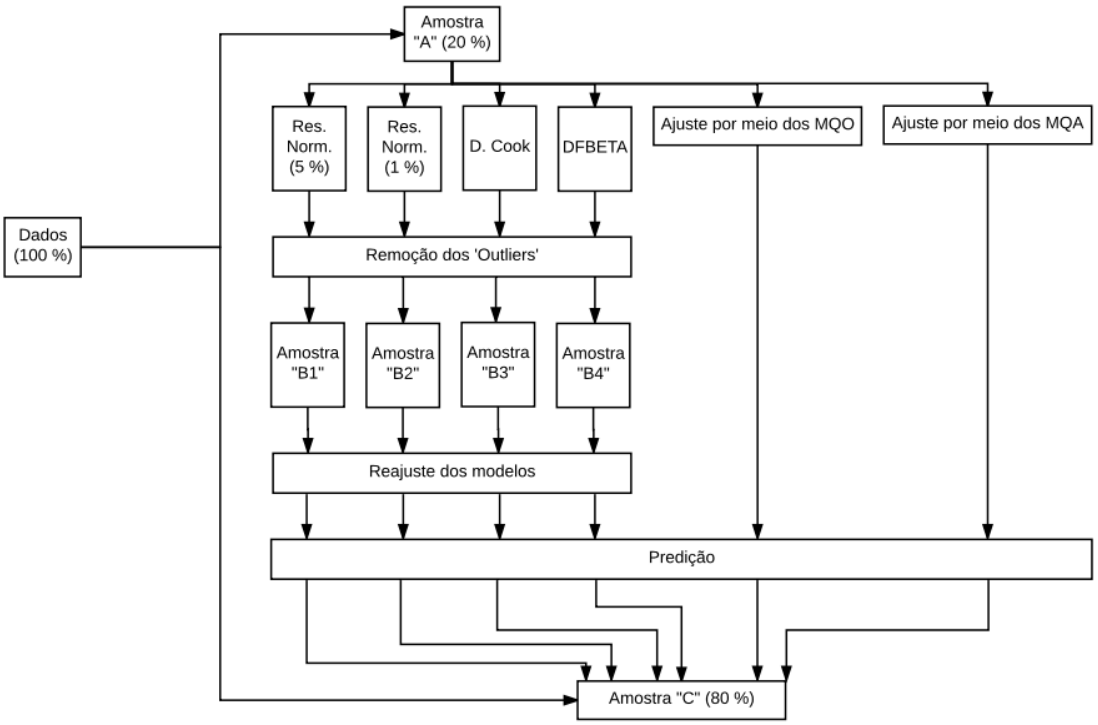

Figura 2 - Fluxograma da avaliação das abordagens.
Fonte: O autor (2016).

Onde:

MQO = Mínimos Quadrados Ordinários e MQA = Mínimos Quadrados Aparados

Para evidenciar o efeito das abordagens no desempenho preditivo dos modelos de acordo com a contaminação relativa da amostra, plotou-se gráficos da redução relativa do RMSE do modelo reajustado pós-tratamento (em comparação com o ajuste sem a remoção dos *outliers*) para cada uma das abordagens em função da contaminação relativa da amostra. A correlação entre as variáveis foi avaliada por meio do teste de correlação de Spearman em um nível de 5 % de probabilidade.

6 RESULTADOS E DISCUSSÃO

6.1 PERFORMACE DOS MODELOS

6.1.1 Modelo de Curtis

Tabela 2: AIC dos modelos de Curtis.

Forma de ajuste	AIC
$\ln(ht) = \beta_0 + \beta_1 DAP^{-1} + \varepsilon$	-88.761,14
$\ln(ht) = \beta_0 + \beta_1 DAP^{-1} + \ln(hdom_{talhão}) + \varepsilon$	-106.083,39
$\ln(ht) = \beta_0 + \beta_1 DAP^{-1} + \ln(hdom_{parcela}) + \varepsilon$	-116.184,31

Fonte: O autor (2016).

Por meio dos valores de AIC dos modelo de Curtis em sua forma tradicional e modificada com inclusão da variável logarítimo neperiano da altura dominante em nível de talhão e parcela, é possível selecionar o modelo modificado com o logaritmo neperiano da altura dominante em nível de parcela.

Desta forma, o gráfico de distribuição residual está plotado (Figura 4) apenas para o modelo selecionado.

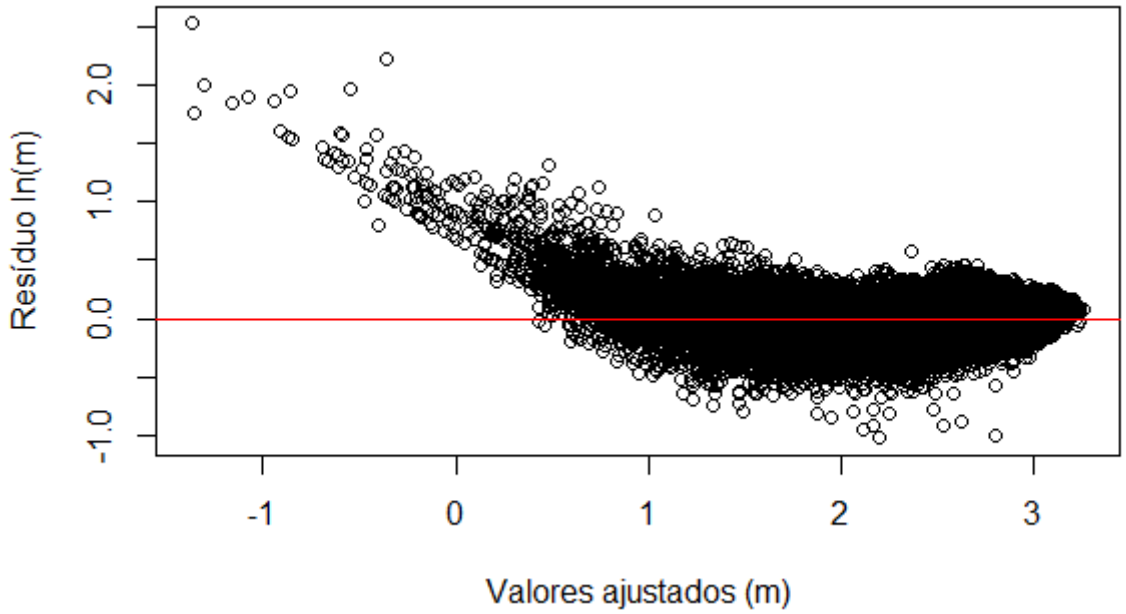

Figura 3 – Distribuição residual do modelo de Curtis modificado.
Fonte: O autor (2016).

É possível observar a partir da distribuição residual do modelo de Curtis modificado uma forte tendência de subestimação para árvores com menores valores de altura, resultando em valores não condizentes com o conjunto de dados, pois o valor mínimo de altura estimado pelo modelo foi de 0,25 metros e não são

mensuradas árvores com altura menor que 1,3 metros. De qualquer forma, o impacto dessas árvores na estimação do volume em nível de povoamento é baixo, por se tratar das árvores das menores classes diamétricas.

6.1.2 Modelo Parábola

Tabela 3: AIC dos modelos Parábola.

Forma de ajuste	AIC
$\ln(ht) = \beta_0 + \beta_1 (DAP) + \beta_2 (DAP^2) + \varepsilon$	342.010,6
$\ln(ht) = \beta_0 + \beta_1 (DAP) + \beta_2 (DAP^2) + \ln(hdom_{talhão}) + \varepsilon$	331.414,8
$\ln(ht) = \beta_0 + \beta_1 (DAP) + \beta_2 (DAP^2) + \ln(hdom_{parcela}) + \varepsilon$	323.715,9

Fonte: O autor (2016).

Por meio dos valores de AIC dos modelos Parábola em sua forma tradicional e modificada com a inclusão da variável logarítimo neperiano da altura dominante em nível de talhão e parcela, é possível selecionar o modelo modificado com o logaritmo neperiano da altura dominante em nível de parcela.

Desta forma, o gráfico de distribuição residual é plotado (Figura 5) apenas para o modelo selecionado.

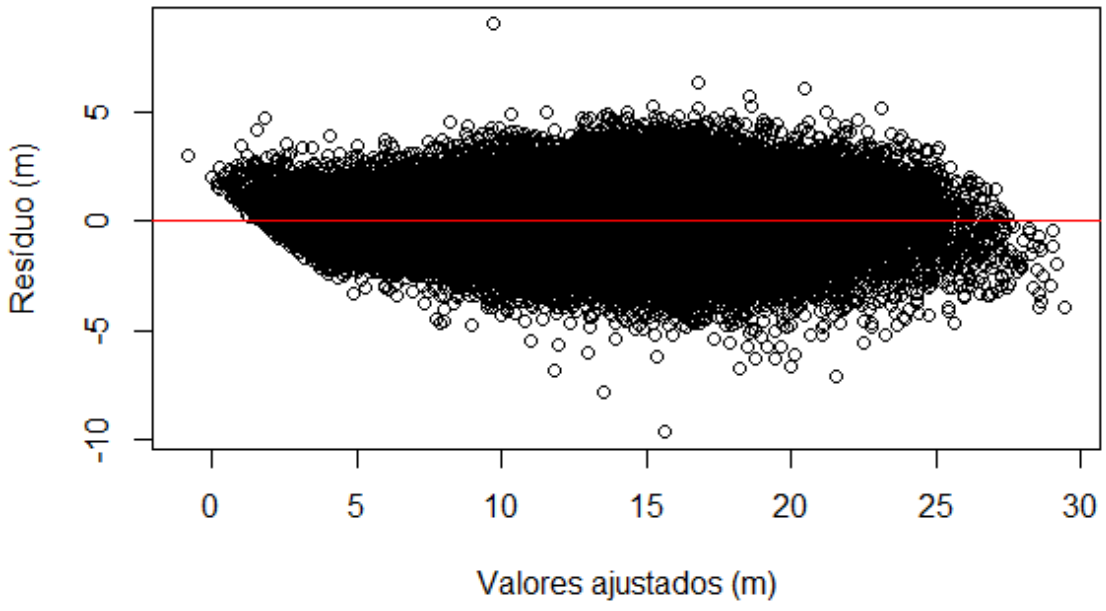

Figura 4 – Distribuição residual do modelo de Parábola modificado.
Fonte: O autor (2016).

É possível observar a partir da distribuição residual do modelo Parábola uma forte tendência de subestimação para árvores com menores valores de altura, resultando em valores não condizentes com o conjunto de dados, pois não são

mensuradas árvores com altura menor que 1,3 metros e o modelo estimou para árvores das menores classes diamétricas, alturas menores que esta. De qualquer forma, o impacto dessas árvores na estimação do volume em nível do povoamento é baixo, por se tratar das árvores das menores classes diamétricas.

6.1.3 Modelo Potência

Tabela 4: AIC dos modelos de Potência.

Forma de ajuste	AIC
$\ln(ht) = \beta_0 + \beta_1 \ln(DAP) + \varepsilon$	-147.861
$\ln(ht) = \beta_0 + \beta_1 \ln(DAP) + \ln(hdom_{talhão}) + \varepsilon$	-159.163
$\ln(ht) = \beta_0 + \beta_1 \ln(DAP) + \ln(hdom_{parcela}) + \varepsilon$	-169.725

Fonte: O autor (2016).

Por meio dos valores de AIC dos modelos de Potência em sua forma tradicional e modificado com a adição da variável logarítimo neperiano da altura dominante em nível de talhão e parcela, é possível selecionar o modelo modificado com o logaritmo neperiano da altura dominante em nível de parcela, por apresentar o menor valor de AIC.

Desta forma o gráfico de distribuição residual é plotado (Figura 6) apenas para o modelo selecionado.

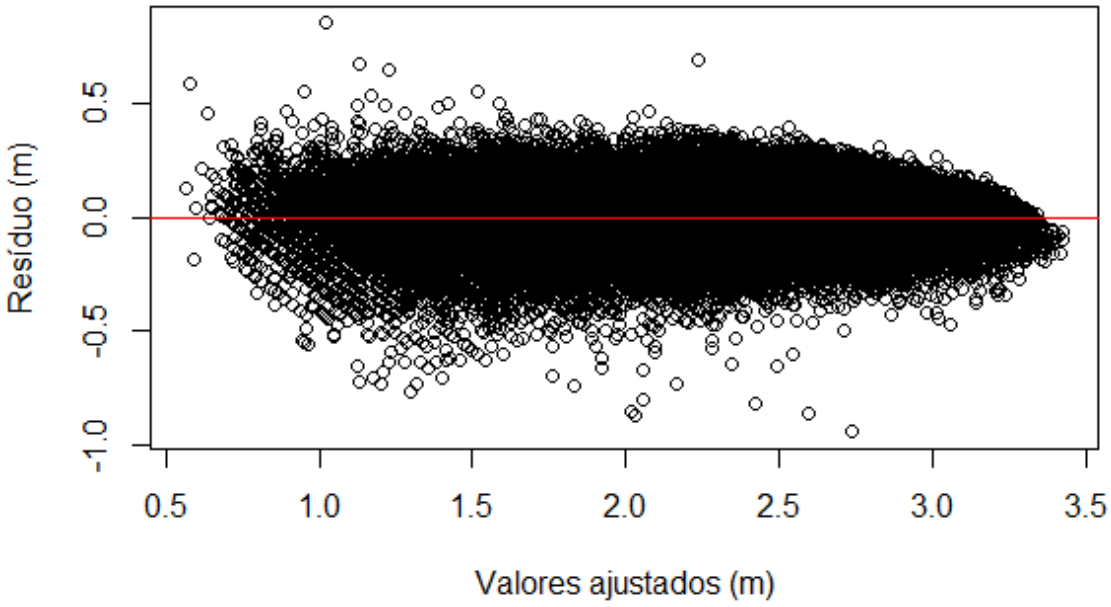

Figura 5 – Distribuição residual do modelo de Potência modificado.
Fonte: O autor (2016).

A partir da distribuição residual do modelo selecionado, é possível observar que o modelo se ajustou bem aos dados. Ao contrário dos modelos ajustados anteriormente, o modelo não estimou altura não condizentes com o conjunto de dados, ou seja, não estimou alturas menores de 1,3 m.

6.1.4 Modelo Embrapa

Tabela 5: AIC dos modelos Embrapa.

Forma de ajuste	AIC
$ht = \beta_0 + \beta_1 DAP + \beta_2 DAP^2 + \beta_3 (DAP.h_{domtalhão}) + \varepsilon$	329.144
$ht = \beta_0 + \beta_1 DAP + \beta_2 DAP^2 + \beta_3 (DAP.h_{domparcela}) + \varepsilon$	318.406

Fonte: O autor (2016).

Por meio dos valores de AIC dos modelos Embrapa utilizando o logarítimo neperiano da altura dominante em nível de talhão e parcela, é possível selecionar o modelo modificado com o logaritmo neperiano da altura dominante em nível de parcela.

Desta forma o gráfico de distribuição residual é gerado apenas para o modelo selecionado (Figura 6).

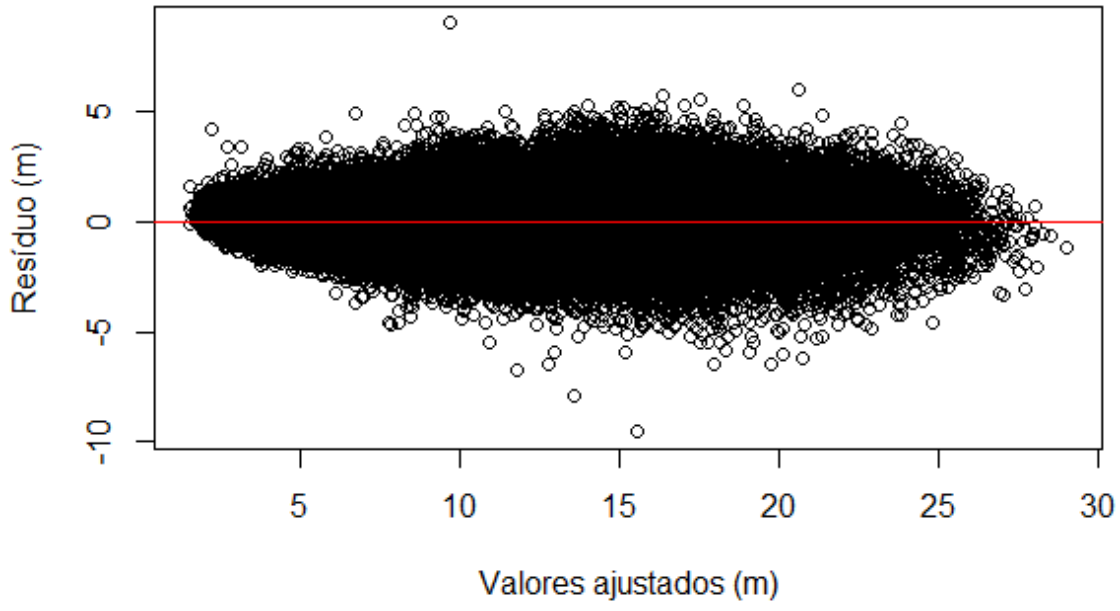

Figura 6 – Distribuição residual do modelo Embrapa.
Fonte: O autor (2016).

É possível observar a partir da distribuição residual do modelo Embrapa uma leve tendência de subestimação para árvores com menores valores de altura,

resultando em valores estimados não condizentes com o conjunto de dados, pois não são mensuradas árvores com altura menor de 1,3 metros e o modelo estimou para árvores das menores classes diamétricas, alturas menores esta. De qualquer forma, o impacto dessas árvores na estimação do volume em nível do povoamento é baixo, por se tratar das árvores das menores classes diamétricas.

6.1.5 Modelo Exponencial

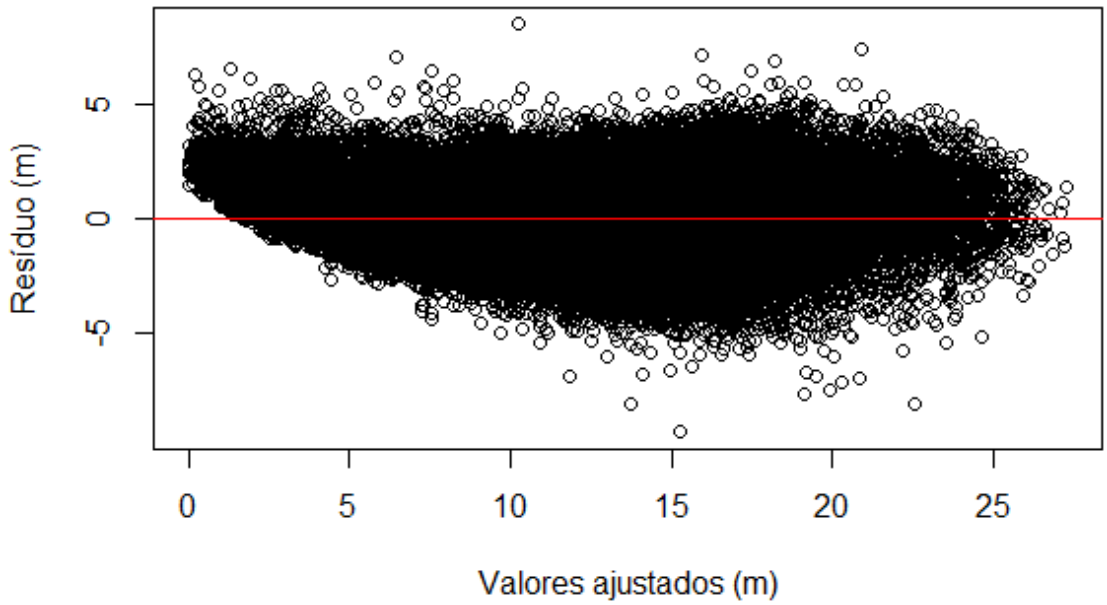

Figura 7 – Distribuição residual do modelo Exponencial.
Fonte: O autor (2016).

A partir da distribuição residual, é possível observar que o modelo subestimou a altura para as árvores das menores classes diamétrica, estimando alturas menores de 1,3 m. Por ser um modelo não linear onde a altura dominante não é natural no modelo, não se ajustou nenhum modelo variante, não sendo necessário calcular os valores de AIC, já que estes são utilizados no presente trabalho apenas para comparar modelos tradicionais com os modelos modificados.

6.1.6 Modelo de Pienaar

Tabela 6: RMSE relativo dos modelos de Pienaar.

Forma de ajuste	RMSE (%)
$\frac{ht}{hdom_{talhão}} = \beta_1 \left(1 - \beta_2\, e^{-\beta_3 \frac{DAP}{Dg_{talhão}}}\right) + \varepsilon$	12,27
$\frac{ht}{hdom_{parcela}} = \beta_1 \left(1 - \beta_2\, e^{-\beta_3 \frac{DAP}{Dg_{parcela}}}\right) + \varepsilon$	14,37

Fonte: O autor (2016)

Por meio dos valores de RMSE relativo dos modelos de Pienaar utilizando o diâmetro médio quadrático e a altura dominante em nível de talhão e parcela, respectivamente, é possível selecionar o modelo de Pienaar com os parâmetros de povoamento em nível de talhão.

Desta forma, o gráfico de distribuição residual é gerado (Figura 8) apenas para o modelo selecionado.

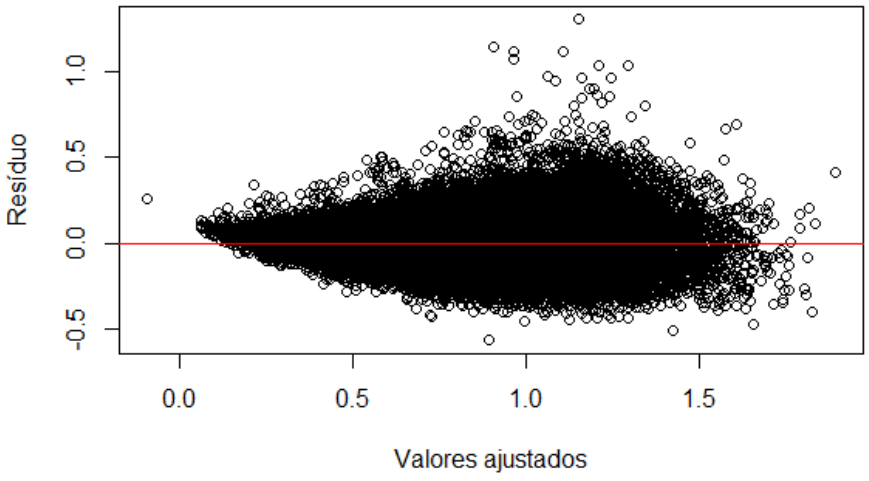

Figura 8 – Distribuição residual do modelo de Pienaar.
Fonte: O autor (2016).

É possível observar a partir da distribuição residual do modelo de Pienaar uma leve tendência de subestimação para árvores com menores valores de altura, resultando na estimação de alturas não condizentes com o conjunto de dados, pois não são mensuradas árvores com alturas menores que 1,3 metros e o modelo estimou para árvores das menores classes diamétricas, alturas menores que 1,3

metros. De qualquer forma, o impacto dessas árvores na estimação do volume em nível do povoamento é baixo, por se tratar das árvores das menores classes diamétricas. Ainda pode-se notar uma dispersão ligeiramente maior para os valores ajustados na classe de 1 metro.

6.2 SELEÇÃO DO MODELO DE RELAÇÃO HIPSOMÉTRICA

Tabela 7: RMSE relativo dos modelos de relação hipsométrica.

Modelo	RMSE (%)
Curtis modificado com a altura dominante em nível de parcela	12,35
Parábola modificado com a altura dominante em nível de parcela	9,98
Exponencial	12,32
Potência modificado com a altura dominante em nível de parcela	10,02
Embrapa com a altura dominante em nível de parcela	9,57
Pienaar com a altura dominante e Dg em nível de talhão	12,27

Fonte: O autor (2016).

Por meio do RMSE relativo (9,57%) e da distribuição residual do modelo, escolheu-se o modelo Embrapa, utilizando a altura dominante em nível de parcela.

Em relação a inclusão da variável altura dominante nos modelos lineares, notou-se uma melhoria significativa em todos os modelos com a inclusão desta variável, devido esta fornecer uma informação que auxiliou o modelo na explicação da altura, pois a relação hipsométrica não é igual para sítios diferentes.

Este resultado, também foi encontrado por Leite (2003).Segundo o autor, a adição da variável altura dominante no modelo, contribui significativamente para a redução da soma dos quadrados, principalmente por permitir representar diferentes capacidades produtivas dos locais onde se encontram as parcelas, pois as relações entre o DAP e a altura total da árvores podem diferir entre parcelas localizadas em áreas boas, médias e ruins.

6.3 AVALIAÇÃO DO IMPACTO DOS OUTLIERS NO MODELO DE RELAÇÃO HIPSOMÉTRICA

Tabela 8: RMSE relativo e contaminação relativa da amostra para todas as abordagens.

Abordagem	RMSE (%)	Outliers (%)
Testemunha	10,54	-
Resíduo padronizado > 2	10,51	4,97
Resíduo padronizado > 3	10,53	0,56
Cook's Distance	10,52	3,51
DFBETA	10,53	3,79
Regressão robusta	11,32	-

Fonte: O autor (2016).

Pode-se notar a partir da Tabela 8 que a abordagem utilizando resíduo padronizado maior que dois como critério na identificação dos *outliers* foi o método que identificou o maior número de *outliers* na amostra, acusando uma contaminação de 4,97 % da amostra.

A partir do RMSE relativo da predição dos modelos ajustados frente as diferentes abordagens, pode-se se notar que todas as abordagens envolvendo a remoção e reajuste dos *outliers* resultaram em uma diminuição do RMSE, porém o resultado é negligenciável devido a diferença ser extremamente pequena, onde o menor RMSE relativo é pertencente a abordagem 1.

É possível observar por meio das figuras 9, 10, 11 e 12 a falta de correlação da redução do RMSE relativo em relação à contaminação da amostra de acordo com as diferentes abordagens do presente trabalho.

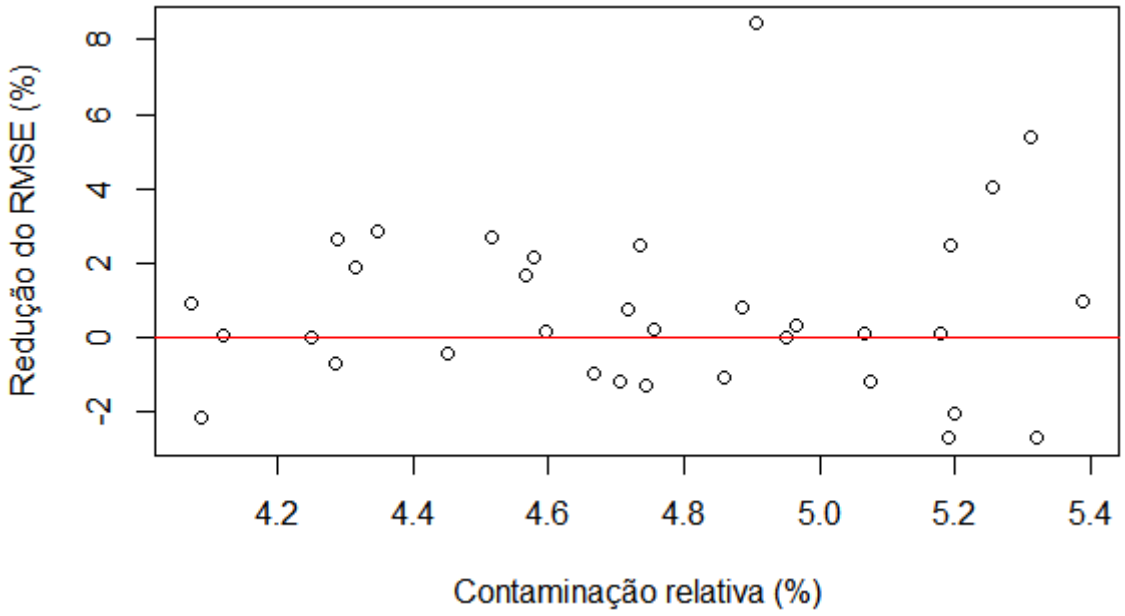

Figura 9 – Relação da redução do RMSE relativo com a contaminação relativa da amostra para a abordagem envolvendo a detecção de *outliers* via resíduo padronizado maior que dois. Fonte: O autor (2016).

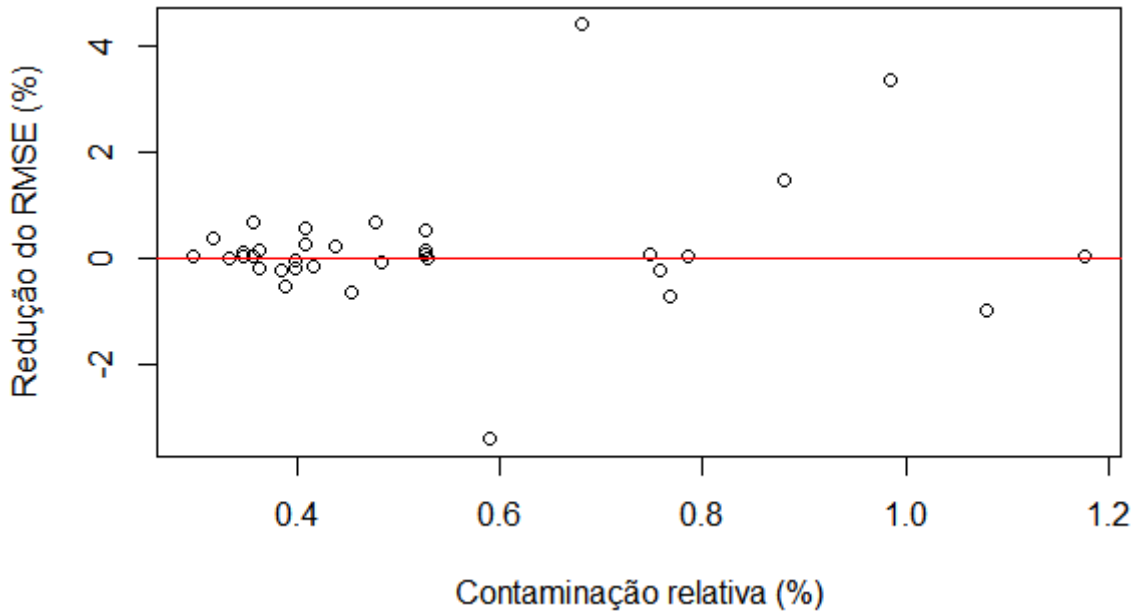

Figura 10 – Relação da redução do RMSE relativo com a contaminação relativa da amostra para a abordagem envolvendo a detecção de *outliers* via resíduo padronizado maior que três. Fonte: O autor (2016).

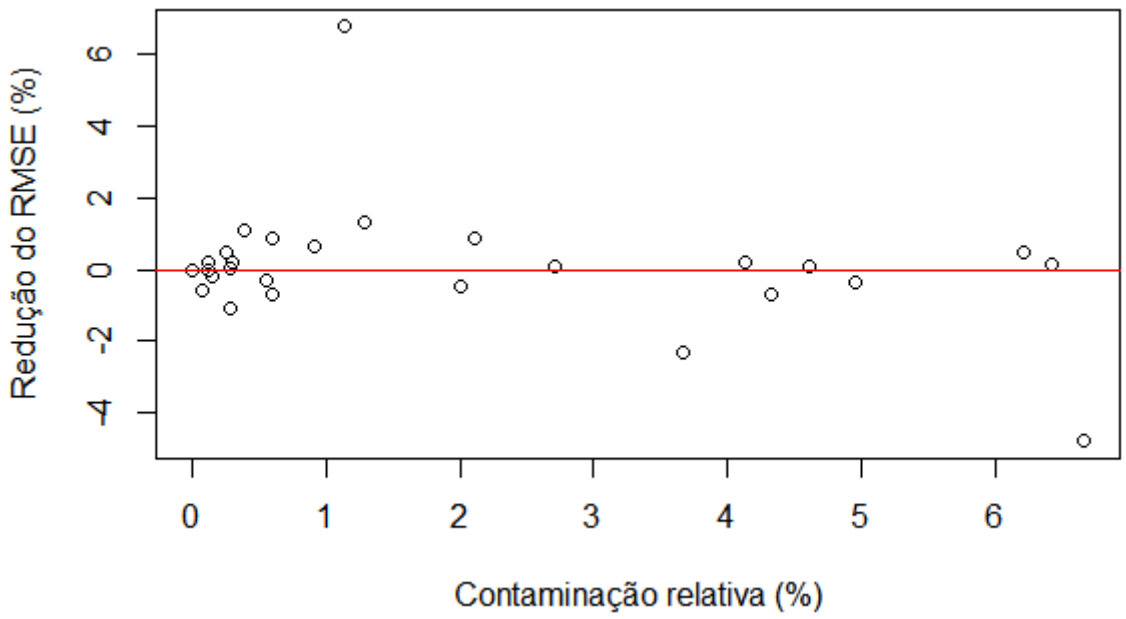

Figura 11 – Relação da redução do RMSE relativo com a contaminação relativa da amostra para a abordagem envolvendo a detecção de *outliers* via Distância de Cook.
Fonte: O autor (2016).

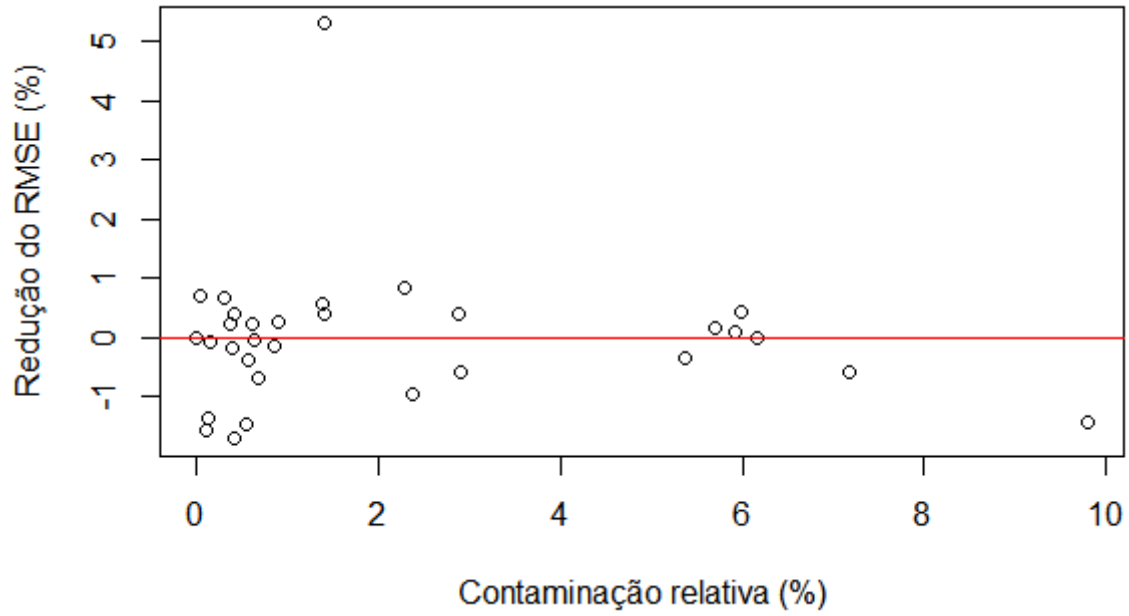

Figura 12 – Relação da redução do RMSE relativo com a contaminação relativa da amostra para a abordagem envolvendo a detecção de *outliers* via DFBETA.
Fonte: O autor (2016).

Para evidenciar a falta de correlação entre estas duas variáveis, foi realizado o teste de correlação de Spearman (devido a amostra não atender os pressupostos para utilização da correlação de Pearson), onde para nenhuma abordagem a correlação foi significativa em um nível de 5% de probabilidade (Tabela 9).

Tabela 9: Coeficiente de correlação de Spearman para a relação entre a redução do RMSE relativo e a contaminação relativa da amostra para cada abordagem.

Abordagem	Coeficiente de correlação (ρ)
Resíduo padronizado > 2	0,2880281^{ns}
Resíduo padronizado > 3	0,3348 ^{ns}
Cook's Distance	0,0149013 ^{ns}
DFBETA	-0,01497555 ^{ns}

Fonte: O autor (2016).

Onde:

ns = não significativo ao nível de 5% de probabilidade pelo teste de correlação de Spearman e * = significativo ao nível de 5% de probabilidade pelo teste de correlação de Spearman.

6.4 AVALIAÇÃO DA PERFORMANCE DO MODELO AJUSTADO PELO MQA

Em relação ao desempenho do modelo onde se utilizou como estimador o método dos mínimos quadrados aparados, este teve um desempenho pior que o modelo testemunha (mínimos quadrados ordinários), isto ocorre devido o modelo utilizar apenas uma fração (aproximadamente 50 %) da amostra de ajuste, ocasionando em perda de informação importante para descrever a relação da altura da árvore com o diâmetro da mesma, resultando em coeficientes não confiáveis.

Apesar de que segundo BARBIERI (2012), o desempenho inferior de um estimador robusto em comparação com um estimador tradicional pode ocorrer quando não existe contaminação da amostra de ajuste, isso ocorreu mesmo em amostras cuja contaminação relativa era maior (Figura 13), não havendo nenhum padrão de melhoria no desempenho do estimador robusto para amostras com maior contaminação relativa.

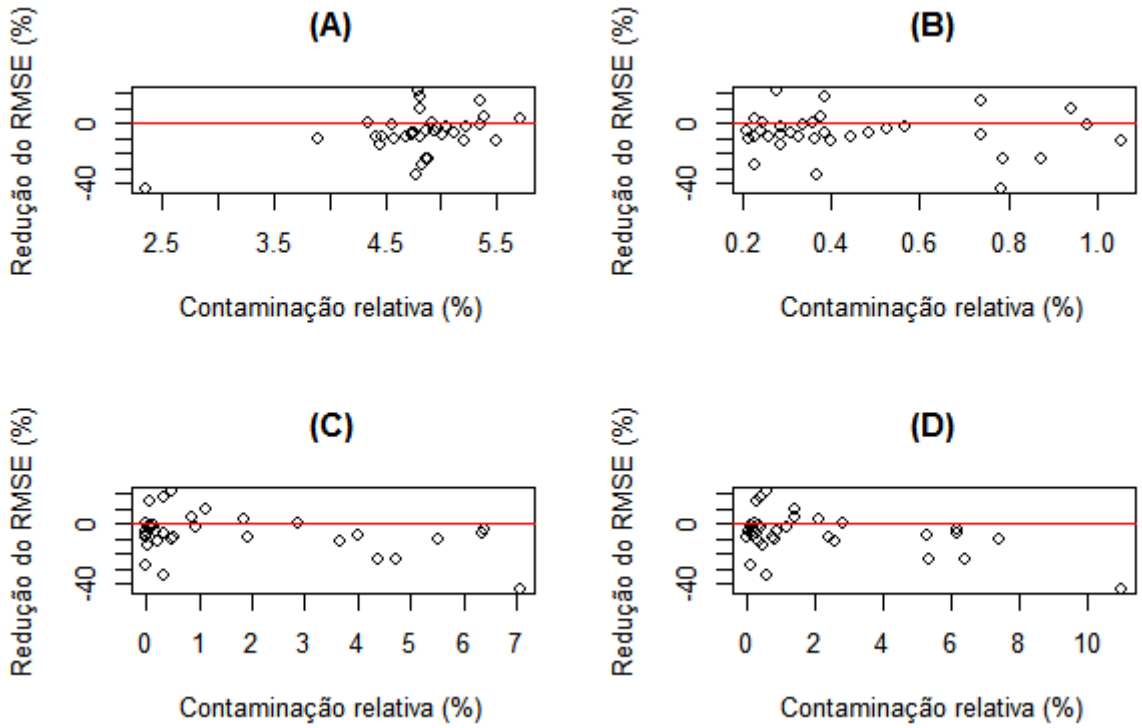

Figura 13 – Relação da redução do RMSE relativo utilizando o estimador robusto com a contaminação relativa da amostra acusada na abordagem 1 (A), 2 (B), 3(C) e 4 (D).
Fonte: O autor (2016).

7 CONSIDERAÇÕES FINAIS

É possível concluir a partir dos resultados expostos que nenhum método de detecção de *outliers* seguido da remoção dos mesmos e reajuste dos modelos teve um efeito significativo no desempenho preditivo do modelo de relação hipsométrica. A utilização de um estimador de alto ponto de ruptura como o método dos mínimos quadrados aparados, culmina em grande perda de informação no modelo, acarretando em predições piores que o modelo de relação hipsométrica ajustado com e sem a remoção dos *outliers*.

Por fim, é importante salientar que antes de realizar o ajuste de relação hipsométrica, deve-se realizar a consistência dos dados, removendo dados que não correspondam a classe de dados onde o modelo será utilizado (utilizar árvores quebradas no modelo para predizer árvores normais ou vice-versa), caso contrário o modelo irá gerar estimativas enviesadas.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at statical model identifation. **IEEE Transactions on Automatic**. Tokyo. V.19, n. 6, p. 716-723, Dez 1974.

ANGELI, A. **Tectonagrandis.** (Supervisão e orientação do Prof. J. L. Stape, Departamento de Ciências Florestais - ESALQ/USP. Atualizado em 05/05/2003). Disponível em: <http://www.ipef.br/identificacao/tectona.grandis.html>. Acesso em: 20 mai. 2015.

BARBIERI, Natália B. **Estimação Robusta para o Modelo de Regressão Logística**. 2012. 59 f. Monografia (Bacharel em Estatística) – Departamento de Estatística, Instituto Federal do Rio Grande do Sul, Porto Alegre, 2012.

BARNET, V.; LEWIS, T. *Outliers in Statical Data*.Universityof Sheffield. p. 188, 1977.

BATISTA, J. L. F.; COUTO, H. T. Z. do; MARQUESINI, M. Desempenho de modelos de relações hipsométricas: estudo de três tipos de floresta. Scientia Forestalis, n.60, p.149- 163, dez. 2001.

BATISTA, J. L. F. Análise de Regressão Aplicada. Piracicaba, 2004. 265 p.

BARROS, A. L. B. P. da. REVISITANDO O PROBLEMA DE CLASSIFICAÇÃO DE PROBLEMAS NA PRESENÇA DE *OUTLIERS* USANDO TÉCNICAS DE REGRESSÃO ROBUSTA, Tese Doutor em Engenharia Teleinformática, Fortaleza, 2013 , p.152, Universidade Federal do Ceará

BULHÕES, S. R. da, LIMA, C. M. V. Comparação de Estimadores de Regressão, 2008. p. 6. Disponível em: <http://www.ime.unicamp.br/sinape/sites/default/files/Resumov1.pdf>. Acesso em 21 mai. 2015.

CALDEIRA, M. V. W.; SCHUMACHER, M. V.; SCHEEREN, L. W.; BARICHELLO, L. R.; WATZLAWICK, L. F. Relação Hipsométrica para *Acaciamearnsii* com Diferentes Idades. EMBRAPA FLORESTAS, Disponível em:<http://ainfo.cnptia.embrapa.br/digital/bitstream/CNPF-2009-09/33611/1/pag57-68.pdf>. Acesso em 25 Mai. 2015.

CARDOSO, D. J. **Avaliação da influência dos fatores de sítio, idade, densidade e posição sociológica na relação hipsométrica para *Pinus taeda* nas regiões central e sudoeste do Paraná.** 1989. 115 f. Dissertação (Mestrado em Engenharia Florestal) – Setor de Ciências Agrárias, Universidade Federal do Paraná, Curitiba.

CHAGAS, E. N. do. EFICIÊNCIA DE ESTIMADORES ROBUSTOS A OBSERVAÇÕES DISCREPANTES EM REGRESSÃO MULTIVARIADA COM A APLICAÇÃO NA ANÁLISE SENSORIAL DE CAFÉ, Tese de Doutorado, Lavras, 2011,p 95.

CHERNICK, M. R.; FRIIS, R. H. Introductory biostatistics for the health sciences: modern applications including bootstrap. New Jersey: John Wiley& Sons, 2003. 406p.

CHICARELU, L. S.; OLIVEIRA, M. C. N. de; POLIZEL, A.; NEPOMUCENO, A. L. A presença de *Outliers* interfere no Teste F e no teste de comparações múltiplas de médias. Disponível em: <http://www.alice.cnptia.embrapa.br/alice/bitstream/doc/574908/1/ID29960.pdf>. Acesso: em 20 mai. 2015.

CUNHA, U. S. da, MACHADO, S. A. do, FILHO, A. F. USO DE ANÁLISE EXPLORATÓRIA DE DADOS E DE REGRESSÃO ROBUSTA NA AVALIAÇÃO DO CRESCIMENTO DE ESPÉCIES COMERCIAS DE TERRA FIRMA NA AMAZÔNIA, Revista árvore, Viçosa – MG , 2002, v.26, n.4, p. 391-402. 2002.

CURTIS, R. O. 1967. Height-diameter and height-diameter age equations for second-growth Douglas fir. Forest Science 13(4):365–375.

DRAPER, N.R.; SMITH, H. Applied regression analysis. 3nd.ed. New York: Wiley, 1998. 706 p.

DRESCHER, R. CRESCIMENTO E PRODUÇÃO DE *TECTONA GRANDIS* LINN EM POVOAMENTOS JOVENS DE DUAS REGIÕES DO ESTADO DE MATO GROSSO – BRASIL, 2004. 144 f. Tese de Doutorado, Santa Maria, 2004. P. 144. (Doutor em Engenharia Florestal) – Área de Concentração de Manejo Florestal, Universidade Federal de Santa Maria, 2012. Disponível em: <http://cascavel.ufsm.br/tede/tde_arquivos/10/TDE-2006-12-01T142839Z-255/Publico/RONALDODRESCHER.pdf>. Acesso em: 20 mai. 2015.

FARCOMENI, A.; VENTURA, L.An overview of robust methods in medical research. Statistical methods in medical research,p. 21-33, 2012.

FERROLI, Paulo Cesar M. et al. Método paramétrico aplicado em design de produtos. Revista Produção On-Line, Florianópolis, v. 7, n. 3, nov. 2007. Disponível em: <http://www.periodicos.ufsc.br/index.php/producaoonline/article/viewFile/4858/4201>. Acesso em: 17 ago. 2008.

FLÓREZ, J. B. CARACTERIZAÇÃO TECNOLÓGICA DA MADEIRA JOVEM DE TECA (*Tectonagrandis*L.f.). Dissertação de mestrado (Mestre). Programa de Pós Graduação em Ciência e Tecnologia da Madeira, Universidade Federal de Labras, Lavras, 2012. 85 f. <http://www.prpg.ufla.br/ct-madeira/wp-content/uploads/2012/07/JeimyBlanco-BDTD.pdf>. Acesso: em 20 mai. 2015.

FONSECA, W. Manual de productores de teca en Costa Rica. Disponível em: <http://www.sirefor.go.cr/Documentos/Reforestacion/2004_Fonseca_ManualProductoresTeca.pdf>. Acesso em: 20 mai. 2015.

FIGUEIREDO, E.O. Reflorestamento com Teca (TectonagrandisL.f.) no Estado do Acre. 2001. 29p. Disponível em: <catuaba.cpafac.embrapa.br/pdf/doc65.pdf>. Acesso em: 20 mai. 2015.

FINGER, C. A. G. Biometria Florestal. Santa Maria.Universidade Federal de Santa Maria, 2006. 284 p.

FILHO, D. B. F.; ROCHA, E. C. da; JÚNIOR, J. A. S. da; PARANHOS, R.; NEVES, J. A. B.; SILVA, M. B. da. Desvendando os Mistérios do Coeficiente de Correlação de Pearson: O retorno. Leviathan, N. 8, p. 66-95, 2014.

GNANADESIKAN, R. Methods for statistical data analysis of multivariate observations. New Jersey: John Wiley, 1997. 353 p.

GRUBBS, F. E., Sample Criteria for Testing Outlying Observations. Ann. Math. Statist.Vol 21.Number 1. 1950, p. 27-58.

ISNOWSKI, J. W.; SIMPSON, J. R.; MONTGOMERY, D. C.; RUNGER, G. C. Resampling methods for variable selection in robust regression Computational Statistics & Data Analysis, v. 43, p. 341-55, 2002.

LADRACH, W.Manejo de plantaciones de la teca para productos sólidos. Disponível em: <http://www.istf-bethesda.org/specialreports/teca_teak/teca.pdf>. Acessoem 20 mai. 2015.

LEITE, H.G.; ANDRADE, V.C.L. 2003. Importância das variáveis altura dominante e altura total em equações hipsométricas e volumétricas. *Revista Árvore*, 27(3):301-310.

LOETSCH, F.; ZOEHRER, F.; HALLER, K. E. Forest inventory.München: BLV, 1973.

v.2, 469p.

KUBOYAMA, F. A. Q. Crescimento de *Tectonagrandis* EM DOIS POVOAMENTOS NO MINICÍPIO DE MIMOSO DO SUL, ESPÍRITO SANTO. 2012. 32 f. Monografia (Engenharia Florestal) – Departamento de Ciência Florestais e da Madeira, Universidade Federal do Espírito Santo, 2012. Disponível em:

<http://www.florestaemadeira.ufes.br/sites/www.florestaemadeira.ufes.br/files/TCC_Filipe%20Akira%20Querino%20Kuboyama.pdf>. Acesso em: 20 mai. 2015.

MACHADO, H. C. da, Detecção de Dados Atípicos e Métodos de Regressão com Alto Ponto de Ruptura. 1997, Campinas, UNICAMP, Dissertação Mestrado em Estatística.

NERO, T. P. P. COMPARAÇÃO DE MODELOS LINEARES E NÃO LINEARES EM RELAÇÕES HIPSOMÉTRICAS PARA CLONES DE *Eucalyptus spp*., NO PÓLO GESSEIRO DO ARARIPE-PE. Dissertação. Pernanbuco, 2012, p. 75.

PORTAL ACTION, 3.4.3 PONTOS INFLUENTES. Disponível em: <http://www.portalaction.com.br/analise-de-regressao/343-pontos-influentes>. Acesso: em 20 mai. 2015.

ROCHA, R. B.; VIEIRA, A. H.; SPINELLI, V. M.; VIEIRA, J. R. Caracterização de fatores que afetam a germinação de Teca (*Tectonagrandis*): temperatura e escarificação.Revista Árvore, Viçosa, v. 35, n.2, mar. 2011. Disponível em: <http://www.scielo.br/scielo.php?pid=S0100-76220110002000005&script=sci_arttext>. Acesso em: 20 mai. 2015.

ROUSSEEUW, P. J. ROBUST REGRESSION AND OUTLIER DETECTION, New Yor, 1987, p. 347.

RONDON NETO, R.M., MACEDO, R.L.G. & TSUKAMOTO FILHO, A.D. Formação de povoamentos florestais com Tectonagrandis L.F. (Teca). Boletim Técnico – Universidade Federal de Lavras. Série Extensão Ano VII, nº33. Lavras-MG, 1998. 29p.

SCHUMACHER, F. X. 1939. A new growth curve and its application to timber yield studies. Journal of Forestry 37:819 – 820.

STOFFELSS, A. ; SOEST J. V. 1953. The main problems in sample plots. Ned. Boschb. Tijdschr. 25:190 – 199.

APÊNDICES

APÊNDICE A – Script dos ajustes e testes envolvendo as abordagens propostas no presente trabalho.

```
#Carrega os dados e remove as linhas em que a altura da árvore é ausente

load("C:/Users/Marco Machado/Dropbox/TCC/Metodologia/2-RelaHipso.rda")
RelaHipso <- RelaHipso[!is.na(RelaHipso$ALT),]
summary (RelaHipso)

#Carrega os pacotes
require(nlme)
#install.packages("outliers")
require (outliers)
#install.packages("robustbase")
require (robustbase)
#install.packages("robustHD")
require(robustHD)

# Função para retornar o erro ponderado de cada modelo

erroponderado <- function(modelo,escala){
e <- vector()
dp <- vector()
for (i in 1:length(modelo)){
  dp[i] <- length (RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT)
    ifelse(escala==1,e[i] <- (sqrt(sum((RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT
     -exp(predict(modelo,RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],])))^2)/(dp[i]-
length(coef (modelo))
       )))/mean(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT),
     ifelse(escala==2,e[i] <-
(sqrt(sum((RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT
                  -(predict(modelo,RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],])))^2)/(dp[i]-
length(coef (modelo))
                  )))/mean(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT),
           ifelse(escala==3,e[i] <-
(sqrt(sum((RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT
                        -
(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$MhDomTal*predict(modelo,RelaHipso[R
elaHipso$CODPROJETO==names(modelo)[i],])))^2)/(dp[i]-length(coef (modelo))
                        )))/mean(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT),
```

```
            e[i] <- (sqrt(sum((RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT
                              -
(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$MhDomPar*predict(modelo,RelaHipso[R
elaHipso$CODPROJETO==names(modelo)[i],]))^2)/(dp[i]-length(coef (modelo))
                              )))/mean(RelaHipso[RelaHipso$CODPROJETO==names(modelo)[i],]$ALT)
       )))
 }
erro <- sum(e*100*dp)/sum(dp)
return(erro)
}

# Ajuste e avaliação dos modelos
#Modelo Curtis
mod.log <- lmList(log(ALT) ~ I(1/DAP)|CODPROJETO, data = RelaHipso)
mod.log2 <- lmList(log(ALT) ~ I(1/DAP)+log(MhDomTal)|CODPROJETO, data = RelaHipso)
mod.log3 <- lmList(log(ALT) ~ I(1/DAP)+log(MhDomPar)|CODPROJETO, data = RelaHipso)
AIC(mod.log,mod.log2,mod.log3)
#mod.log3
erroponderado(mod.log3,1)
plot(fitted.values(mod.log3),residuals(mod.log3),xlab="Valores ajustados (m)",ylab="Resíduo ln(m)")
abline (h=0,col="red",lwd=1)
summary (mod.log3)

#Parabola
mod.parab <- lmList(ALT ~ DAP + I(DAP^2)|CODPROJETO, data = RelaHipso)
mod.parab2 <- lmList(ALT ~ DAP + I(DAP^2)+log(MhDomTal)|CODPROJETO, data = RelaHipso)
mod.parab3 <- lmList(ALT ~ DAP + I(DAP^2)+log(MhDomPar)|CODPROJETO, data = RelaHipso)
AIC(mod.parab,mod.parab2,mod.parab3)
#mod.parab3
plot(fitted.values(mod.parab3),residuals(mod.parab3),xlab="Valores ajustados (m)",ylab="Resíduo
(m)")
abline (h=0,col="red",lwd=1)
summary (mod.log3)
erroponderado (mod.parab3,2)

#Potência 2
mod.P2 <- lmList(log(ALT) ~ log(DAP)|CODPROJETO, data = RelaHipso)
mod.P22 <- lmList(log(ALT) ~ log(DAP)+log(MhDomTal)|CODPROJETO, data = RelaHipso)
mod.P23 <- lmList(log(ALT) ~ log(DAP)+log(MhDomPar)|CODPROJETO, data = RelaHipso)
AIC (mod.P2,mod.P22,mod.P23)
#mod.P23
summary (mod.P23)
```

```
erroponderado(mod.P23,1)
plot(fitted.values(mod.P23),residuals(mod.P23),xlab="Valores ajustados (m)",ylab="Resíduo (m)")
abline(col="red",h=0)


#Embrapa
embrapa <- lmList(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomTal)|CODPROJETO, data = RelaHipso)
embrapa2 <- lmList(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar)|CODPROJETO, data = RelaHipso)
AIC(embrapa,embrapa2)
#embrapa2
erroponderado(embrapa2,2)
plot(fitted.values(embrapa2),residuals(embrapa2),xlab="Valores ajustados (m)",ylab="Resíduo (m)")
abline(col="red",h=0)
summary (embrapa2)

#Ajuste e avaliação dos modelos não lineares

#Exponencial
mod.Exp <- nlsList(ALT ~ exp(b0 + b1/DAP)|CODPROJETO, data = RelaHipso,
          start = list(b0 = 20, b1 = 15),nls.control(maxiter=500))
erroponderado(mod.Exp,2)
plot(fitted.values(mod.Exp),residuals(mod.Exp),xlab="Valores ajustados (m)",ylab="Resíduo (m)")
abline(col="red",h=0)

#Pienaar
Piennar <- nlsList(I(ALT/MhDomTal) ~ a1 * (1 - a2 * exp(-a3 * I(DAP/QTal)))|CODPROJETO,
          data = RelaHipso, start = list(a1 = 1.64, a2 = 0.99, a3 = 0.78))
Piennar2 <- nlsList(I(ALT/MhDomPar) ~ a1 * (1 - a2 * exp(-a3 * I(DAP/QPar)))|CODPROJETO,
           data = RelaHipso, start = list(a1 = 3.348296, a2 = 0.99, a3 = 1.88852))
#Piennar 2

erroponderado(Piennar,3)
erroponderado(Piennar2,4)

plot(fitted.values(Piennar),residuals(Piennar),xlab="Valores ajustados (m)",ylab="Resíduo (m)")
abline(col="red",h=0)

plot(fitted.values(Piennar2),residuals(Piennar2),xlab="Valores ajustados (m)",ylab="Resíduo (m)")
abline(col="red",h=0)

## O modelo weib foi o modelo selecionado
```

```
# Ajuste, detecção e remoção de outliers e predição

################################################################################
####

# Criação dos objetos necessários para o funcionamento do script
load("C:/Users/Marco Machado/Dropbox/TCC/Metodologia/2-RelaHipso.rda")
RelaHipso <- RelaHipso[!is.na(RelaHipso$ALT),]
require("nlme")
require (robustbase)
proj <- lmList(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar)|CODPROJETO, data = RelaHipso)
data <- RelaHipso
Aj <- RelaHipso[1,]
etestl <- vector()
etestnl <- vector()
el1 <- vector()
el2 <- vector()
el3 <- vector()
el4 <- vector()
elr <- vector()
elr99 <- vector()
enlr <- vector()
etestlg <- vector()
etestnlg <- vector()
el1g <- vector()
el2g <- vector()
el3g <- vector()
el4g <- vector()
elrg <- vector()
elrg99 <- vector()
enlrg <- vector()
dim <- vector()
d1 <- vector()
d2 <- vector()
d3 <- vector()
d4 <- vector()
d1g <- vector()
d2g <- vector()
d3g <- vector()
d4g <- vector()

# Todo o procedimento abaixo será rodado 5 vezes
```

```
for (j in 1:34){
RelaHipso <-  data[data$CODPROJETO==names(proj)[j],]
dim[[j]] <- dim(RelaHipso)[1]
for (i in 1:5){
# Criação da amostra de ajuste (20 %) e predição (80%), onde a primeira será formada por 3 classes de diâm.
# com o objetivo de evitar extrapolação no momento da predição
  x <- sample (1:length(RelaHipso$DAP),length(RelaHipso$DAP))
  x <- RelaHipso[x[1:length(RelaHipso$DAP)],]
# Criação de um objeto contendo o números de indivíduos de cada classe
  p <- floor((length(x$DAP)*0.2)/3)
  ifelse(p/length(which(levels(cut (x$DAP,3))[3]==cut (x$DAP,3)))<0.7,
  p3 <- p,p3<-floor(length(which(levels(cut (x$DAP,3))[3]==cut (x$DAP,3)))*0.7))
  ifelse(p/length(which(levels(cut (x$DAP,3))[1]==cut (x$DAP,3)))<0.7,
        p1 <- p,p1<-floor(length(which(levels(cut (x$DAP,3))[1]==cut (x$DAP,3)))*0.7))


  Aj[1:p1,] <- x[levels(cut (x$DAP,3))[1]==cut (x$DAP,3),][1:p1,]
  x <- x[-(which(levels(cut (x$DAP,3))[1]==cut (x$DAP,3))[1:p1]),]
  Aj[(p1+1):(p1+p),] <- x[levels(cut (x$DAP,3))[2]==cut (x$DAP,3),][1:p,]
  x <- x[-(which(levels(cut (x$DAP,3))[2]==cut (x$DAP,3))[1:p]),]
  Aj[(p1+p+1):(p1+p+p3),] <- x[levels(cut (x$DAP,3))[3]==cut (x$DAP,3),][1:p3,]
  x <- x[-(which(levels(cut (x$DAP,3))[3]==cut (x$DAP,3))[1:p3]),]

#Ajuste do modelo linear e não linear convencionalmente e por meio de regressão robusta, respectivamente
  ml <- lm(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar), data = Aj)
  mlr <- ltsReg(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Aj)
  mlr99 <- lqs(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Aj,method = "lqs")
  #mnlr <- nlrob(ALT ~ b0 * (1 - exp(-b1 * DAP^b2)), data = Aj, start = list(b0 = 10,b1 = 0.07,b2 = 0.9),control = list(maxiter = 500))
#Detecção dos outliers e criação de novos objetos sem estes
  ## Resíduo patrodinizado (95,5 %)
  Ajabd1 <- Aj[abs(rstandard(ml))<2,]
  d1[i] <- dim(Ajabd1)[1]/dim(Aj)[1]
  ## Resíduo patrodinizado (99,7 %)
  Ajabd2 <- Aj[abs(rstandard(ml))<3,]
  d2[i] <- dim(Ajabd2)[1]/dim(Aj)[1]
  ## Cook's distance
  Ajabd3 <- Aj[cooks.distance(ml)<4/(p*3),]
  d3[i] <- dim(Ajabd3)[1]/dim(Aj)[1]
  ## DFBETA (para todos os parâmetros)
```

```
  Ajabd4 <- Aj[abs(dfbetas(ml)[,1])<2/sqrt(p*3),]
  Ajabd4 <- Aj[abs(dfbetas(ml)[,2])<2/sqrt(p*3),]
  Ajabd4 <- Aj[abs(dfbetas(ml)[,3])<2/sqrt(p*3),]
  d4[i] <- dim(Ajabd4)[1]/dim(Aj)[1]
#Reajuste dos modelos nos dados sem os outliers
  rml1 <- lm(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Ajabd1)
  rml2 <- lm(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Ajabd2)
  rml3 <- lm(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Ajabd3)
  rml4 <- lm(ALT ~ DAP + I(DAP^2)+I(DAP*MhDomPar),data=Ajabd4)
  pred <- x

# Predição e cálculo do RMSE (%) para cada abordagem
## Testemunha linear
  etestl[i] <- (sqrt(mean((pred$ALT - (predict(ml,pred)))^2))/mean(pred$ALT))*100
## Abordagens envolvendo a detecção dos outliers
  el1[i] <- (sqrt(mean((pred$ALT - (predict(rml1,pred)))^2))/mean(pred$ALT))*100
  el2[i] <- (sqrt(mean((pred$ALT - (predict(rml2,pred)))^2))/mean(pred$ALT))*100
  el3[i] <- (sqrt(mean((pred$ALT - (predict(rml3,pred)))^2))/mean(pred$ALT))*100
  el4[i] <- (sqrt(mean((pred$ALT - (predict(rml4,pred)))^2))/mean(pred$ALT))*100
  elr[i] <- (sqrt(mean((pred$ALT - (coef(mlr)[[1]]+coef(mlr)[[2]]*(pred$DAP)+
                                +coef(mlr)[[3]]*(pred$DAP^2)+
                               coef(mlr)[[4]]*(pred$DAP*pred$MhDomTal)))^2))/mean(pred$ALT))*100
  elr99[i] <- (sqrt(mean((pred$ALT - (coef(mlr99)[[1]]+coef(mlr99)[[2]]*(pred$DAP)+
                               +coef(mlr99)[[3]]*(pred$DAP^2)+
                               coef(mlr99)[[4]]*(pred$DAP*pred$MhDomTal)))^2))/mean(pred$ALT))*100
}
etestlg[j] <-mean (etestl)
el1g[j] <-mean(el1)
el2g[j] <-mean(el2)
el3g[j]<-mean(el3)
el4g[j]<-mean(el4)
elrg[j]<-mean(elr)
elrg99[j]<-mean(elr99)
d1g[j] <- mean (d1)
d2g[j] <- mean (d2)
d3g[j] <- mean (d3)
d4g[j] <- mean (d4)
#elrg[j]<-mean(enlr)
}
```

#Resultado final do desempenho dos modelos e do percentual de contaminação da amostra de acordo

```
#com cada método de detecção

sum((etestlg*dim))/sum(dim)
sum((el1g*dim))/sum(dim)
(1-(sum((d1g*dim))/sum(dim)))*100
sum((el2g*dim))/sum(dim)
(1-(sum((d2g*dim))/sum(dim)))*100
sum((el3g*dim))/sum(dim)
(1-(sum((d3g*dim))/sum(dim)))*100
sum((el4g*dim))/sum(dim)
(1-(sum((d4g*dim))/sum(dim)))*100
sum((elrg*dim))/sum(dim)
sum((elrg99*dim))/sum(dim)

# Relação da redução do RMSE % com a contaminação relativa da amostra

plot ((1-d1g)*100,(1-(el1g/etestlg))*100, ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)")
abline (h=0,col="red")
plot ((1-d2g)*100,(1-(el2g/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)")
abline (h=0,col="red")
plot ((1-d3g)*100,(1-(el3g/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)")
abline (h=0,col="red")
plot ((1-d4g)*100,(1-(el4g/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)")
abline (h=0,col="red")

# Teste de correlação de Spearman
cor.test(((1-d1g)*100),((1-(el1g/etestlg))*100),method = "spearman")
cor.test(((1-d2g)*100),((1-(el2g/etestlg))*100),method = "spearman")
cor.test(((1-d3g)*100),((1-(el3g/etestlg))*100),method = "spearman")
cor.test(((1-d4g)*100),((1-(el4g/etestlg))*100),method = "spearman")

# Desempenho do estimador robusto para diferentes níveis de contaminação
par(mfrow=c(2, 2))
plot ((1-d1g)*100,(1-(elrg/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)",main="(A)")
abline (h=0,col="red")
plot ((1-d2g)*100,(1-(elrg/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)",main="(B)")
```

```
abline (h=0,col="red")
plot ((1-d3g)*100,(1-(elrg/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)",main="(C)")
abline (h=0,col="red")
plot ((1-d4g)*100,(1-(elrg/etestlg))*100,ylab= "Redução do RMSE (%)", xlab = "Contaminação relativa (%)",main="(D)")
abline (h=0,col="red")
```