ESTIMAÇÃO DO VOLUME DE PINUS CARIBAEA VAR. BAHAMENSIS VIA MODELO DE REGRESSÃO RIDGE

[1] Mário Diego Rocha Valente,
[1]Estatístico, Analista de Trânsito do DETRAN/Pará, mario.valente@detran.pa.gov.br

RESUMO

O objetivo deste trabalho foi utilizar um método estatístico de regressão linear múltipla com penalização (Regressão Ridge) para estimar o volume total de madeira da espécie florestal *Pinus caribaea* var. *bahamensis*, em situações com presença de multicolinearidade entre variáveis preditoras. Os dados utilizados foram oriundos de um experimento conduzido pela empresa Duratex Florestal S/A, no Estado de São Paulo, no ano de 2017. Foram analisadas 250 árvores, cujos volumes foram mensurados a partir de sete variáveis independentes. A presença de multicolinearidade foi identificada por meio da matriz de correlação entre as variáveis explicativas e dos fatores de inflação da variância (VIF). Para contornar esse problema, utilizou-se a Regressão Ridge, cuja abordagem considera a estrutura de correlação entre os preditores. O modelo ajustado atendeu às suposições teóricas exigidas, apresentou interpretação prática facilitada, bom ajuste aos dados (R^2 de 95%) e alta capacidade preditiva, demonstrando-se adequado para estimativas de volume de madeira de *Pinus* na região estudada.

Palavras-Chave: Multicolinearidade, Regressão *Ridge*, Volume de Pinus.

ABSTRACT

The objective of this study was to apply a multiple linear regression method to estimate the total wood volume of the forest species *Pinus caribaea* var. *bahamensis* in the presence of multicollinearity. The data were obtained from an experiment conducted by Duratex Florestal S/A in the state of São Paulo, Brazil, in 2017. A total of 250 pine trees were analyzed, and their volumes were quantified based on measurements of seven independent variables. The existence of multicollinearity was confirmed through the correlation matrix of the independent variables and the variance inflation factor (VIF). To address this issue, Ridge Regression was employed, which incorporates the correlation structure among the predictors. The resulting model met all the theoretical assumptions, demonstrated ease of interpretation and application, and provided a good fit to the data (R^2 = 95%) along with strong predictive performance. Therefore, it can be effectively used to estimate pine volume in the study region.

KEYWORDS: Multicollinearity, Ridge Regression, Pine Volume**.**

INTRODUÇÃO

O manejo florestal, seja em ocorrência natural ou plantada, tem como primeiro insumo alta dose de capital financeiro, visto que se obriga adquirir a terra e protegê-la para o seu uso econômico. Para as florestas plantadas ainda precisa-se instituir estudos para poder cultivá-las (levantamento topográfico, calagem e adubação do solo, estudo das microbacias e disponibilidade de água) (BINOTI *et al.*, 2013).

As florestas plantadas são as principais fontes de suprimento de madeira das cadeias produtivas nos segmentos industriais como: a produção de celulose, papel, madeira serrada, chapas e painéis reconstituídas, siderúrgica a carvão vegetal, energia e produtos de madeira sólida (FONSECA, 2009).

De acordo com Silvestre *et al.* (2014), diversos pesquisadores estão buscando alternativas que possibilitem estimativas da produção de povoamento florestais, a partir desta, planeja-se desde o abastecimento da indústria até a qualificação de áreas a serem plantadas, além de possibilitar cálculos de viabilidade econômica que determinam as idades de rotação nas quais se tem maior rentabilidade, o que se relacionam as receitas das empresas florestais.

Para Mendonça *et al.* (2015), a estimativa do volume em povoamentos florestais é de suma importância, pois a partir dela é que será determinada a produção. Os modelos devem ser ajustados de forma a representar as variações dos povoamentos florestais como: espécie que tem efeito significante no volume sólido de madeira empilhada, o sítio, a densidade, a forma da árvore que é influenciada pelo ambiente e características genéticas das espécies e a idade. Assim, para os autores, a identidade de modelos é uma ferramenta que possibilita avaliar se as fontes de variação do povoamento têm influência significativa nas equações geradas para estimar o volume.

Segundo Miranda *et al.* (2016), a medição de todas as árvores de uma floresta, com a finalidade de conhecer seus volumes, é muitas vezes uma tarefa impraticável, dessa forma, é utilizado na maioria das vezes um modelo estatístico para tal fim. Assim, para os autores, a identidades de modelos é uma ferramenta que possibilita avaliar se as fontes de variação do povoamento têm influência significativa nas equações geradas para estimar o volume.

As técnicas estatísticas surgem nesse cenário como importantes fontes de produção de conhecimento, principalmente para estimação do volume comercial, em que o uso de equações de volume e de relações hipsométricas em inventários florestais vem-se constituindo em operação rotineira para cálculo de volume de madeira em pé e estimativa da altura das árvores através da relação diâmetro a altura do peito e altura (VALENTE *et al.*, 2011).

Nesse contexto, faz-se necessário a utilização de modelos estatísticos que possibilitem estimar o volume com base em medições mais simples (CERQUEIRA *et al.*, 2017). A grande motivação de aplicar a Regressão *RIDGE* foi à utilização de todas as variáveis mensuradas, mesmo sendo variáveis altamente correlacionadas. Porém, é muito utilizada, equivocadamente, a regressão linear diretamente nessas variáveis explicativas. Podendo, surgir problemas quando o modelo tiver mais de uma variável ou quando as variáveis não tenham uma relação linear com a variável resposta, logo as estimativas não serão confiáveis.

O objetivo deste estudo visa estabelecer uma metodologia que permita estimar o volume da unidade amostral do Pinus por meio de modelos de regressão na existência da multicolinearidade, conforme noções de (KIBRIA, 2003).

REVISÃO DE LITERATURA

Multicolinearidade

As aplicações nas ciências florestais, freqüentemente encontram-se variáveis independentes que estão correlacionadas entre elas mesmas e, também, com outras variáveis que não estão incluídas no modelo, mas estão relacionadas às variáveis dependentes (NETER *et al.*, 1996).

O fato de muitas funções de regressões diferentes proporcionarem bons ajustes para um mesmo conjunto de dados é porque os coeficientes de regressão atendem a várias amostras em que as variáveis independentes são altamente correlacionadas. Assim, os coeficientes de regressão estimados variam de uma amostra para outra quando as independentes estão altamente correlacionadas. Isso leva à informação imprecisa a respeito dos coeficientes verdadeiros, sendo esse fenômeno chamado de multicolinearidade (NETER *et al.*, 1996).

Segundo Pereira *et al.* (2014), a principal conseqüência da existência de colinearidade entre as variáveis é observada na inferência relacionada com as estimativas dos parâmetros, uma vez que os erros padrão das estimativas são inflacionadas, resultando conseqüentemente, em intervalos de confiança com grandes amplitudes e, naturalmente, menos preciso.

De acordo com HAIR *et al.* (2005), SILVA *et al.* (2009) e VALENTE *et al.* (2011), o indício mais claro da existência da multicolinearidade é quando o coeficiente de explicação (R^2) é bastante alto, mas nenhum dos coeficientes da regressão é estatisticamente significativo segundo o teste t convencional, ou seja, pode indicar sua exclusão do modelo, mesmo que exista uma forte relação linear desta com a variável resposta.

A multicolinearidade pode acarretar sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo (CHAGAS *et al.*, 2009). Assim, uma forma de medi-la é utilizar o Fator de Inflação da Variância ("Variance Inflation Factor"- VIF), definido por $VIF(\hat{\beta}_j) = \frac{1}{(1-R_j^2)}$, sendo $R_j^2 =$ coeficiente de correlação múltipla, resultante da regressão de X_j nos outros p-1 regressores. Quanto mais forte for a dependência linear de X_j nos regressores e, por conseguinte mais forte a colinearidade, maior será o valor de R_j^2. Logo se diz que $V(\hat{\beta}_j) = \sigma^2 C_{ij}$ é "inflacionada" pela quantidade $(1-R_j^2)^{-1}$. Dessa maneira define-se o fator de inflação da variância para $\hat{\beta}_j$ como: $VIF(\hat{\beta}_j) = \frac{1}{(1-R_j^2)}$, $j = 1,2,...,p$. O VIF é igual a 1 (um) quando o valor de $R^2 = 0$, isto é, quando x_i não é colinear com as outras variáveis X. Quando $R^2 \neq 0$, então VIF é maior que 1, indicando uma variância inflacionada para $\hat{\beta}$. Se o valor de R^2 for próximo de 1, existe uma alta correlação entre a variável X_i e as demais variáveis, com isso, o valor de VIF será alto, apontando para o envolvimento dessas covariáveis na multicolinearidade. Nesse

contexto, o VIF representa o incremento da variância devido a presença de colinearidade, obtido para cada variável na diagonal inversa da matriz de correlação (PEREIRA *et al.*, 2014).

A principal questão está em diagnosticar o grau de severidade da multicolinearidade (PEREIRA *et al.*, 2014). Quanto maior for o valor do VIF, mais severa será a multicolinearidade (CHAGAS *et al.*, 2009). Alguns autores como, por exemplo, Khalaf e Shukur (2005), Mardikyan e Cetim (2008), Petrini *et al.* (2009), Dorugade e Kashid (2010) e SILVA *et al.* (2013) sugerem que, se qualquer valor de VIF > 10, a multicolinearidade causará efeitos nos coeficientes de regressão e será um problema. Outros autores consideram esse valor muito formal e sugerem que o VIF não deve exceder quatro ou cinco unidades. (MYERS; MONTGOMERY, 2002) e (MONTGOMERY *et al.*, 2006).

Kibria (2003), Chagas *et al.* (2009) e Pereira *et al.* (2014) ressaltam que, o desempenho do Modelo depende do conhecimento que o pesquisador tenha acerca da relação entre o tamanho amostral e do efeito da multicolinearidade presente nas covariáveis a serem utilizadas no modelo de regressão envolvidas no experimento.

Modelo de Regressão *Ridge*

A regressão *Ridge*, conhecida também como, de Cume, Cumeeira ou em Crista, proposta por Hoerl e Kennard (1970a), é um dos vários métodos propostos para remediar os problemas de multicolinearidade, alterando o método de mínimos quadrados para permitir estimadores viesados dos coeficientes de regressão.

Quando um estimador tem um viés pequeno e é substancialmente mais preciso que o estimador não viesado, este pode ser escolhido desde que tenha grande probabilidade de estar próximo do valor verdadeiro (Hoerl; Kennard 1970b). Assim, a probabilidade do estimador *ridge* β estar próximo do valor verdadeiro α é muito maior que para o estimador não viesado de mínimos quadrados ordinários (NETER *et al.*, 1996).

Uma medida da combinação do efeito do viés e da variação amostral é o valor esperado do quadrado do desvio do estimador $\hat{\beta}$ e do valor verdadeiro β. Esta medida é chamada de Erro Médio Quadrático (EMQ), e pode ser escrita como,

$$E(\hat{\beta}-\beta)^2 = V(\hat{\beta}) + [E(\hat{\beta})-\beta]^2. \quad (1)$$

Dessa maneira, o EMQ é igual à variância do estimador mais o viés ao quadrado. Note que se o estimador for não viesado, o erro médio quadrático é igual ao estimador da variância. Pelo método de mínimos quadrados, o coeficiente β pode ser estimado como,

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad (2)$$

e as estimativas e suas variâncias poderão ser incertas na presença de multicolinearidade. A regressão *ridge* consiste na adição de coeficientes $k \geq 0$ a diagonal principal da matriz de correlações $(X'X)^{-1}$, causando um decréscimo na variância das estimativas. Desta maneira, o estimador *ridge* de β é obtido por,

$$\hat{\beta} = (X'X + K)^{-1}X'Y. \quad (3)$$

sendo $K = diagonal\ (k_1, k_2, ..., k_p)$, $K_i \geq 0$, onde um procedimento bastante usado é $k = kI$, $k \geq 0$. O estimador *ridge* é na verdade uma família de estimadores, onde *k* é um valor pequeno que deve ser escolhido a critério do pesquisador e I a matriz indicadora. Em geral, aumenta-se gradativamente o valor de *k* até que os estimadores dos coeficientes tornam-se estáveis, não variam. Se a escolha for $k_i = 0$, para todo *i*, tem-se o estimador de Mínimos Quadrados (NETER; WASSERMAN, 1974), (DRAPER; SMITH, 1981) e (ELIAN, 1998).

Quando os dados possuem traços de multicolinearidade sempre existe um valor para o parâmetro *k,* no qual os estimadores de Regressão *Ridge* produzem um Quadrado Médio do Erro (QME) menor do que o QME produzido pelos Estimadores de mínimos quadrados ordinários (HOERL; KENNARD, 1970a), (NETER; WASSERMAN, 1974), (DRAPER; SMITH, 1981), (ELIAN, 1998).

A função estimada pela Regressão *Ridge* produz predições com novas observações que tendem a serem mais precisas do que as feitas pela função estimada pelo método de mínimos quadrados ordinários, quando as variáveis independentes são correlacionadas e a nova observação segue o mesmo padrão de multicolinearidade, está precisão na predição de novas observações é favorecida pela Regressão *Ridge*, especialmente quando a multicolinearidade é forte (NETER; WASSERMAN, 1974), (DRAPER; SMITH, 1981), (ELIAN, 1998).

Métodos para Determinação do Valor do Parâmetro K

Um valor ideal para o parâmetro *K*, o qual resulta em um menor QME que o obtido pelo Método de Mínimos Quadrados Ordinários (MQO) depende do vetor de parâmetro β desconhecido e da variância do erro σ^2 também desconhecida (HOERL; KENNARD, 1970a). Conseqüentemente, *K* precisa ser determinado empiricamente ou obtido dos dados, e não é possível determinar o valor ideal do parâmetro *ridge K*. Muitos métodos têm sido propostos para obter os valores apropriados, mas não existe um consenso de qual método é o mais adequado. Assim, o parâmetro de cumeeira *K* será estimado a partir de dois métodos: Gráfico do Traço de Cume (*Ridge Trace Plot*) e Gráfico do Fator de Inflação da Variância (*Variance Inflation Factor Plot*).

Um dos obstáculos principais em utilizar a regressão *ridge* está em escolher um valor de *k*. O traço de cume é um esboço dos valores de (*p*-1) coeficientes estimados de regressão *ridge* padronizados para diferentes valores de *K*, usualmente entre 0 e 1. Feito o traço, pode-se examinar um valor de *K* onde as estimativas se estabilizam. Hoerl e Kennard (1970b) desenvolveram um gráfico bidimensional do valor de cada coeficiente versus k, mostrando como os valores de $\hat{\beta}$ variam em função dos valores de k, ou seja, a partir do gráfico o analista escolhe um valor para *K* que os coeficientes da regressão tendem a ser mais precisos, que o MQO, quando os dados estão sob o efeito da multicolinearidade.

Segundo Chagas *et al.* (2009), o objetivo é escolher um valor de k a partir do qual as estimativas dos parâmetros sejam relativamente estáveis, gerando uma série de coeficientes com menor soma dos quadrados do resíduo do que a solução clássica. Assim, na medida em que se aumenta o valor de k, a soma de quadrados dos resíduos também aumentará, sugerindo iniciar com valores pequenos de k e ir aumentando gradativamente até que os coeficientes se estabilizem.

Para Hoerl e Kennard (1970a), o Fator de Inflação da Variância, mostra a variabilidade em função do valor de *K*, ou seja, à medida que se atribui valores para *K* que estabilizam os coeficientes de regressão *ridge*, a variabilidade diminui, removendo a multicolinearidade. O traço do cume pode também ser usado para sugerir variáveis(s) que podem ser retiradas do modelo. Algumas variáveis cuja estimativa do parâmetro é instável a cada mudança do valor de *K* ou que decresce para zero são candidatos para anulação.

Atualmente na literatura, existem várias medidas corretivas para suavizar os efeitos provocados pela multicolinearidade, outros métodos são propostos desde simples às mais complexas, tais como, Ampliação do Tamanho da Amostra e Remoção das Variáveis (HAIR *et al.*, 2005), utilização de Modelo de Regressão por Componente Principais (SILVA *et al.*, 2009), Modelo de Regressão por Análise Fatorial (VALENTE *et al.*, 2011), Regressão com Variáveis Latentes (MALHOTRA, 2011) e Regressão via Redes Neurais (RODRIGUES *et al.*, 2010) e (LEAL *et al.*, 2015).

MATERIAIS E MÉTODOS

Localização e Área de Estudo

Os dados foram provenientes de um experimento conduzido pela Empresa Duratex Florestal S/A localizada na cidade de São Paulo, SP em 2022. A área de estudo consiste em plantio de *Pinus caribaea var. bahamensis*, cujas idades variam de 9,5 a 13,2 anos, caracterizando talhões maduros, com espaçamento empregado de 4x3m.

Levantamento dos Dados

Para a coleta das variáveis dendrométricas do povoamento, realizou-se inventário florestal por meio do método da área fixa com processo de amostragem aleatória simples. Para realização da cubagem empregou-se o método direto de mensuração do volume pela fórmula de Smalian 250 árvores-amostra, onde foram medidos os diâmetros ao longo do tronco com circunferências das secções a cada 2 metros (ALMEIDA *et al.*, 2016). A variável dependente foi o Volume real (m³), e as variáveis independentes consideradas inicialmente e, explicativas do volume, foram DAP (Diâmetro à Altura do Peito) coletado com fita métrica, ou seja, mede-se, na verdade a Circunferência a Altura do Peito (CAP), a 1.30 m do solo, para posteriormente, ser convertida em DAP, a Idade em anos, a Altura comercial foi considerada da base do corte (0.15 cm do solo) até o ponto onde houve a remoção das copas das árvores derrubadas em metros, o Índice da Área Foliar (IAF), Diâmetro Angular da Folha (DAF) e a Medida da Clareira (GAP) em cm, e Área Basal das árvores.

Análise Estatística

Para identificação de multicolinearidade nas estimativas dos coeficientes de regressão, considerou-se, como regra de decisão, o VIF para os estimadores superior a cinco, com base na afirmativa de Montgomery *et al.* (2006). Para estimativa de $VIF(\hat{\beta}_j)$. Foi utilizado também o gráfico do VIF proposto por Hoerl e Kennard (1970a, b). Foi aplicado o Teste de White para verificar a presença de heterocedasticidade dos erros de regressão e o Teste de Durbin-Watson, para

avaliar a autocorrelação de primeira ordem nos resíduos segundo as noções de (SILVA *et al.*, 2011); (SILVA e SANTANA, 2014).

A regressão *ridge* foi obtida por meio do método de seleção das variáveis do tipo *Stepwise Standard*, em que as variáveis foram introduzidas uma a uma e verificando o seu grau de contribuição para o modelo. Os parâmetros da equação de regressão foram estimados a partir das inter-relações da variável dependente (Volume) e das variáveis independentes (DAP, IDADE, ALTURA, IAF, DAF, GAP e ÁREA BASAL). As estimativas foram obtidas por meio do Software Estatístico chamado *Statistical Package for Social Science* - SPSS versão 25, sendo utilizado método de mínimos quadrados que consistem no procedimento matemático para minimizar os erros quadráticos para calcular as estimativas dos parâmetros da regressão.

RESULTADOS E DISCUSSÃO

Inicialmente, observou-se a ordem de grandeza das variáveis, visando detectar discrepâncias que pudessem causar problemas na análise, avaliou-se a viabilidade da análise de Regressão Linear a partir da matriz de correlações. O primeiro passo é um exame visual das correlações, identificando as que são estatisticamente significantes, com isso, verificou-se que, existe um número substancial de correlações maiores que 0,30 (GORSUCH, 1983 e HAIR *et al.*, 2009), sugerindo possíveis relações entre as variáveis.

Tabela 1. Matriz de Correlação referente aos Dados de PINUS da Empresa Duratex Florestal em 2022.

Variáveis	Volume	IAF	DAF	GAP	IDADE	DAP	Altura	Área Basal
VOLUME	1							
IAF	0, 088	1						
DAF	0, 014	-0, 233	1					
GAP	-0, 068	**-0, 946**	0, 440	1				
IDADE	**0, 708**	0, 012	0, 198	0, 063	1			
DAP	**0, 744**	0, 068	0, 132	-0, 036	0, 865	1		
ALTURA	**0, 894**	0, 030	0, 093	-0, 005	0, 801	0, 858	1	
ÁREA BASAL	**0, 738**	0, 470	0, 052	-0, 415	0, 305	0, 525	0, 518	1

De acordo a Tabela 1 verifica-se que, as correlações das variáveis independentes (Idade, DAP, Altura e Área Basal) com a variável dependente (Volume) são moderadas acima de 0,70 e, a variável mais correlacionada com o Volume de madeira é a Altura (r = 0,89). Observaram-se ainda uma alta correlação inversa entre as variáveis independentes IAF e GAP (r = -0,94), esses resultados se complementam, pois IAF e GAP são variáveis radiométricas obtidas por meio de um instrumento e são por natureza inversamente proporcional.

Um modelo de regressão linear múltipla foi ajustado para o Volume de *Pinus,* utilizando o procedimento passo a passo, ou seja, à medida que se introduziu uma variável por vez no modelo, a natureza de cada uma delas também mudava. Por exemplo, quando se colocou a variável DAP no modelo, esta mudou o sinal das

outras variáveis, indícios de multicolinearidade nos dados (HAIR *et al.*, 2009). Porém, para confirmar a presença dos efeitos será aplicada a medida VIF.

Contudo, de acordo com a Tabela 2, observa-se a presença de multicolinearidade nos dados, ou seja, verifica-se que, os valores do VIF, em sua maioria são maiores que 10, indicando uma forte existência de colineraridade entre as variáveis independentes.

Tabela 2. Ajuste inicial do Modelo de Regressão Linear Múltiplo para determinação da equação de Volume de Pinus da Empresa Duratex Florestal em 2022.

Variáveis	Coeficientes	Teste t	P-valor	VIF
Constante	-94, 611	-0, 703	0, 488	-
Índice de Área Foliar (IAF)	3, 251	0, 109	0, 914	**18, 102**
Distribuição Angular da Folha (DAF)	-1, 249	-0, 848	0, 404	**22, 125**
Medida de Clareira (GAP)	75, 823	0, 185	0, 855	**20, 175**
Idade	7, 819	0, 838	0, 409	6, 008
Diâmetro Altura do Peito (DAP)	-5, 816	-1, 058	0, 299	**17, 21**
Altura	14, 115	5, 119	**0, 000**	4, 604
Área Basal	2, 053	1, 305	0, 203	2, 551

Ajuste do Modelo de Regressão *Ridge*

Realizou-se a aplicação de um modelo de regressão específico para dados sob o efeito e influência da multicolinearidade chamado Regressão *Ridge* com as variáveis originais, para obter estimadores mais confiáveis e, assim, contornar a multicolinearidade (GAZOLA, 2002).

Com base nos dados dendrométricos analisa-se o volume das árvores em função de suas características, com isso, inicialmente construiu-se um gráfico chamado Traço de Cume (Ridge Trace Plot), na qual se utiliza os coeficientes de regressão em função do valor de um pequeno viés *K*, onde se tem uma noção de qual será o melhor valor para estabilizar os estimadores e ter uma precisão maior para o modelo de regressão estimado.

Assim, de acordo com a Figura 1, verifica-se que, os estimadores começam a estabilizarem usando inicialmente um K = 0,1; a variável altura das árvores é uma possível candidata a ser retirada do modelo, pois apresenta um distanciamento significativo das outras variáveis.

De acordo com os dados anteriormente mencionados, construiu-se um gráfico denominado VIF conforme a Figura 2, onde se utiliza o coeficiente de correlação múltipla de regressão linear de X_i em relação às variáveis explanatórias restantes, na qual se utiliza os diversos valores de VIF em função de pequenos valores de k, com isso, utilizando um k = 0,1 os estimadores começam a estabilizar os coeficientes no modelo de regressão *ridge*, conjuntamente, à medida que se usa um k= 0,1 os VIF ficam próximos ou iguais a um, tendo um coeficiente de correlação múltiplo igual à zero, indicando que cada variável x_i não é colinear com as demais variáveis x_i, diminuindo também a variabilidade dos estimadores, aumentando a precisão e removendo a multicolinearidade.

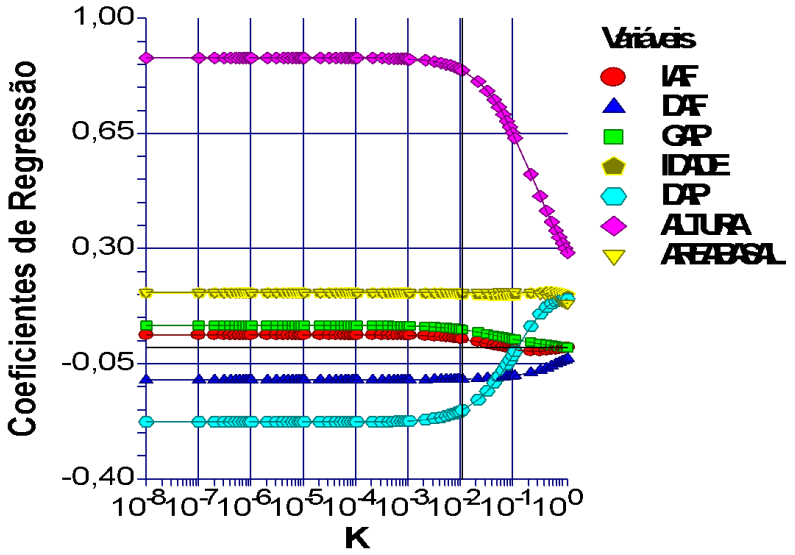

Figura 1. Traço de Cume do Volume de Pinus da Empresa Duratex Florestal em 2022.

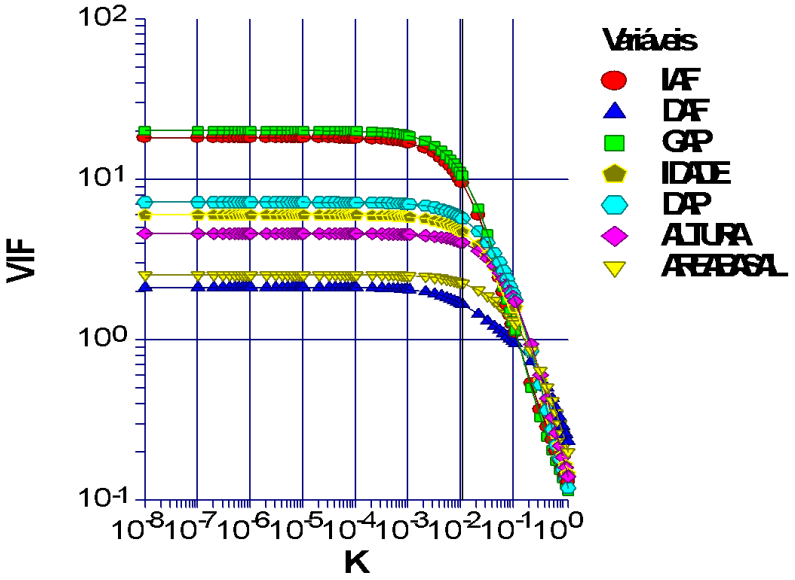

Figura 2. Estimativas dos Fatores de Inflação da Variância em função dos valores de k, no estudo do volume de Pinus da Empresa Duratex Florestal em 2022.

A análise de regressão foi realizada usando o conjunto de dados (sete variáveis independentes, sendo apenas quatro significativas) com as variáveis quantitativas. Foram rodadas inúmeras simulações, com o método *Ridge Regression Stepwise Standard* (GAZOLA, 2002), para identificar o melhor conjunto de variáveis que farão parte do modelo. O valor de k = 0,1 estabilizou os coeficientes, o R^2 ajustado e o quadrado médio do erro, isto é, tornou os valores das estimativas dos coeficientes constantes e manteve as variáveis significativas. Para Cruz *et al.* (2014), quanto maior for o valor de k, mais enviesadas são as informações obtidas pelo modelo estimado.

Ao ajustar-se um modelo de regressão linear múltipla, verificou-se que os mesmos foram significativos devido ao valor do nível descritivo (*p*) ser menor que o nível de significância de 0,05 como mostra a Tabela 3.

Tabela 3. Ajuste Inicial do Modelo de Regressão *Ridge* para Determinação da Equação de Volume de Pinus da Empresa Duratex Florestal em 2022.

Variáveis	Coeficientes	Teste t	P-Valor	VIF
Constante	- 96, 031	0, 487	0, 001	-
Idade	7, 693	0, 409	0, 012	3, 721
DAP	-4, 887	0, 299	0, 017	3, 740
Altura	13, 513	0,105	0, 009	4, 049
Área Basal	2, 055	0,202	0, 002	2, 258

Após analisar a significância individual de cada fator de acordo com o teste *t* partiu-se para analisar a significância geral do modelo de regressão, aplicando-se a Análise de Variância (ANOVA) aos dados das árvores. A análise de variância mostra que se rejeita a hipótese de não haver regressão, isto é, o modelo é significante a um nível de 1%, e conclui-se que pelo menos uma das variáveis explanatórias está relacionada com o valor do Volume, conforme a Tabela 4.

Tabela 4. Análise de Variância para a Significância da Equação de Volume de Pinus da Empresa Duratex Florestal, em 2022.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Teste F	P-valor
Regressão	7	115.554,1	16507,73	17, 277	0, 0000
Resíduo	28	26.752,26	955, 4378		
Total	35	142.306,4			

Analisando-se as medidas de ajuste, observa-se que, os valores do coeficiente de correlação múltiplo (*R* múltiplo), de explicação (R^2) e explicação ajustado (R^2 ajustado), são consideravelmente altos, indicando alta correlação da variável dependente com as variáveis independentes, e alta explicação da variável dependente pelas variáveis independentes, indicando que o modelo ajustado explica bem a variabilidade do valor do Volume das árvores. (SANQUETTA, *et al.*, 2016).

Tabela 5. Estatísticas Referentes ao Ajuste da Equação de Volume de Madeira de Pinus de dados dendrométricos, da Empresa Duratex Florestal, em 2022.

Medidas de Ajustes	Coeficientes
R múltiplo	0, 951
R^2	0, 965
R^2 ajustado	0, 952

Em suma, o coeficiente de determinação múltiplo, que representa a proporção da variação em *Y* (*Volume*) que é explicada a partir do conjunto de variáveis explanatórias selecionadas, apresentou um valor igual a 95% da variação no volume, pode ser explicado a partir da variação nas variáveis e 5% do volume são explicados por outras variáveis que não constam no modelo, de acordo com a Tabela 5. Portanto, a equação de regressão linear múltipla para o volume de Pinus que descreve o relacionamento entre o valor do volume e quatro variáveis independentes é:

VOLUME = - 96.03131 +7.693503 * IDADE - 4.887759 * DAP + 13.51347 * ALTURA + 2.055151 * ÁREA BASAL

Assim, a equação estimada para o volume de Pinus em função de suas características pode ser usada para predizer o volume de outras árvores sem o problema da multicolinearidade.

As suposições do modelo com a equação de regressão ajustada podem ser verificadas a partir da análise de resíduos (SANQUETTA *et al.*, 2017), a fim de procurar evidências sobre violações das suposições de homocedasticidade, independência e normalidade referentes aos dados dendrométricos da empresa Duratex Florestal, em 2022. A suposição de normalidade pode ser verificada a partir da figura 3 construída pelos resíduos padronizados *versus* os respectivos valores teóricos da distribuição normal, que mostra a normalidade dos erros quando os pontos se distribuem em torno de uma linha reta**.**

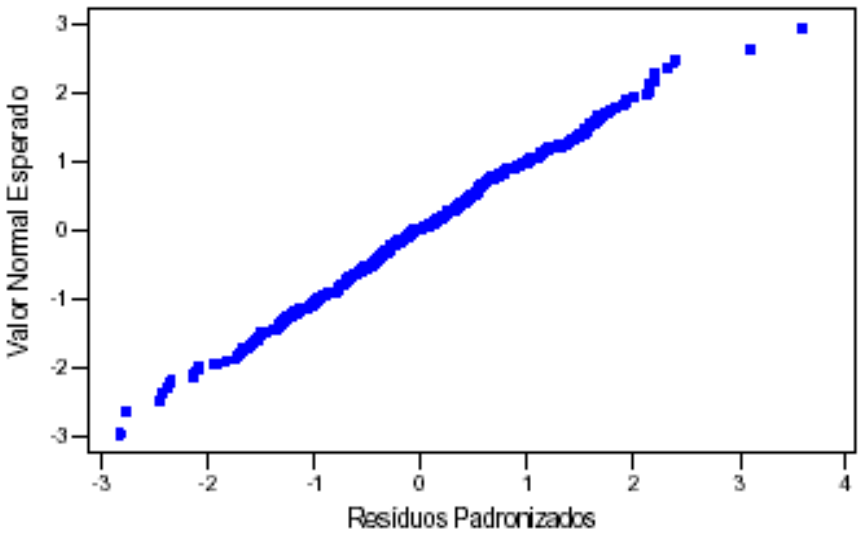

Figura 3. Resíduos padronizados *versus* valores esperados da normal para o volume de pinus da Empresa Duratex Florestal, em 2022.

O teste para Homocedasticidade de White foi realizado nos modelo de regressão *ridge*, cujos resultados indicam a rejeição da hipótese de presença de heterocedasticidade erros, o que indica que a variância dos erros e constante, para todas as observações (SCHNEIDER et al, 2009), (SILVA e SANTANA, 2014).

A suposição de variância constante ou homogeneidade de variância é verificada facilmente a partir da Figura 4, que apresenta os pontos distribuídos aleatoriamente em torno de uma reta horizontal que passa pela origem, sem qualquer padrão, Esta disposição dos pontos indica que a suposição de variância constante é razoável.

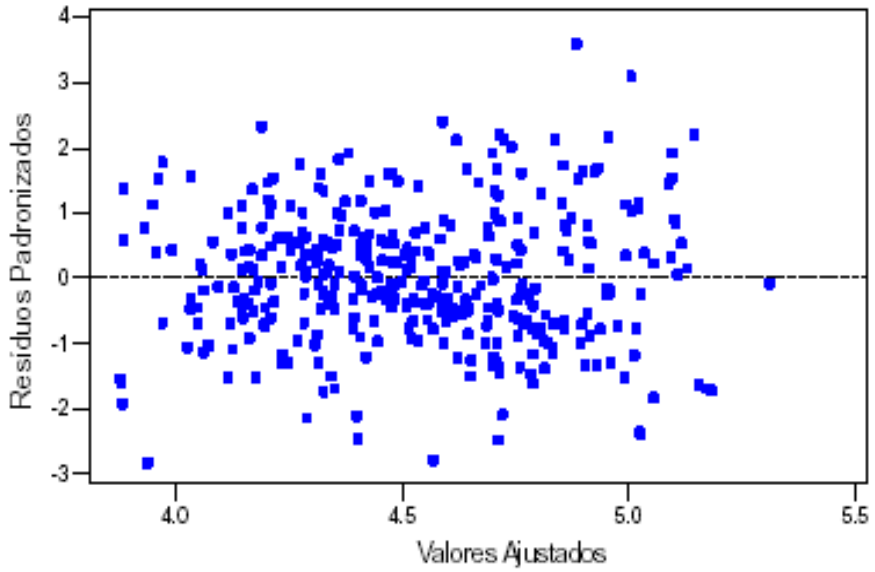

Figura 4. Resíduos ajustados *versus* resíduos padronizados para o volume de pinus da Empresa Duratex Florestal, em 2022.

O teste de Autocorrelação de Durbin-Watson foi realizado no modelo de regressão *ridge*, indicando que os resíduos não são autocorrelacionados contemporaneamente, portanto, as estimativas são não viesadas e eficientes (SCHNEIDER et al., 2009), (SILVA e SANTANA, 2014).

A existência de autocorrelação dos erros pode ser investigada a partir do gráfico dos resíduos versus sequência no tempo (ou sequência de coleta de dados), pode-se verificar a independência dos resíduos quando se distribuem aleatoriamente, em torno de zero, conforme a Figura 5.

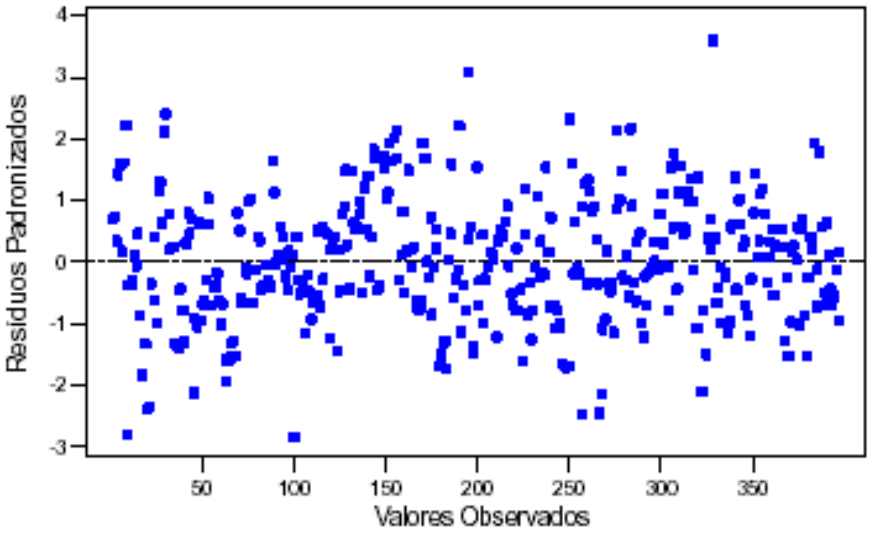

Figura 5. Valores Observados versus resíduos padronizados para o volume de pinus da Empresa Duratex Florestal, em 2022.

Validação do Modelo Ajustado

Utilizando os dados observados em 14 pontos de amostragem, selecionados aleatoriamente em fragmentos da Empresa Duratex Florestal no Município de São Paulo, em 2022, e que não foram utilizados nos ajustes do modelo avaliado, procedeu-se à validação do modelo selecionado para estimar o volume total de pinus. Verificou-se que, o modelo de regressão *ridge* aplicado aos dados após a identificação da multicolinearidade pelo VIF, obteve grande poder de explicação e predição, pois os volumes reais observados na amostra ficaram próximos do volume estimado, sendo o modelo proposto o mais adequado para estimar os volumes totais das árvores da espécie do tipo Pinus *Caribaea var. Bahamensi*.

CONSIDERAÇÕES FINAIS

O modelo escolhido não apresentou problemas de heterocedasticidade e as estimativas dos parâmetros foram todas significantes a 1%. O problema de multicolinearidade entre as variáveis originais foi solucionado pela metodologia proposta. O modelo de regressão *ridge* obtido pode ser utilizado para estimar o volume de *pinus caribaea var. bahamensis* na mesma região em estudo.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, D. L. C. S.; SILVA, F. R., SANTOS, A. F. A.; GARCIA, M. L.; WOJCIECHOWSKI, J. C. Determinação de equação volumétrica e hipsométrica para um plantio de tectona grandis L. f. em Alta Floresta, MT**. Ciências Agroambientais**, Alta Floresta, v. 14, n. 2, p. 1-9, 2016. Disponível em: https://periodicos.unemat.br/index.php/rcaa/article/view/1266/1526.

CHAGAS, E. N.; MENEZES, C. C.; CIRILLO, M. A.; BORGES, S. V. Método "Ridge" em Modelo de superfície de resposta: Otimização de condições experimentais na elaboração de doce de goiaba. **Revista Brasileira de Biometria**, São Paulo, v. 26, n. 4, p. 71-81, 2009. Disponível em: http://jaguar.fcav.unesp.br/RME/fasciculos/v27/v27_n1/A5_Elcio.pdf.

CRUZ, C. D.; CARNEIRO, P. C. S.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético. 3. Ed. revisada e ampliada. Viçosa, Ed. UFV, v.2, 668p, 2014.

CERQUEIRA, C. L.; LISBOA, G. S.; FRANÇA, L. C. J.; MÔRA, R.; MARQUES, G. M.; SALLES, T. T.; BRIANEZI, D. Modelagem da altura e volume de Tectona grandis L. F. na mesorregião Nordeste do Pará. **Nativa**, Sinop, v. 5, esp., p. 606-611, 2017. Disponívelem:http://periodicoscientificos.ufmt.br/ojs/index.php/nativa/article/view/5037.

DRAPER, N. R.; SMITH, H. Applied Regression Analysis. New York: John Wiley & Sons, 1981.

DORUGADE, A. V.; KASHID, D. N. Alternative Method for Choosing Ridge Parameter for Regression. **Applied Mathematical Sciences**, v. 4, n. 9, p. 447-456, 2010. Disponível em: http://www.m-hikari.com/ams/ams-2010/ams-9-12-2010/dorugadeAMS9-12-2010.pdf

ELIAN, Silva. N. Análise de Regressão. São Paulo: IME, 1998.

FONSECA, F. H. Agenda Estratégica do Setor de Florestas Plantadas. Câmara Setorial de Sivilvicultura, Brasília, 36p, 2009.

GAZOLA, Sebastião. Construção de um Modelo de Regressão para Avaliação de Imóveis, Dissertação de Mestrado, UFSC - Florianópolis: 2002.

GORSUCH, R. L. Factor Analysis. New Jersey: Lawrence Erlbaum, 1983.

HOERL, A. E., KENNARD, R. W. Ridge Regression: Applications to nonorthogonal problems. Disponível em: **Technometrics**, v. 12, n. 1, p. 69-82, 1970a. https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488635#.XL3sQTBKiM8.

HOERL, A. E., KENNARD, R. W. Ridge Regression: Biased estimation for nonorthogonal problems. **Technometrics**, v. 12, n. 1, p. 55-68, 1970b. Disponível em:https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634#.XL3sfTBKiM8.

HOERL, A. E.., KENNARD, R. W.; BALDWIN, K. F. Ridge Regression: some simulation. **Communications in Statistics-Theory and Methods**, v. 4, n. 2, p.

105-124, 1975. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/03610927508827232.

HAIR JR, J. F. *et al.* Análise multivariada de dados. 5. ed. Porto Alegre: Bookman, 2005.

KIBRIA, B M Golam. Performance of some new ridge regression estimators. **Communications in Statistics-Simulation and Computation**, v. 32, n. 2, p. 419-435, 2003. Disponível em: https://www.tandfonline.com/doi/abs/10.1081/SAC-120017499.

KHALAF, G.; SHUKUR, G. Choosing Ridge Parameter for Regression Problem. **Communications in Statistics-Theory and Methods**, v. 34, n. 3, p. 1177-1182, 2005. Disponível em: https://www.tandfonline.com/doi/abs/10.1081/STA-200056836

MALHOTRA, N. K. Pesquisa de marketing: uma orientação aplicada. 3ª Ed. Porto Alegre: Bookman, 2001. 719p.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to linear regression analysis. 4. Ed. New York: John Wiley & Sons, 2006. 612p.

MARDIKYAN, Sona.; CETIN, Eyup. Efficient Choice of Biasing Constant for Ridge Regression. **International Journal Comtemporaneon Math Sciences**, v. 3, n. 11, p. 527-536, 2008. Disponível em: https://pdfs.semanticscholar.org/784a/570d3dd8c219313d5374fb48b2dcb9d7bb94.pdf.

MENDONÇA, A. R.; PACHECO, G. R.; VIEIRA, G. C.; ARAUJO, M. S.; INTERAMNENSE, M. T. Identidades de modelos para estimativa do volume de pinus. **Nativa**, Sinop, v. 03, n. 04, p. 281-286, 2015. Disponível em: http://dx.doi.org/10.14583/2318-7670.v03n04a10.

MIRANDA, D. L. C.; ANGELIN, T. B.; LISBOA, G. S.; SILVA, F.; GOUVEIA, D. M.; CONDÉ, T. B.; SILVA, C. S. Modelos estatísticos para estimativas de árvores de Parkia *gigantocarpa* Ducke, em plantios experimentais em Mato Grosso. **Nativa**, Sinop, v.04, n.01, p.1-6, 2016. Disponível em: http://dx.doi.org/10.14583/2318-7670.v04n01a01.

NETER, J.; WASSERMAN, W. Applied linear statistical models. Illinois: Richard D. Irwin, 1974.

NETER, J.; WASSERMAN, W.; KUTNER, M. H.; NACHTSHELM, C. J. Applied Linear Regression Models. 3ª ed., Times Mirror Hiher Group, Inc., Boston, 1996.

PEREIRA, Gislene Araujo; MILANI, Letícia Lima; CIRILLO, Marcelo Ângelo. Uso de alguns estimadores ridge na análise estatística de experimentos em entomologia. **Revista Ceres**, v. 61, n. 3, 2014. Disponível em: http://www.scielo.br/scielo.php?pid=S0034737X2014000300006&script=sci_abstract&tlng=pt.

RODRIGUES, E. F.; OLIVEIRA, T. R.; MADRUGA, M. R.; SILVEIRA, A. M. Um Método para determinar o volume comercial do Schizolobium amazonicum (Huber) Ducke utilizando redes neurais artificiais. **Revista Brasileira de Biometria**, v. 28, n. 1, p. 16-23, 2010. Disponível em: http://jaguar.fcav.unesp.br/RME/fasciculos/v28/v28_n1/A2_Eraldo.pdf

SANQUETTA, C.R; SANQUETTA, M.N.E; BASTOS, A; QUEIROZ, A; CORTE, A.P.D. Estimação de volumes de Araucaria angustifolia (Bertol.) O. Kuntze por fatores de forma em classes diamétricas e modelos de regressão. **Enciclopédia Biosfera**, v. 13 n. 23 p. 588-597, 2016. Disponível em: http://www.conhecer.org.br/enciclop/2016a/agrarias/estimacao.pdf.

SANQUETTA, C.R; DOLCI, M; CORTE, A.P. D; SANQUETTA, M.N.I; PELLISSARI, A.L. Estimativa da altura e do volume em povoamento jovens de restauração em Rondônia. **BIOFIX Scientific Journal**, v. 2 n. 2 p. 23-31, 2017. Disponível em: http://dx.doi.org/10.5380/biofix.v2i2.54124

SCHNEIDER, P.R; SCHNEIDER P.S.P; SOUZA, C.A.M. Análise de regressão aplicada à Engenharia Florestal. 3. ed. Santa Maria: UFSM, CEPEF; 2009.

SILVA, A. V. L.; *et al.* Alternativa de modelo linear para estimação da biomassa verde de Bambusa vulgaris Schrad. ex JC Wendl na existência de multicolinearidade**.** **Ciência Florestal**, v. 19, n. 2, p. 207-214, 2009. Disponível em: http://www.scielo.br/pdf/cflo/v19n2/1980-5098-cflo-19-02-00207.pdf.

SILVA, E. N.; SANTANA, A.C.; QUEIROZ, W.T; SOUSA, R.J. Estimação de equações volumétricas para árvores de valor comercial em Paragominas, Estado do Pará. **Amazônia Ciência e Desenvolvimento**, volume, 07, n.13, julho/dezembro, 2011. Disponível em: http://www3.bancoamazonia.com.br/images/arquivos/institucional/biblioteca/revista_amazonia/edicao13/n13_estimacao_de_equacoes.pdf.

SILVA, E. N.; SATANA, A. C. Modelo de regressão para estimação do volume de árvores comerciais, e, florestas de Paragominas. **Revista Ceres**, Viçosa, v. 61, n. 5, p. 631-636, 2014. Disponível em: http://www.ceres.ufv.br/ojs/index.php/ceres/article/view/4153.

SILVESTRE, R.; BONAZZA, M.; STANG, M.; LIMA, G. C. P.; KOEPSEL, D. A.; MARCO, F. T.; CIANOSCHI, L. D.; SCRIOT, R.; MORÊS, D. F. Equações volumétricas em povoamentos de Pinus taeda L. no município de Lajes–SC. **Nativa**, Sinop, v.02, n.01, p.1-5, 2014. Disponível em: http://periodicoscientificos.ufmt.br/ojs/index.php/nativa/article/view/1402 http://dx.doi.org/10.14583/2318-7670.v02n01a01**.**

VALENTE, M. D. R.; PINHEIRO, J. G.; QUEIROZ, W. T. Modelo de predição para o volume total de Quaruba (Vochysia inundata ducke) via análise de fatores e regressão. **Revista Árvore**, Viçosa, v. 35, n. 2, p. 307-317, 2011. Disponível em: http://www.scielo.br/pdf/rarv/v35n2/a15v35n2.pdf.