UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS AGRÁRIAS
CURSO DE ENGENHARIA FLORESTAL

EDUARDO EMÍLIO NADOLNY DE LACERDA

INFLUÊNCIA DA DETECÇÃO DE OUTLIERS NA MODELAGEM DO VOLUME DE ESPÉCIES TROPICAIS DA AMAZÔNIA

CURITIBA
2017

EDUARDO EMÍLIO NADOLNY DE LACERDA

INFLUÊNCIA DA DETECÇÃO DE OUTLIERS NA MODELAGEM DO VOLUME DE ESPÉCIES TROPICAIS DA AMAZÔNIA

Artigo apresentado como requisito parcial à conclusão do curso de Engenharia Florestal do Setor de Ciências Agrárias, da Universidade Federal do Paraná.

Orientador: Prof. Allan Libanio Pelissari

CURITIBA

2017

INFLUÊNCIA DA DETECÇÃO DE *OUTLIERS* NA MODELAGEM DO VOLUME DE ESPÉCIES TROPICAIS DA AMAZÔNIA

Eduardo Emílio Nadolny de Lacerda

RESUMO

As dificuldades de se obter estimativas precisas do volume de árvores oriundas de florestas tropicais, comprometem as atividades de manejo e de conservação. Essa dificuldade deve-se às observações anormais, designadas por *outliers*, os quais apresentam comportamento distinto, e podem representar erros não amostrais ou eventos discrepantes em uma população, cuja exclusão deve seguir regras e técnicas de detecção. Este estudo tem como objetivo analisar a influência das diferentes técnicas de detecção e exclusão de *outliers* no ajuste de modelo de volume comercial individual de espécies tropicais da Floresta Amazônica, a fim de orientar a exclusão das discrepâncias sem comprometer a manutenção das medidas de tendência e variabilidade dos dados. Para isso, foi utilizado um banco de dados de 5.321 árvores cubadas na Floresta Nacional do Jamari, Brasil, cujas técnicas "Box-Plot", "Grubbs", "Peirce" e "Z-Score" foram aplicadas para detectar e, posteriormente, excluir *outliers* para a modelagem do volume pelo modelo Schumacher-Hall. Nos quatro diferentes testes de *outliers* utilizados, ocorre a remoção de dados na extremidade maior ou nas classes de maiores diâmetros, alturas e consequentemente, volumes. Alguns possibilitam maior exclusão de dados, como no caso do teste de Peirce, ao passo que outros geram exclusão menor, como o teste Z-Score. Contudo, os testes de detecção de *outliers* aplicados não indicam de forma satisfatória a exclusão dos dados discrepantes para a composição de uma amostra adequada ao ajuste de modelos de volume individual para espécies da Floresta Amazônica.

Palavras chave: Floresta Amazônica, Equações de volume, Modelo de Schumacher e Hall

INFLUENCE OF OUTLIERS DETECTION IN THE VOLUME MODELING OF TROPICAL SPECIES FROM THE AMAZON FOREST

ABSTRACT

The difficulties of obtaining accurate estimates of volume of trees from tropical forests compromise management and conservation activities. This difficulty is due to abnormal observations, referred to as outliers, which present distinct behavior, and may represent non-sample errors or discrepant events in a population whose exclusion must follow rules and detection techniques. This study aims to analyze the influence of different outliers' detection and exclusion techniques on the adjustment of individual commercial volume model of tropical species in the Amazon Forest, in order to guide the exclusion of discrepancies without compromising the maintenance of trend measures and data variability. For this, a database of 5,321 trees cubed in the Jamari National Forest, Brazil, was used. The "Box-Plot", "Grubbs", "Peirce" and "Z-Score" techniques were applied to detect and, exclude outliers for volume modeling from the Schumacher-Hall model. In the four different outliers' tests used, the removal of data in the larger end or in the classes of larger diameters, heights and, consequently, volumes, occurs. Some allow greater data exclusion, as in the case of the Peirce test, while others generate smaller exclusion, such as the Z-Score test. However, the tests of detection of applied outliers do not indicate satisfactorily the exclusion of the discrepant data for the composition of a sample adapted to the adjustment of individual volume models for species of the Amazon Forest.

Keywords: Amazon forest, Volume equations, Schumacher and Hall model

INTRODUÇÃO

Devido à redução intensa das áreas de floresta nativa, sobretudo pela exploração madeireira ilegal e pelo avanço da fronteira agrícola, a gestão sustentável das florestas tropicais da Amazônia tornou-se imprescindível (GUTIERREZ-VELEZ e MACDICKEN,2008). Para isso, a compreensão teórica das florestas tropicais depende fundamentalmente da qualidade dos dados coletados sobre as mesmas (PHILLIPS et al., 2002), uma vez que as diversas atividades de manejo e conservação das florestas tropicais requerem estimativas seguras e precisas do volume e biomassa das árvores (COLE e EWEL, 2006; RIBEIRO et al., 2014).

No entanto, diversos fatores podem dificultar a qualidade das estimativas da produção das florestas naturais (AKINDELE e LEMAY, 2006; THAINES et al., 2010; IGBINOSA e AMOO, 2014). Quando considerada a complexidade e a diversidade das formações florestais amazônicas, a carência de equações específicas para quantificação do volume das árvores comerciais ainda é evidente (MOURA, 1994; ROLIM et al., 2006; HIRAMATSU, 2008). As florestas tropicais possuem considerável heterogeneidade, na composição de espécies, nas idades, na forma e na estrutura das árvores, que dificulta o desenvolvimento de equações estatisticamente precisas para estimativa de volume de toda a população.

De acordo com Cysneiros (2016) em seu trabalho de modelagem volumétrica para a Floresta Nacional do Jamari, devido à constatada heterogeneidade e extensão do banco de dados, averiguou-se ser necessária uma etapa de pré-processamento para filtrar possíveis dados discrepantes, denominados também de *outliers*.

Os *outliers* são dados que apresentam comportamento distinto dos demais, podendo representar erros não amostrais ou eventos discrepantes em uma população (GRUBBS, 1969), como também podem representar indicativos importantes sobre modelos matemáticos, como: incompatibilidade com o conjunto de dados e omissão de variáveis importantes (CUNHA et al.,

2002). Além disso, na análise de dados ambientais, a ocorrência de *outliers* pode ser comum em que essas discrepâncias têm impacto expressivo na interpretação de análises estatísticas (SABINO et al., 2014). No entanto, esses dados não podem ser excluídos automaticamente, existindo técnicas e regras para sua detecção e tomada de decisão (DRAPPER e SMITH, 1998; HODGE e AUSTIN, 2004).

A influência das diferentes técnicas de detecção de *outliers* é constantemente empregada em diversas áreas das ciências, como na análise ambiental da qualidade da água (SABINO et al., 2014), séries espaço-temporais de precipitação (SILVA, 2012), para modelos de regressão logit binominal (NEPOMUCENA e CIRILLO, 2009), para regressão logística (BEDRICK e HILL, 1990), entre outros. Dessa forma, importantes estudos foram desenvolvidos visando aprimorar as técnicas de detecção e exclusão desses valores, como: Grubbs (1969), Barnett e Lewis (1994) e Hodge e Austin (2004), a fim de orientar o tratamento de dados para a realização de análises estatísticas.

Este estudo teve como objetivo analisar a influência das diferentes técnicas de detecção e exclusão de *outliers*, no ajuste de modelo de regressão para volume comercial individual de espécies tropicais da Floresta Amazônica brasileira, a fim de orientar a exclusão das discrepâncias sem comprometer a manutenção das medidas de tendência e de variabilidade dos dados. Para isso, foi testado o teste de *Grubbs* a 95 e 99% de probabilidade, Critério de *Peirce*, teste *Z*-score e *Box-plot*, sobre o ajuste de um modelo de estimativa do volume de árvores.

MATERIAL E MÉTODOS

Base de dados

O banco de dados utilizado no presente estudo compreende a 5.231 árvores cubadas nos anos de 2014 e 2015, em uma Unidade de Manejo Florestal na Floresta Nacional do Jamari, Brasil, no Norte do estado de Rondônia, entre as coordenadas geográficas 09º 00' 00'' S a 09º 30' 00'' S e 62º44' 05'' W a 63º 16' 64'' W (ZACHOW, 1991). Na área predominam as tipologias de Floresta Ombrófila, Aberta e Densa (RADAM, 1988), com diversas espécies de interesse florestal. O método de Smalian (1) foi adotado para a cubagem do volume comercial, correspondido entre a base do fuste até o ponto de inversão morfológica.

$$v = \frac{g1+g2}{2}.l \quad (1)$$

Para a estimativa do volume individual das árvores foi empregado o modelo de Schumacher e Hall (2) ajustado para a mesma base de dados por Cysneiros (2016), selecionado dentre 11 modelos de volume tradicionais. Esse modelo é consagrado na área florestal por apresentar boa aderência as diversas tipologias florestais (ROLIM et al., 2006; COLPINI et al., 2009; THAINES et al., 2010; TONINI & BORGES, 2015). Outra vantagem apresentada por este modelo é a individualização das três possíveis variáveis de seleção de *outliers* no ajuste de modelos para estimativa do volume de árvores.

$$Ln\, v = -8{,}2734 + 1{,}8040\,.\, Ln\, d + 0{,}7629\,.\, Ln\, h \qquad (2)$$

Em que: v = volume comercial sem casca (m³); d = diâmetro altura do peito (cm), medido a 1,3 m do solo; h = altura comercial (m); ln = logaritmo neperiano; εi = erro associado.

Tratamentos de *outliers*

Os tratamentos avaliados nesta pesquisa foram compostos pelas seguintes técnicas de detecção de *outliers*: a) teste de *Grubbs* a 95 e 99% de probabilidade, b) Critério de *Peirce*, c) gráfico *Box-plot* e d) teste Z-score, descritos abaixo. Esses tratamentos foram aplicados por variável de seleção, sendo o diâmetro à altura do peito em centímetros (*d*), a altura comercial em metros (*h*) e o volume individual comercial em metros cúbicos (*v*). Os resultados gerados pelos diferentes tratamentos foram comparados com uma testemunha, composta pela totalidade das amostras.

Gráfico Box-plot

O gráfico Box-plot é frequentemente empregado em análises estatísticas para explicitar o centro e a dispersão de uma distribuição de dados, além de identificar valores discrepantes da distribuição (TRIOLA, 2005). Nesse gráfico (Figura 1), a caixa central é formada pelo 1º e 3º quartil, em que a linha divisória corresponde ao valor da mediana ou 2º quartil da distribuição.

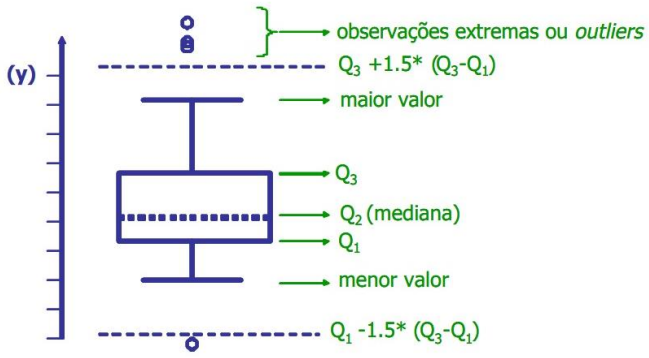

Figura 1: Gráfico Box-Plot (CONCEIÇÃO et al., 2010).

As hastes laterais desses gráficos correspondem aos limites inferiores (3) e superiores (4) da distribuição, assim, todos os valores localizados fora desses limites, são classificados como *outliers*. O box-plot pode ser considerado um método simples de detecção de *outliers*, pois verifica as discrepâncias apenas nos extremos das distribuições.

$$L.I. = Q1 - 1{,}5\,(Q3 - Q1) \qquad (3)$$

$$L.S. = Q3 + 1{,}5\ (Q3 - Q1) \qquad (4)$$

Em que: Q1 = primeiro quartil da distribuição de dados; Q3 = terceiro quartil.

Teste de Grubbs

No teste de Grubbs os valores suspeitos de serem *outliers* foram comparados à média da distribuição (GRUBBS, 1969). Para isso, é utilizado o desvio padrão como denominador ao invés da amplitude, como no teste de Dixon para conjuntos de dados com até dez. Assim, o valor G (5) é calculado para todos os valores suspeitos dessa distribuição comparado com o G crítico, sendo *outliers* quando $G > G_{crítico}$. De acordo com Alfassi et al.,(2005), os seguintes passos são aplicados: 1) ordenar os dados em ordem crescente, 2) calcular a média da amostra 3) determinar a diferença do valor suspeito em relação à média, 4) calcular o desvio padrão com base nos resultados dos passos anteriores, 5) obter o valor de G para os dados suspeito do conjunto, conforme indica a equação (5), e por fim comparar o valor de G, com o valor crítico (ELLISON et al., 2009). Esse teste foi aplicado para os níveis de 95% e 99% de significância.

$$G = \frac{d_i}{\sigma} \qquad (5)$$

Em que: d_i = diferença entre o valor suspeito e a média; σ = desvio padrão.

Teste Z-score

O valor Z-score (6) é uma medida de posição de um determinado valor em uma distribuição de dados. Esse valor indica o quanto uma medida se distância da média em termos de desvio padrão. Seus valores variam de -3 a 3, que, corresponde a 99,72 % da área ocupada em uma distribuição normal dos dados. O valor 3 indica que um determinado valor está três desvios padrões acima da média, enquanto um valor -3 indica está três desvios padrões abaixo da média. Um Z-score muito alto significa que um valor está muito fora de um padrão de dados,

podendo ser considerado um *outlier*. Esse teste [e calculado segundo a equação (6) somente aos dados suspeitos de serem *outliers*.

$$Z = \frac{y-\bar{y}}{\sigma} \qquad (6)$$

Em que: y = valor observado; $\bar{y}$ = média dos valores observados; σ = desvio padrão.

Conforme Sarabando (2010), com n indicando o tamanho da amostra de dados analisada, tem-se que se $n \geq 1000$, como no caso do banco de dados deste estudo, então: se $-3.3 \geq$ Z-score ≥ 3.3 o valor é considerado um *outlier*, caso contrário, o valor não é um *outlier*.

Critério de Peirce

Menos utilizado que as outras técnicas citadas, o critério de Peirce se baseia na teoria da probabilidade para a identificação de valores discrepantes em uma distribuição. Nesse teste, é calculada a diferença máxima permitida (7), entre um determinado valor e a média da distribuição, sendo considerado *outlier*, os valores superiores a R ($|x_i - \bar{x}| > |x_i - \bar{x}|max$).

De acordo com Rossi (2003), os seguintes passos são aplicados: 1) calcula-se a média (x) e o desvio padrão (s) da amostra de dados sendo analisada; 2) para quaisquer medidas de dados suspeitas, obtém-se a diferença entre o valor suspeito e a média da amostra de dados; 3) em seguida calcula-se a distância máxima permitida. Obtém-se o valor de *R* correspondente ao tamanho de conjunto de dados, a partir da tabela de valores críticos; 4) considera-se como *outlier* os valores que forem maiores que a distância máxima permitida; 5) se isso ressaltar na identificação de algum *outlier*, assumir o caso de duas observações suspeitas e reaplicar o teste, mantendo os valores originais da média, desvio padrão e tamanho da amostra. Repete-se os cálculos em sequência crescente conforme o número de possibilidades de valores duvidosos até que não haja mais dados que precisem ser eliminados; 6) posteriormente, eliminam-se os dados

que forem identificados como *outliers,* calcula-se novamente a média e desvio padrão do novo conjunto de dados reduzido e retorna ao passo 2; 7) repete-se a aplicação do método até que não sejam identificados novos *outliers*.

$$R = \frac{|y_i - \bar{y}|max}{\sigma} \qquad (7)$$

Em que: y = valor observado; $\bar{y}$ = média dos valores observados; σ = desvio padrão; *max* = máximo valor.

Análise dos tratamentos

Para a avaliação do efeito dos diferentes tratamentos, foi analisada a correlação linear entre as variáveis e as estatísticas descritivas do conjunto de dados, para os diferentes tratamentos e variáveis de seleção. Com o objetivo de avaliar as tendências de distribuição dos dados foram construídos histogramas por tratamento e variável de seleção.

Foi empregado o método de estratificação de distribuição diamétrica, que utiliza valores empíricos ou se apoia em fórmulas matemáticas para a definição do número e da amplitude adequada de classes, de acordo com características da amostra (CYSNEIROS, 2016). Para isso foi empregada a fórmula de Sturges para definição do número e intervalo de classes para cada variável (8), considerando todas as amostras (testemunha), em que foram selecionados treze classes de diâmetro e treze classes de altura.

$$K = 1 + 3{,}322\,.\,Ln\ (n) \qquad (8)$$

Em que: K = número de classes; e n = número de indivíduos amostrados.

Após o ajuste do modelo de volume de Schumacher-Hall por meio de regressão linear, foi avaliado o efeito dos tratamentos sobre as estatísticas de ajuste e precisão, como: coeficiente de determinação (9), erro padrão relativo da estimativa (10) e a análise gráfica dos resíduos.

Todos os cálculos foram realizados por meio do *software* SAS 9.0 e com auxílio de planilhas eletrônicas do Microsoft Excel.

$$R^2 = \left[\left(1 - \left(\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)^2\right)\right] \qquad (9)$$

$$Syx\ \% = \left[\frac{\left(\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{(n-p)}\right)}{\bar{y}}\right].100 \qquad (10)$$

Em que: y_i = valor observado; $\widehat{y}_i$= valor estimado pelo modelo; n = número de observações; p = número de coeficientes do modelo; e $\bar{y}$ = média dos valores observados da variável dependente.

Com base nos resultados apresentados por cada método, foram sugeridas estratégias para a exclusão dos valores discrepantes, de forma a preconizar a manutenção das medidas de tendência e variabilidade dos dados. Essas estratégias visam aumentar a correlação entre variáveis, indo de encontro à melhoria dos ajustes e estimativas volumétricas.

RESULTADOS

Tratamentos de *outliers*

Os testes de identificação de *outliers* aplicados apresentaram influência diferenciada sobre as estatísticas descritivas das variáveis de seleção (Tabela 1). As alterações mais drásticas para o conjunto de dados foram realizadas considerando o *d* e *v* como variáveis de seleção, com redução evidente da amplitude (λ) dos dados. Quanto à variabilidade dos dados (σ, σ^2 e CV%), a variável de seleção *h* surtiu menor efeito, onde a maior redução do CV% e da σ^2 foi ocasionada pela variável de seleção *v*.

Dentre os testes aplicados, o critério de Peirce foi o mais seletivo, removendo um conjunto maior de observações, como 1.472 na variável *d*, 1.201 na variável *h* e 414 na variável *v*, causando, assim, as maiores alterações no banco de dados, a maior influência nas estatísticas descritivas e diminuição drástica da amplitude dos dados (Tabela 1). Por outro lado, o teste Z-score gerou as alterações mais suaves dentre os testes, removendo no máximo 2,16 % dos dados na variável *v*, e um mínimo de 0,21% dos dados na variável *h*, correspondendo a somente 113 e 11 dados respectivamente. As médias dos dados não sofreram significativa alteração em relação à testemunha (Tabela 1).

Tabela 1: Influência dos testes de identificação de *outliers* nas estatísticas descritivas do conjunto de dados para as variáveis de seleção diâmetro a 1,3 m *d*, altura comercial *h* e volume comercial individual *v.*

Variável	Tratamento	n	out	out %	$\bar{y}$	σ	σ^2	CV%	λ
d	Testemunha	5231	-	-	81,54	22,44	503,67	27,52	195,00
	Box-Plot	5020	211	4,03	78,65	17,31	299,49	22,00	77,00
	Grubbs 95%	5000	231	4,42	78,46	17,07	291,42	21,76	75,73
	Grubbs 99%	5092	139	2,66	79,42	18,35	336,90	23,11	89,00
	Z-score	5153	78	1,49	80,18	19,53	381,36	24,35	98,00
	Peirce	3759	1472	28,14	70,49	9,42	88,83	13,37	38,81
h	Testemunha	5231	-	-	20,44	5,46	29,82	26,72	38,18
	Box-Plot	5217	14	0,27	20,39	5,39	29,04	26,43	31,02
	Grubbs 95%	5064	167	3,19	20,21	5,02	25,25	24,86	21,46
	Grubbs 99%	5202	29	0,55	20,37	5,34	28,47	26,19	27,20
	Z-score	5220	11	0,21	20,40	5,40	29,18	31,35	26,48
	Peirce	4030	1201	22,96	18,15	3,73	13,93	20,57	19,42
v	Testemunha	5231	-	-	7,90	5,69	32,48	72,18	54,33
	Box-Plot	4885	346	6,61	6,76	3,51	12,31	51,93	16,19
	Grubbs 95%	4955	276	5,28	6,94	3,75	14,04	54,00	17,72
	Grubbs 99%	5082	149	2,85	7,27	4,28	18,34	58,95	21,24
	Z-score	5118	113	2,16	7,38	4,49	20,14	60,79	23,53
	Peirce	4817	414	7,91	6,61	3,31	10,98	50,11	14,87

n = número de amostras, out = número de amostras excluídas, out% = porcentagem das amostras excluídas, $\bar{y}$ *= média das amostras,* σ *= desvio padrão da amostra,* σ^2 *= variância da amostra*, CV% *= coeficiente da variação e* λ *= amplitude da amostra.*

A distribuição de frequência das variáveis de seleção, antes e após aplicação dos testes (Figura 2), evidenciou uma tendência comum dos critérios de detecção de *outliers*, caracterizada pela remoção apenas dos maiores valores, localizados nas últimas classes.

Essa tendência também pode ser observada pela redução drástica da amplitude em todos os métodos de detecção de *outliers* (Figura 2), indicando que os valores nas extremidades de maior valor ou nas classes de maiores diâmetros foram excluídos e assim, diminuindo a amplitude dos dados (Tabela 2).

As amostras foram divididas em treze classes de diâmetro e altura e catorze classes de volume. Em todos os testes de detecção de *outliers*, as duas primeiras classes de diâmetro foram mantidas inalteradas, enquanto nas últimas seis classes todos os dados foram considerados *outliers* e excluídos (Tabela 2). Nas classes de altura, no teste de Grubbs 95% e 99% a primeira classe ($5 < 8$ m) teve dados excluídos, enquanto nos testes de Box-plot, Z-score e Peirce tiveram novamente somente as duas últimas classes com dados excluídos (Tabela 2). Para as classes de volume em todos os testes de *outliers*, as últimas oito classes tiveram todos os dados excluídos do total das amostras, enquanto as primeiras classes se mantiveram inalteradas.

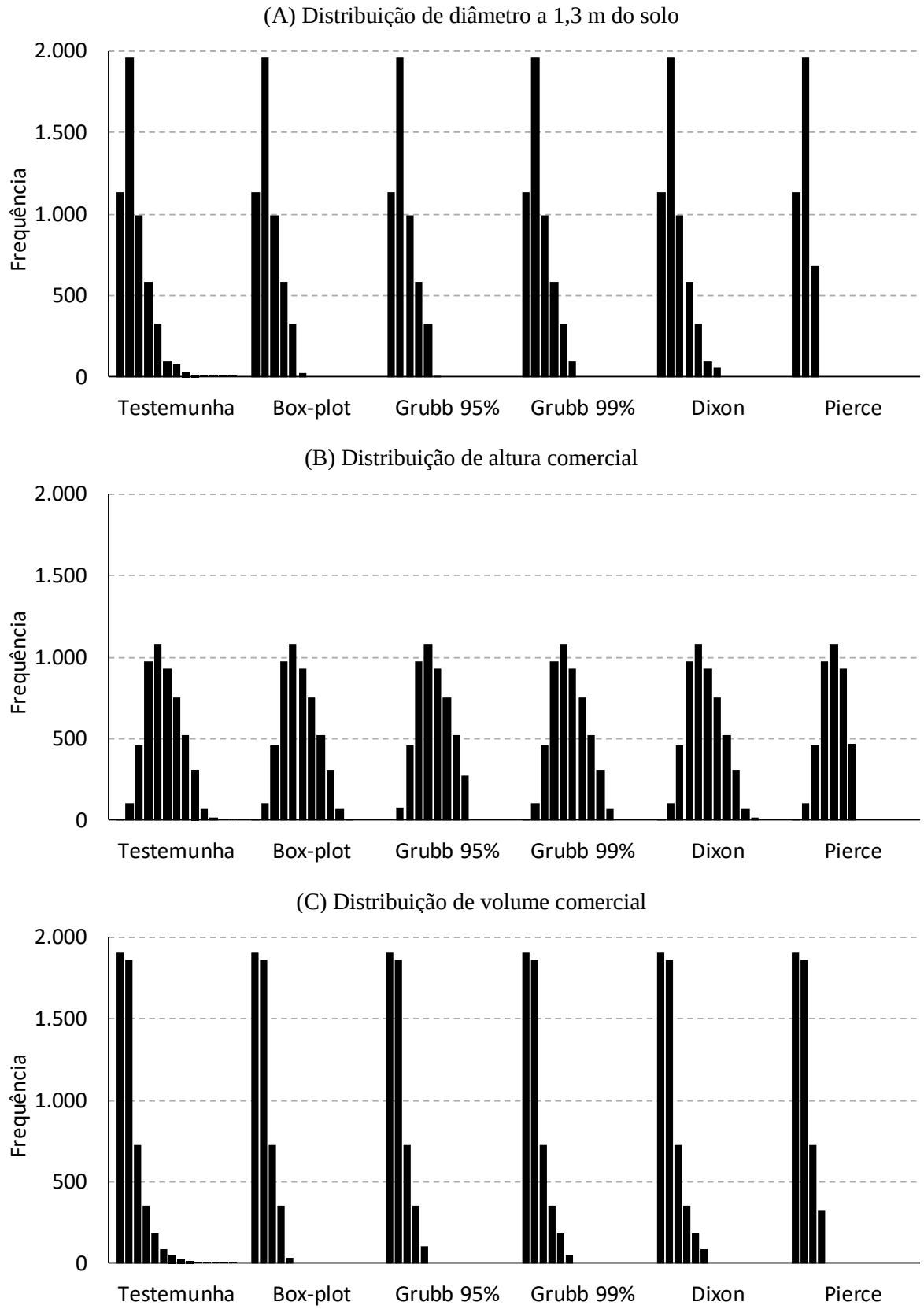

Figura 2: Influência dos testes de identificação de *outliers* na distribuição de frequência das variáveis de seleção *d*, *h* e *v*.

Tabela 2: Influência dos testes de identificação de *outliers* nas diferentes classes nas variáveis *d*, *h* e *v*.

	Classe	Testemunha	Box-plot	Grubbs 95%	Grubbs 99%	Z-Socre	Peirce
d	050<065	1136	1136	1136	1136	1136	1136
	065<080	1961	1961	1961	1961	1961	1961
	080<095	995	995	995	995	995	682
	095<110	580	580	580	580	580	0
	110<125	323	323	323	323	323	0
	125<140	97	25	5	97	97	0
	140<155	78	0	0	0	61	0
	155<170	32	0	0	0	0	0
	170<185	11	0	0	0	0	0
	185<200	8	0	0	0	0	0
	200<215	7	0	0	0	0	0
	215<230	2	0	0	0	0	0
	230<245	1	0	0	0	0	0
	Classe	Testemunha	Box-plot	Grubbs 95%	Grubbs 99%	Z-Socre	Peirce
h	05<08	10	10	0	5	10	10
	08<11	106	106	75	106	106	106
	11<14	457	457	457	457	457	457
	14<17	971	971	971	971	971	971
	17<20	1083	1083	1083	1083	1083	1083
	20<23	933	933	933	933	933	933
	23<26	750	750	750	750	750	470
	26<29	525	525	525	525	525	0
	29<32	304	304	270	304	304	0
	32<35	69	69	0	68	69	0
	35<38	17	9	0	0	12	0
	38<41	5	0	0	0	0	0
	41<44	1	0	0	0	0	0
	Classe	Testemunha	Box-plot	Grubbs 95%	Grubbs 99%	Z-Socre	Peirce
v	01<05	1909	1909	1909	1909	1909	1909
	05<09	1861	1861	1861	1861	1861	1861
	09<13	726	726	726	726	726	726
	13<17	354	354	354	354	354	321
	17<21	180	35	105	180	180	0
	21<25	88	0	0	52	88	0
	25<29	54	0	0	0	0	0
	29<33	27	0	0	0	0	0
	33<37	15	0	0	0	0	0
	37<41	6	0	0	0	0	0
	41<45	5	0	0	0	0	0
	45<49	3	0	0	0	0	0
	49<53	2	0	0	0	0	0
	53<57	1	0	0	0	0	0

Considerando a correlação entre as variáveis, a aplicação dos testes nas diferentes variáveis de seleção também apresentou resultados variados (Tabela 3), com aumento da correlação *d* x *h* para as variáveis de seleção *d* e *h*. Por outro lado, houve redução da correlação *d* x *v* para as variáveis de seleção de *outliers d* e *v*, com aumento dessa correlação apenas quando considerado *h* como variável de seleção. O critério de Peirce ocasionou aumento expressivo na correlação entre as variáveis na maioria dos casos, com a maior influência dentre os testes aplicados (Tabela 3).

Tabela 3: Influência dos testes de identificação de *outliers* na correlação entre as variáveis

Tratamento	Correlação entre variáveis								
	d			h			v		
	d x *h*	*d* x *v*	*h* x *v*	*d* x *h*	*d* x *v*	*h* x *v*	*d* x *h*	*d* x *v*	*h* x *v*
Testemunha	0,124	0,786	0,348	-	-	-	-	-	-
Box-Plot	0,141	0,731	0,392	0,124	0,786	0,345	0,090	0,731	0,429
Grubbs 95%	0,137	0,728	0,390	0,129	0,790	0,335	0,096	0,745	0,418
Grubbs 99%	0,133	0,746	0,379	0,126	0,788	0,343	0,102	0,766	0,388
Z-score	0,126	0,760	0,367	0,124	0,786	0,346	0,105	0,772	0,377
Peirce	0,150	0,549	0,517	0,156	0,802	0,329	0,082	0,719	0,437

Ajuste do modelo de Schumacher-Hall

O ajuste do modelo de volume apresentou erros de estimativa elevados (Tabela 4), devido à alta variabilidade das variáveis empregadas. A utilização das variáveis de seleção *d* e *h* para a detecção de *outliers* apresentou pouco efeito no ajuste do modelo, com redução pouco expressiva do erro padrão da estimativa (Syx%). Porém, ao considerar o *v* como variável de seleção o efeito foi positivo, gerando redução de até 10% do erro padrão da estimativa.

O método de Peirce, ao considerar o *d* como variável de seleção, apresentou um considerável efeito negativo em relação ao coeficiente de determinação (R^2). Porém ao se

considerar *h* como variável de seleção se observou um efeito positivo mínimo. Em todos os testes de identificação de outliers, considerando a variável *v*, se observou uma redução do coeficiente de determinação, ocasionado pela retirada das variáveis dependentes das últimas classes diminuindo sua variabilidade total. Porém essa redução paras as três variáveis de seleção é mínima em todos os testes aplicados.

Tabela 4: Influência dos testes de identificação de *outliers* e variáveis de seleção no ajuste do modelo de volume de Schumacher-Hall para uma floresta natural na Amazônia**.**

Variável	Tratamento	R^2	Syx %
d	Testemunha	0,687	40,424
	Box-Plot	0,638	38,920
	Grubbs 95%	0,636	38,960
	Grubbs 99%	0,655	38,811
	Z-score	0,670	38,803
	Peirce	0,512	34,795
h	Testemunha	0,687	40,424
	Box-Plot	0,687	40,460
	Grubbs 95%	0,682	40,619
	Grubbs 99%	0,686	40,428
	Z-score	0,687	40,437
	Peirce	0,695	41,476
v	Testemunha	0,687	40,424
	Box-Plot	0,643	31,046
	Grubbs 95%	0,655	31,774
	Grubbs 99%	0,665	34,131
	Z-score	0,667	35,089
	Peirce	0,634	30,304

Analisando o efeito dos testes de *outliers* na distribuição gráfica dos resíduos gerados pelo ajuste do modelo (Figura 3), foi possível observar a mesma tendência explicitada na análise dos histogramas, onde foram excluídas principalmente as árvores de maiores diâmetros. Assim, os maiores resíduos presentes nas classes intermediárias foram mantidos na base de dados após a aplicação dos testes, não alterado a variabilidade e heterocedasticidade dos resíduos. É também observado nas classes de diâmetro com a maior concentração de árvores, uma

tendência a superestimativas com resíduos perto de -400% mesmo após os testes de *outliers* (FIGURA 3).

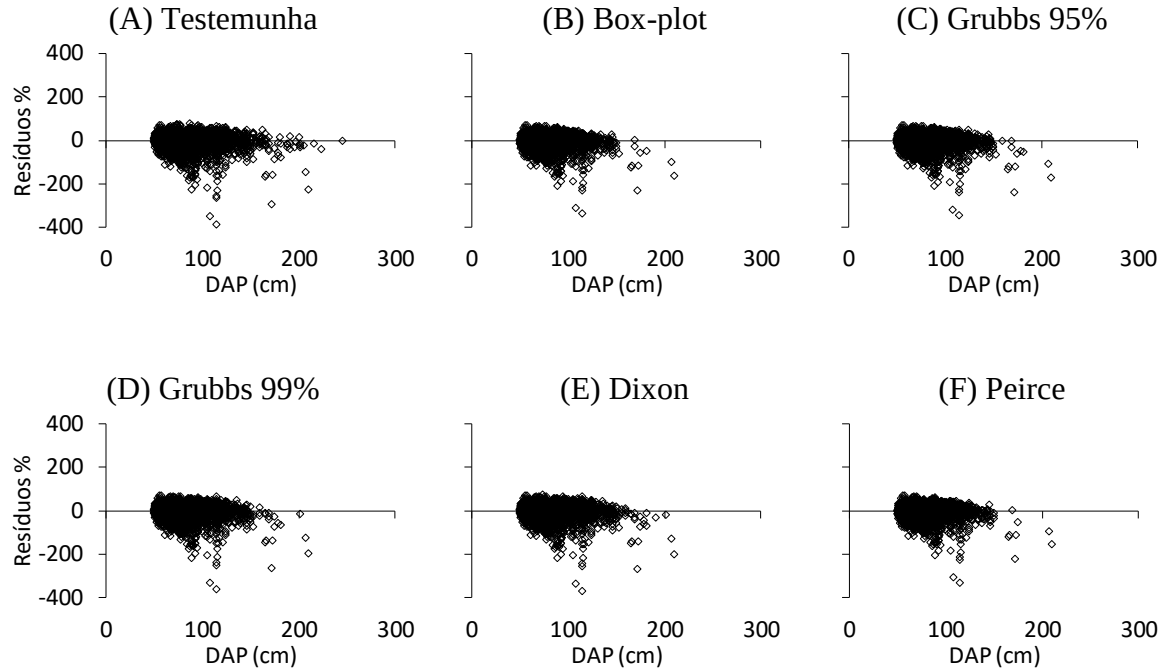

Figura 3: Distribuição dos resíduos para os testes de *outliers* aplicados considerando o *v* como variável de seleção.

DISCUSSÃO

A análise gráfica dos resíduos é um critério na seleção de modelos volumétricos e indica a qualidade de um ajuste permitindo analisar se os resíduos tendem a estar mais ou menos dispersos da reta de regressão. Como os maiores resíduos presentes nas classes intermediárias foram mantidos na base de dados após a aplicação dos testes, e se mantiveram bastante dispersos em relação à reta, não ocorreu alteração na heterocedasticidade dos resíduos. Com isso, os dados das classes diamétricas centrais continuaram bastante dispersos, enquanto os de classes maiores foram quase que totalmente eliminados.

Em outro critério de avaliação da qualidade de ajustes,), houve uma redução do coeficiente de determinação (R^2) na maioria dos testes e variáveis, o que apresenta, consequentemente, redução da qualidade. Para o erro padrão da estimativa (Syx%), a diferença foi insignificante para a variável *h* em todos os testes. Já para as variáveis *d* e *v* o teste de Peirce mostrou-se como o melhor com redução de até 10 % na variável *v.* Porém, esse foi o teste que mais excluiu unidades da amostra nas três variáveis, diminuindo consideravelmente a associação dos dados estimados com os observados, apresentando, assim um menor erro.

A qualidade do ajuste se mostra ruim, mesmo após a aplicação dos quatro testes de *outliers,* quando comparado com outros trabalhos de ajuste de equações de volume desenvolvidos na Amazônia brasileira. Moura (1994) obteve na amplitude máxima um R^2aj. de 0,96 com 68 espécies dentre 710 árvores cubadas e um erro Syx % de 12,08 na amplitude mínima, Rolim et al. (2006) obtiveram um R^2 de 0,99 e um Syx% de 4,64 com uma amostra composta por 55 árvores cubadas.

Ainda que os altos valores de R^2aj. e baixos valores de Syx% estejam relacionados a amostras com poucas árvores cubadas em relação aos deste estudo, devido à sua alta

variabilidade de uma floresta natural, os testes de *outliers* não aprimoraram significativamente a qualidade do ajuste, e tão pouco reduziram satisfatoriamente o erro.

Os métodos utilizam-se de estratégias para detectar e eliminar valores com uma diferença muito grande da normalidade, comparando com valores absolutos. Porém nas florestas tropicais da Amazônia, é natural a ocorrência de algumas árvores com diâmetros muito maiores ao restante da população, não significando que essa medição esteja errada. Nos testes de *outliers,* tanto nas tabelas, como na distribuição gráfica dos resíduos gerados pelo ajuste do modelo, foi possível observar a mesma tendência explicitada na análise dos histogramas, onde foram excluídas principalmente as árvores de maior diâmetro.

Essas metodologias aplicam-se a outros objetivos de estudo, como no caso de tratamento de dados de energia elétrica (DA SILVA, 2011), onde um dado muito acima ou abaixo de uma média possivelmente será um dado discrepante. Para o uso na área de modelagem de volume para florestas naturais, esses métodos não se mostraram eficientes e proporcionam apenas a remoção de um agrupamento de dados na extremidade, mantendo aqueles nas classes intermediárias, e inalterando a heterocedasticidade, a elevada dispersão e variabilidade e a tendência de superestimação.

Este trabalho aplicou quatro testes separadamente em três diferentes variáveis como mostrado anteriormente. É importante aplicar testes multivariados para a detecção de *outiers,* visando avaliar a variabilidade em conjunto das variáveis *d*, *h* e *v,* verificando a possível diferença de resultado, assim como aplicar a metodologia dos quatro testes de *outliers* nos resíduos dos ajustes das equações de volume, a fim de comparar os resultados com os obtidos na aplicação para uma única variável.

CONCLUSÕES

Os quatro testes de *outliers* utilizados neste estudo, proporciona a remoção de dados na extremidade maior ou nas classes de maiores diâmetros, alturas e, consequentemente, volumes. Alguns possibilitam maior exclusão de dados, como no caso do teste de Peirce ao passo que outros geram menor remoção, o caso do Z-Score. Contudo, os testes de detecção de *outliers* aplicados não indicam de forma satisfatória a exclusão dos dados discrepantes para a composição de uma amostra adequada ao ajuste de modelos de volume individual da Floresta Amazônica.

Os elevados erros nas estimativas de volume em florestas tropicais estão relacionados à grande variabilidade de diversos fatores encontrada em estudos com um elevado número de dados em florestas tropicais, naturais e inequiâneas. Observa-se que quanto menor é o número de unidades na amostra, menor é o erro. Os testes de *outliers* fazem isso, reduzem a quantidade de dados em uma amostra, porém não se observa uma redução igualmente proporcional dos erros.

AGRADECIMENTOS

Os autores agradecem a empresa Amata Brasil S.A., por fornecer os dados utilizados neste estudo. Gostaria de agradecer o Vinicius Cysneiros pela ajuda, disponibilidade e por ter contribuído enormemente com informações e dados para este trabalho.

Também gostaria de agradecer o professor Allan Pelissari pela paciência, conselhos, instrução, e compreensão durante o desenvolvimento do trabalho. Sua orientação foi de fundamental importância para a elaboração do presente artigo.

E por fim, gostaria de agradecer minha família pela incessante ajuda, em particular aos meus tios florestais, Mármonn e Maurício Nadolny pelo apoio durante meu período acadêmico.

REFERÊNCIAS

AKINDELE, S. O.; LEMAY, V. M. Development of tree volume equations for common timber species in the tropical rain forest area of Nigeria. **Forest Ecology and Management,** v. 226, p. 41-48, 2006.

BARNETT, V; LEWIS, T. ***Outliers*** **in Statistical Data**. 3 ed. New YorkJohn Wiley & Sons, 1994.

BEDRICK, E. J.; HILL, J. R. Outlier tests for logistic regression: A conditional approach. **Biometrika**, v. 77, p. 815-827, 1990.

COLE, T. G.; EWEL, J. J. Allometric equations for four valuable tropical tree species. **Forest Ecology and Management**, v. 229, p. 351-360, 2006.

COLPINI, C.; TRAVAGIN, D. P.; SOARES, T. S.; SILVA, V. S. M. Determinação do volume, do fator de forma e da porcentagem de casca de árvores individuais em uma Floresta Ombrófila Aberta na região noroeste de Mato Grosso. **Acta Amazonica**, v. 39, n. 1, p. 97-104, 2009.

CYSNEIROS, V. C. **Estratégias para modelagem do volume comercial em florestas tropicais**. 2016. 117 f. Dissertação (Mestrado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2016.

CUNHA, U. S.; MACHADO, S. A.; FIGUEIREDO FILHO, A. Uso de análise exploratória de dados e de regressão robusta na avaliação do crescimento de espécies comerciais de terra firme da Amazônia. **Revista Árvore**, v. 26, n. 4, p. 391 – 402, 2002.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3ª ed. New York: John Wiley & Sons, Inc. 1998. 704 p.

ELLINSON, S. L. R.; BARWICK, V. J.; FARRANT, T. J. D. **Practical Statistics for the**

Analytical Scientist. 2. Ed. A Bench Guide. 2009. 283 p.

GRUBBS, F. E. Procedures for Detecting Outlying Observations in Samples. **Technometrics**, v. 11, n. 1, p. 13-14, 1969.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies**. Artificial Intelligence Review**, v. 22, n. 2, p. 85-126, 2004.

IGBINOSA, A. H.; AMOO, O. B. Appropriate Volume Functions for Leguminosae Family in Two Tropical Rainforests in Cross River State, Nigeria. **Journal of Environment and Ecology**, v. 5, n. 2, p. 206–221. 2014.

NEPOMUCENA, C. M.; CIRILLO, M. A. Estratégias para detectar *outliers* em dados de proporção. **Revista Brasileira de Biometria**, v. 27, n. 4, p. 538-547, 2009.

RIBEIRO, R. B. S.; GAMA, J. R. V.; MELO, L. O. Seccionamento para cubagem e escolha de equações de volume para a Floresta Nacional do Tapajós. **Cerne**, v. 20, n. 4, p. 605-612, 2014.

ROLIM, S. G.; COUTO, H. T. Z.; JESUS, R. M.; FRANÇA, J. T. Modelos volumétricos para a Floresta Nacional do Taipé-Aquirí, Serra dos Carajás. **Acta Amazonica**, v. 36, n. 1, p. 106-114, 2006.

ROSS, S. M. Peirce's Criterion for the Elimination of Suspect Experimental Data. **Journal of Engineering Technology**, v.20, n.2, p. 38-41, 2003.

SABINO, C. V. S.; LAGE, L. V.; ALMEIDA, K.C. B. Uso de métodos estatísticos robustos na análise ambiental. **Engenharia Sanitária e Ambiental**, v. 1, p. 87 – 94, 2014.

SILVA, A. N. **Detecção de *outliers* em séries espaço-temporais: análise da precipitação em Minas Gerais.** 2012. 69 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2012.

THAINES, F.; BRAZ, E. M.; MATTOS, P. P.; THAINES, A. A. R. Equações para a estimativa de volume de madeira para a região da bacia do Rio Ituxi, Lábrea, AM. **Pesquisa Florestal Brasileira**, v. 30, n. 64, p. 283-289, 2010.

TONINI, H.; BORGES, R. A. Equação de volume para espécies comerciais em Floresta Ombrófila Densa no Sul de Roraima. **Pesquisa Florestal Brasileira**, v. 35, n. 82, p. 11-17, 2015.

TRIOLA, M. F. **Introdução à Estatística**. 9 ed. Rio de Janeiro, 2005. 656 p.

DA SILVA, M. D. **Um Modelo Para Análise e Tratamento de Dados de Demanda de Energia Elétrica,** 2011. Dissertação (Bacharelado em Ciência da Computação) – Universidade Federal de Alfenas, Alfenas, 2011.