ADRIANO VICTOR LOPES DA SILVA

ALTERNATIVAS E COMPARAÇÕES DE MODELOS LINEARES PARA ESTIMAÇÃO DA BIOMASSA VERDE DE *Bambusa vulgaris* NA EXISTÊNCIA DE MULTICOLINEARIDADE

RECIFE-PE – Fevereiro /2008.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ALTERNATIVAS E COMPARAÇÕES DE MODELOS LINEARES PARA ESTIMAÇÃO DA BIOMASSA VERDE DE *Bambusa vulgaris* NA EXISTÊNCIA DE MULTICOLINEARIDADE

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração:

Modelagem Estatística e Computacional (com ênfase nas áreas agrárias, biológicas e humanas).

Orientador: Prof. Dr. Rinaldo Luiz Caraciolo Ferreira

Co-orientador: Prof. Phd. José Antônio Aleixo da Silva

RECIFE-PE – Fevereiro/2008

FICHA CATALOGRÁFICA

S586p Silva, Adriano Victor Lopes da
Alternativas e comparações de modelos lineares para estimação da biomassa verde de *Bambusa vulgaris* na existência de multico - linearidade / Adriano Victor Lopes da Silva. -- 2008.
55 f. : il.

Orientador : Rinaldo Luiz Caraciolo Ferreira
Dissertação (Mestrado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco. Departamento de Es – tatística e Informática.
Inclui bibliografia.

CDD 574. 018 2

1. Análise multivariada
2. Regressão
3. Bambu
I. Ferreira, Rinaldo Luiz Caraciolo
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ALTERNATIVAS E COMPARAÇÕES DE MODELOS LINEARES PARA ESTIMAÇÃO DA BIOMASSA VERDE DE *Bambusa vulgaris* NA EXISTÊNCIA DE MULTICOLINEARIDADE

ADRIANO VICTOR LOPES DA SILVA

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 26/02/2008 pela Comissão Examinadora.

Orientador:

Prof. Dr. Rinaldo Luiz Caraciolo Ferreira
UFRPE

Banca Examinadora:

Prof(a). Dr(a). Borko D. Stosic
UFRPE

Prof(a). Dr(a). Manoel Raimundo de Sena Junior
UFRPE

Prof(a). Dr(a). Tatijana Stosic
UFRPE

Dedicatória

Dedico este trabalho aos meus pais:
Janildo Lopes da Silva e Maria de Lourdes Lopes da Silva.

Agradecimentos

A Deus, pelo dom da vida;

Aos professores do programa de pós-graduação em biometria;

A meus pais, irmã, avó e tios pelo estímulo e apoio incondicional;

À UFRPE, pela valiosa acolhida no treinamento em nível de pós-graduação;

A CAPES, que fez possível a conclusão deste trabalho;

Ao meu orientador, pelo apoio e esclarecimento;

Ao Prof. Dr. José Antônio Aleixo, pela valiosa e acertada orientação acadêmica;

À Agrimex na pessoa do engenheiro florestal Hugo Gutiérrez, que ofereceu seus plantios artificiais de bambu na tomada dos dados necessários para a execução deste trabalho.

Aos amigos de curso, em especial: Rosangela Silveira, Moacy Cabral, Luiz Henrique Gama, Vanessa Santos e Juliana Silva, pelo companheirismo e solidariedade.

Aos amigos Abraão David, Adriana Rachel, Bárbara Magdala, Carlos Heitor, Eduarda Gabrielle, Gustavo Arruda, Hemilio Ferdandes, Ícaro Jônatas, Juliette Noadya, Julio Fujimaki, Leandro Tavares, Marcella Florêncio, Marcilene Nunes, Marcos Brasil, Marcos Feitoza, Rafaella Paiva, Raul Príncipe, Robson Florêncio, Rodrigo Simeão, Samira Nunes, Taciano Rocha, Thaise Gomes, Thiago Barreto, pela amizade, consideração e incentivo na obtenção do sucesso.

Ao funcionário Marco Antônio, secretário do curso, e à querida amiga Zuleide, pelos préstimos e atenção dispensados.

Resumo

O objetivo deste trabalho foi utilizar métodos estatísticos univariado e multivariado na seleção de variáveis independentes, em modelos matemáticos lineares para a estimativa da biomassa verde da haste principal do bambu, *Bambusa vulgaris,* visando reduzir tempo e custo sem perda de precisão, além de empregar alternativas para estimação na existência de multicolinearidade. Os dados foram provenientes de um experimento conduzido pela empresa Agroindustrial Excelsior S. A. (Agrimex) localizada no Engenho Itapirema na cidade de Goiana – PE. Foram utilizadas 450 hastes de bambu, que tiveram sua biomassa verde quantificada através do peso e 4 variáveis independentes medidas na mesma haste. Inicialmente, verificou-se a existência da multicolinearidade por meio da matriz de correlação das variáveis independentes e pelo fator de inflação da variância. Para seleção das variáveis independentes foram utilizados os métodos: Stepwise e Retenção por K componentes. As alternativas utilizadas foram a Regressão com os componentes principais e Regressão Ridge. No geral, em apenas uma situação os métodos de seleção de variáveis se comportam adequadamente na existência de multicolinearidade entre as variáveis explicativas, exatamente o método multivariado de retenção por K=3 componente pela matriz de covariância, modelo de Spurr. Os métodos alternativos de estimação conduzem respostas semelhantes, mesmo que possuindo estruturas diferentes, no entanto, a regressão com os componentes principais obteve os melhores resultados.

Palavras-chave:. Bambu, Análise Multivariada, Regressão.

Abstract

The objective of this work was to use univariate and multivariate statistical methods on selection of independent variables, in the mathematical linear models, to estimate the green biomass of the main bamboo rod, bambusa vulgaris, pursuing time and cost reduction without loss of precision. The data came from an experiment carried out for the Agroindustrial Excelsior S. A. (Agrimex) company located in the city of Goiana – PE.Quantified by its green biomass weight,450 bamboo rods were used and 4 independent variables measured in the rod. Initially, the effect of the multicollinearity could be verified through the correlation matrix of the independent variables and the varience inflation factors.To select the independent variables two methods were used: Stepwise and K component retention. The alternatives used were component regression and Ridge regression. In general, in only one situation the variable selection methods behave adequately while multicollinearity is present among the independent variables, that is the multivaried method of retention K=3 component for the covariate matrix, model of Spurr. The estimative of alternative methods showed similar responses, however, the principal component regression yields the best results.

Keywords: Bamboo, Multivariate analysis, Regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O bambu é uma planta de grande utilidade industrial, sendo importante alternativa para a produção de biomassa, particularmente para o Brasil, país que intensamente usa para produção de papel e energia (BRITTO et al., 1997).

O colmo de bambu, principal componente utilizado da planta, tem tecido constituído de fibras e vasos. Segundo Montalvão et al. (1984), as presenças de fibras e amido em colmos são as principais propriedades tecnológicas do bambu. Esses componentes caracterizam o uso industrial do bambu, como matéria-prima fibrosa, para produção de celulose para fabricação de papel (AZZINI et al., 2000).

No Brasil a espécie que apresenta maiores áreas de plantio é *Bambusa vulgaris*. A região Nordeste tem a maior área plantada do mundo, nos estados Maranhão, Paraíba e Pernambuco. *Bambusa vulgaris* é, portanto, uma planta essencial ao desenvolvimento florestal do Nordeste brasileiro, usada como matéria prima industrial para a produção e papel de fibra longa, que exibe maior resistência para uso em embalagens (BONILLA,1991).

Apesar do alto potencial produtivo da *bambusa vulgaris*, há poucas pesquisas em silvicultura e manejo. Assim, a produção de biomassa por unidade de área e tempo é pouco estudada no Brasil. Isso é preocupante, implicando decisões sobre espécies, variedades e clones cultivados, tipo de solo, espaçamento e idade de corte (BONILLA, 1991).

Para suprir o aumento da demanda no Nordeste do Brasil, além de procurar elevar a área de cultivo, tem-se, o que é mais critico, diminuido o ciclo de corte raso, que era de 2,5 anos, para um ano, podendo acarretar a deterioração paulatina da produtividade e da sobrevivência das touceiras dos plantios de bambu (MENDES, 2005).

O objetivo deste estudo visa estabelecer uma metodologia que permita quantificar a biomassa verde da unidade amostral de bambu existente em povoamento de *bambusa vulgaris* na cidade de Goiana-PE, por meio de modelos de regressão lineares simples e múltiplos. Especificamente procurou-se:

Comparar os métodos de seleção de variáveis, univariado e multivariado, Stepwise e Retenção por K componentes, respectivamente, e desenvolver modelos de regressão lineares simples ou múltiplos para estimar o peso da haste do bambu, quando as variáveis são correlacionadas, ou seja, existe um alto grau de multicolinearidade.

Utilizar duas técnicas como alternativas para o alto grau de muticolinearidade, a regressão com componentes principais, ou seja, os componentes serão as variáveis explicativas e a regressão Ridge que consiste em fazer uma transformação dos parâmetros estimados.

Para tal, a pesquisa foi desenvolvida em plantios da empresa Agrimex Agroindustrial Excelsior S. A., localizados no Engenho Itapirema a 50 km do Recife na cidade de Goiana-PE, sobre a BR 101 Norte. A empresa usa a matéria prima industrial, *bambusa vulgaris*, para a produção de papel de embalagem de cimento, papel duplex. E pretende melhorar sua produção por meio das estimativas da biomassa verde da haste do bambu.

2. REVISÃO DE LITERATURA

2.1 **Bambu**

O bambu pertence à família das gramíneas e sempre esteve presente na cultura e na vida diária do homem primitivo de todos os continentes com exceção da Europa que não tem o bambu na forma nativa. O uso do bambu no oriente remonta há quase cinco mil anos e há mais de quinhentos anos na América do Sul (VASCONCELLOS, 2007).

Segundo Silva (2005), embora seja uma gramínea, os bambus possuem hábito arborescente e da mesma forma que as árvores apresentam uma parte aérea constituída pelo colmo, folhas e ramificações e outra subterrânea composta pelo rizoma e raiz (Figura 1).

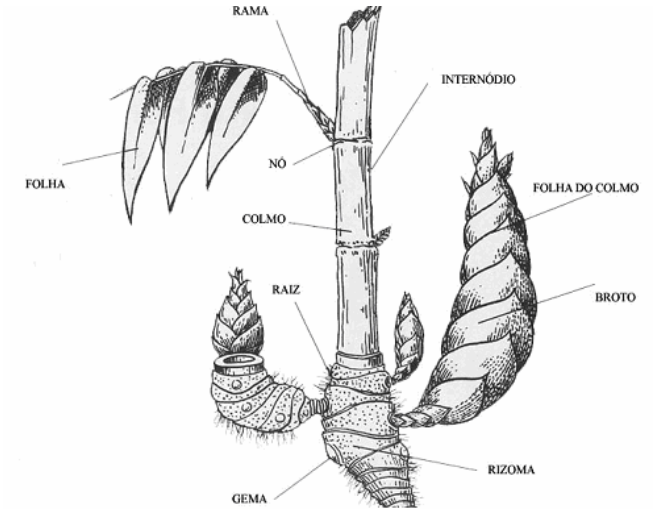

Figura 1. Partes do bambu (SILVA, 2005).

Rizoma é um caule subterrâneo dotado de nós e entrenós com folhas reduzidas a escamas e que se desenvolve paralelamente a superfície do solo. Basicamente, existem dois grupos distintos de bambus quanto ao tipo de rizoma: os que formam touceiras (simpodiais) e os alastrantes (monopodiais) (SILVA, 2005).

As raízes dos bambus partem dos rizomas, se lançam na projeção da copa numa profundidade diretamente proporcional as dimensões de cada espécie (VASCONCELLOS, 2007)

As folhas dos bambus respondem pela função de elaborar as substâncias necessárias ao rápido crescimento dessa planta através do processo da fotossíntese (TEIXEIRA, 2003).

Os colmos são segmentados por nós e os espaços compreendidos entre dois nós são denominados entrenós, que são menores na base, aumentam o seu comprimento na parte mediana e reduzem novamente o tamanho na medida em que vão aproximando do ápice. As paredes dos nós são mais finas que as paredes dos entrenós e recebem o nome de diafragma. Entre as espécies, os colmos diferem pela cor, diâmetro, comprimento, espessura da parede, comprimento dos entrenós e outras características (SILVA,2005).

Os colmos assim como as folhas têm também a capacidade de realizar a fotossíntese, contudo estruturar a parte aérea, armazenar e conduzir a seiva bruta e elaborada constitui-se nas suas principais funções (AZZINI, 1987).

Não há concordância entre os pesquisadores sobre o número de gêneros e espécies de Bambu no mundo, acredita-se que há de 75 a 111 gêneros já em termos de espécies de 1200 a 1600. (KALEY, 2000; LONDOÑO, 2004).

Na América Latina, o Brasil é o país com maior diversidade, reúne 81 % dos gêneros de bambus lenhosos (LONDOÑO,2004).

Segundo Silva (2005), as espécies exógenas mais comuns são: *Bambusa vulgaris Schrad, B. vulgaris var. vittata, B.tuldoides, Dendrocalamus giganteus* e algumas espécies de *Phyllostachys*. Essas espécies, todas de origem asiática, foram trazidas pelos primeiros colonizadores portugueses, posteriormente, pelos orientais e se difundiram facilmente pelo país. Essa dispersão ocorreu de forma tão generalizada que muitos leigos acreditam ser nativa a espécie *Bambusa vulgaris*.

O *Bambusa vulgaris* tem hábito entouceirante, ou seja, os rizomas se desenvolvem formando uma touceira densa e concêntrica. Tal hábito dificulta a determinação de sua produção de forma indireta. O gênero Bambusa possui apenas bambus de rizomas paquimorfos, ou seja, de colmos bem juntos e é utilizado como polpa de papel além de fonte de bebida alcoólica. O papel de bambu tem a mesma qualidade que o papel de madeira e é o uso industrial do bambu de maiores proporções do mundo. O Brasil é o único país das Américas a ter uma indústria de

papel de bambu, com uma grande plantação no Estado do Maranhão (VASCONCELLOS, 2007).

No Nordeste brasileiro são cultivados mais de 100 mil hectares de *Bambusa vulgaris* Schrad. para a produção de papel cartão duplex. O papel feito de celulose de bambu é um dos únicos que, por sua porosidade e resistência, serve para fazer filtros descartáveis para café. Por sua resistência ao rasgo, é escolhido para embalagens como sacos de cimento. As duas unidades de celulose pertencentes a um mesmo grupo se localizam nos municípios de Coelho Neto, no Maranhão e Jaboatão dos Guararapes no Estado de Pernambuco. A capacidade instalada é de 72 000 toneladas/ano e já iniciou um plano de expansão que elevará a produção anual de papel para 144 000 toneladas/ano (ITAPAGÉ, 2007).

2.2 **Propriedades e uso do bambu**

Do ponto de vista agronômico os colmos e folhas do bambu são largamente empregados na produção de papel, desinfetantes, baterias, tecidos, cervejas e outra centena de usos. O bambu pode substituir a madeira em diversas aplicações, e com isso diminuir o impacto ambiental ocasionado pelo corte predatório de árvores essenciais ao equilíbrio do ecossistema (TEIXEIRA, 2003).

O bambu oferece seis vezes mais celulose que o pinheiro que mais rápido cresce. Suas fibras são muito resistentes e tem qualidade igual ou superior à fibra de madeira. (VASCONCELLOS, 2007).

A produção de bambu depende do crescimento, que incluem variáveis como: números de touceiras/ha, número de colmos por touceira e diâmetro e altura de colmos, sendo a produção estimada pela biomassa produzida (BONILLA,1991).

A produtividade do bambu deve ser mensurada de acordo com a sua finalidade e pode ser quantificado pela biomassa, número de brotos ou número de colmos por uma determinada área. No Brasil, praticamente, inexistem trabalhos científicos relativos à produtividade dos bambus (VASCONCELLOS, 2007).

A biomassa do bambu pode ser expressa por massa verde ou massa seca. A massa verde se refere ao material fresco amostrado, contendo uma proporção variável de água. A massa seca corresponde à massa de uma árvore, de um arbusto ou seus componentes, sendo obtido após a secagem do material em estufa. (CALDEIRA, 2003).

Estimativa da Biomassa

Teixeira (2003) definiu a biomassa como a quantidade de material vegetal contida por unidade de área numa floresta e expressa em unidade de massa. Em geral, os componentes utilizados na medição da biomassa são; biomassa vertical acima do solo, composição das árvores e arbustos, composição da serapilheira, troncos caídos (fitomassa morta acima do solo) e composição de raízes (biomassa abaixo do solo).

Segundo Caldeira (2003), o termo biomassa representa a matéria orgânica armazenada em um determinado ecossistema, pois especifica o valor numérico dos componentes presentes, além de ser fundamental nos estudos de ciclagem de nutrientes, conversão de energia, absorção e armazenamento de energia solar e também possibilita tirar conclusões para uma exploração racional dos ecossistemas. Segundo Pardé (1980), expressar a biomassa em matéria seca é vantajoso na aplicação em determinados mercados madeireiros, para a necessidade de explicar a produtividade biológica dos ecossistemas e pela facilidade em comparações e cálculos.

Segundo Husch et al. (1982) o aumento do uso de medições de peso para produtos florestais desenvolveu uma necessidade de estimar o peso da madeira em árvores em pé.

A porção viva acima do solo, em geral, é onde se concentra a maior parte da biomassa, sendo o componente estimado com maior freqüência. A estimativa da biomassa abaixo do solo não é um componente analisado com muita freqüência, na maioria das vezes, requer grandes investimentos financeiros (CALDEIRA, 2003).

Para Teixeira (2003), a quantificação da biomassa florestal pode ser feita por dois métodos, o método direto, no qual há a determinação do peso da biomassa fresca e da biomassa seca e o método indireto, que estima a biomassa por meio de modelos matemáticos a partir de dados de inventários florestais fazendo a relação de variáveis como o volume da madeira, o DAP (diâmetro à altura do peito), altura comercial do tronco, diâmetro da copa e a altura total das árvores.

Para Caldeira (2003), a quantificação da biomassa fornece informações sobre magnitude, qualidade e distribuição dos produtos da floresta que não se encontram nos tradicionais mapas dos ecossistemas.

Segundo Husch et al. (1982), o peso verde ou seco da madeira em pé pode ser estimado de duas maneiras, obtendo o volume individual das árvores de uma tabela de volume convencional ou de medições individuais do fuste, e converter para peso usando uma apropriada relação peso volume. O outro modo é obter o peso das árvores individuais diretamente.

A biomassa de uma árvore expressa em peso pode ser determinada diretamente, por meio da determinação do peso verde de cada componente (SOARES; HOSOKAWA, 1984).

Silva (1996) destacou a alta correlação que o DAP apresenta com o peso dos componentes das árvores, atingindo um coeficiente de determinação (R^2) maior do que 0,95 e distribuição de resíduos aceitáveis para os modelos ajustados. Afirmou também que devido a menor correlação do DAP com os pesos dos galhos e das folhas, faz-se necessário o seu uso na forma quadrática e associada com a altura total para composição do modelo matemático.

A estimativa da biomassa pode ser obtida a custos sensivelmente mais baixos e com a mesma eficiência do que pelos métodos baseados em diâmetro e altura, quando se faz uso do método dos dois diâmetros e da relação hipsométrica, associada à equação linear múltipla (FRANCO, 1996).

Análise de Componente Principal (PCA)

Análise de Componente Principal (PCA) é uma técnica estatística que transforma um conjunto de “p” variáveis em um conjunto com um número menor “k” ($k<p$) de variáveis não-correlacionadas, que explica uma parcela substancial das informações do conjunto original (KENDALL, 1950).

Além da redução da dimensionalidade dos dados, o objetivo principal é o de explicar a estrutura de variância e covariância de um vetor aleatório composto de “p” variáveis aleatórias, através da construção de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si (JOHNSON; WICHERN, 1982).

A suposição de normalidade do vetor aleatório não é requisito necessário para que a técnica de análise de componente principal possa ser utilizada. Entretanto, quando a distribuição de probabilidade do vetor aleatório for normal p-variado, as componentes principais, além de não correlacionadas, são independentes e têm distribuição normal.

Os componentes principais dependem somente da matriz de covariância ou da matriz de correlação do vetor aleatório de interesse. Geometricamente, os componentes principais são as coordenadas dos pontos amostrais em um sistema de eixos obtidos pela rotação do sistema de eixos originais, na direção da variabilidade máxima dos dados (JOHNSON; WICHERN, 1982).

A análise de componentes principais é uma técnica multivariada utilizada para examinar a relação entre grande número de variáveis. O peso relativo de cada variável na composição de cada eixo é medido através de sua correlação com este eixo. Desta forma, a PCA pode ser utilizada como instrumento de seleção de variáveis, na medida em que aquelas com maior peso na construção das componentes, são as que possivelmente melhor represente o conjunto estudado (KENDALL, 1950).

Segundo Johnson e Wichern (1982), PCA fornece um método multivariado para a redução de variáveis em casos em que ocorrem multicolinearidade nas variáveis independentes.

Seja X o vetor de p variáveis originais $x' = (X_1, \ldots, X_p)$, com vetor de médias $\mu = (\mu_1, \ldots, \mu_p)'$, e matriz de covariância $\sum_{pxp}$. Sejam $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ os autovalores da matriz $\sum_{pxp}$, com respectivos autovetores normalizados $e_1, e_2, \ldots, e_p$, em que $e_i = (e_1, e_2, \ldots, e_p)'$, isto é, os autovetores e_i satisfazem as seguintes condições:

$e_i' e_j = 0; \forall i \neq j;$

$e_i' e_i = 1; \forall i = 1,2,\ldots, p;$

$\sum_{kxk} e_i = \lambda_i e_i; \forall i = 1,2,\ldots, p.$

Então, a j-ésima componente principal da matriz $\sum_{pxp}$, j = 1, 2,..., p é definida como:

$$PC_j = e_j' \chi = e_{j1} X_1 + e_{j2} X_2 + \ldots + e_{jk} X_p$$

O valor esperado e variância da j-ésima componente principal são, respectivamente, iguais a:

$$E[PC_j] = e_j'\mu = e_{j1}\mu_1 + e_{j2}\mu_2 + \ldots + e_{jp}\mu_p$$

$$Var[PC_j] = e_j'\sum_{pxp} e_j = \lambda_j$$

Os componentes são não correlacionados, ou seja, $Cov(PC_i, PC_j) = 0; i \neq j$. Cada autovalor λ_i representa a variância de uma componente principal PC_i. Como os autovalores são ordenados em ordem decrescente, a primeira componente é a de maior variabilidade e a p-ésima é a de menor.

A proporção explicada pela j-ésima componente principal é definida como:

$$\frac{\lambda_i}{Traço(\sum_{pxp})} = \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j};$$

Em que $Traço(\sum_{pxp})$ é a variância total de X.

A correlação estimada entre a j-ésima componente principal e a variável aleatória $X_i, i = 1,2,\ldots, p$ é dada de duas formas, através da matriz de covariância ou da matriz de correlação:

I. Utilizando a matriz de covariância:

$$r_{PC_jX_i} = \frac{e_{ji}\sqrt{\lambda_j}}{\sqrt{\sigma_{ii}^2}};$$

Em que σ_{ii}^2 é a variância da variável aleatória X_i.

II. Utilizando a matriz de Correlação:

$$r_{PC_jX_i} = e_{ji}\sqrt{\lambda_j}\ .$$

2.5 **Modelos de Regressão**

Nos inventários florestais, envolvendo espécies arbóreas, usa-se frequentemente a modelagem estatística para fazer estimação da biomassa.

Segundo Bonilla (1991) o uso de equações para quantificar a biomassa é evidente em levantamentos quantitativos porque permite, a partir de medições detalhadas em um número limitado de árvores criteriosamente selecionadas dentro de um povoamento florestal.

Para estimativa da biomassa (peso) é visto com bastante freqüência a utilização do método estatístico de análise de regressão.

Os modelos lineares são aqueles cujos coeficientes são estimados diretamente pelo método de mínimos quadrados (PAULA, 2004).

Análise de regressão é uma metodologia estatística para predizer valores de uma ou mais variáveis respostas (dependentes), baseando-se em um conjunto de valores de variáveis preditoras (independentes). Tem por objetivo descrever através de uma equação matemática a relação existente entre duas ou mais variáveis a partir de “n” observações dessas variáveis (DRAPER; SMITH , 1981).

Em muitas situações a explicação de um fenômeno através de apenas uma variável independente (regressão simples) pode não ser satisfatória, pois essa variável será apenas uma componente influenciando na variação da resposta estudada. Nestes casos deve-se propor uma equação envolvendo mais de uma variável independente, chamado de modelo de regressão múltiplo (PAULA, 2004).

O método mais usual para estimação dos parâmetros do modelo de regressão é conhecido como método de mínimos quadrados. O principio fundamental do método de mínimos quadrados consiste em determinar estimativas dos parâmetros que minimizem o quadrado das distâncias entre os valores observados (valores reais) e os correspondentes ao modelo proposto (valores estimados), onde essas diferenças são definidas como resíduo (DRAPER; SMITH , 1981).

O modelo de Regressão linear múltipla pode ser descrito da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$$

Ou matricialmente como: sugerido por Montgomery et al. (2001).

$$Y = X\beta + \varepsilon$$

Em que,

$$Y=\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \qquad X=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}; \qquad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \qquad \beta=\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix};$$

Em geral, Y é um vetor (nx1) de observações, X é uma matriz (nxp) de níveis das variáveis regressoras, β é um vetor (p+1x1) de coeficientes da regressão e ε é um vetor (nx1) de erros aleatórios.

Para o modelo as seguintes suposições são considerados:

- A variável resposta (dependente) é uma função linear dos parâmetros.
- Os valores das variáveis independentes (preditoras) são fixos.
- Os erros são variáveis aleatórias com média zero e variância constante. $E(\varepsilon_i)=0; Var(\varepsilon_i)=\sigma^2$.
- Os erros são não-correlacionados. $Corr(\varepsilon_i, \varepsilon_j)=0; \forall i \neq j$

A estimação pontual dos parâmetros do modelo pelo método de mínimo quadrado é dada por:

$$\hat{\beta}=\left(X^{'}X\right)^{-1}X^{'}Y$$

Com as propriedades:

$$E(\hat{\beta})=\beta; \qquad Cov(\hat{\beta})=\left(X^{'}X\right)^{-1}\sigma^2$$

A regressão linear simples é um caso particular da regressão linear múltipla quando, k=1, em que número de variáveis explicativas é igual a um.

2.5.1 **Teste de significância da regressão**

Segundo Montgomery et al. (2001), o teste para significância da regressão, é um teste para determinar se há uma relação linear entre a variável resposta, Y, e alguma das variáveis regressoras, $X_1,...,X_p$. Este procedimento é freqüentemente conhecido como teste global de adequacidade do modelo. As hipóteses apropriadas são:

H_0: $\beta_1=\cdots=\beta_p=0$

H_1: $\beta_j \neq 0$ para pelo menos um j.

A rejeição da hipótese nula implica que pelo menos uma variável regressora, $X_1,...,X_p$, contribui significativamente para o modelo.

O procedimento que define a estatística do teste utiliza as seguintes somas de quadrados para a obtenção do teste da razão de verossimilhança:

SQT = Y'Y - $\frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$ {soma de quadrados total};

SQE = Y'Y - $\hat{\beta}$'X'Y {soma de quadrados dos resíduos};

SQR = $\hat{\beta}$' X'Y- $\frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$ {soma de quadrados da regressão}.

As duas últimas somas de quadrados são uma decomposição da soma total, ou seja,

$$SQT = SQR + SQE.$$

A estatística obtida pela razão de verossimilhança é:

$$F_0 = \frac{MQR}{EMQ},$$

e as fórmulas de MQR, média dos quadrados da regressão e EMQ, erro médio quadrático, são fornecidas na Tabela 1.

Sob a hipótese nula que especifica a adequação do modelo, F_0 tem distribuição F de Snedecor com p graus de liberdade no numerador e n-p-1 graus de liberdade no denominador.

A regra de decisão do teste é rejeitar H_0 se

$$F_0 > F_{(\alpha,p,n-p-1)},$$

onde $F_{(\alpha,p,n-p-1)}$ é o percentil da distribuição F de Snedecor a um nível α de significância.

Tabela 1 - Análise de Variância para Significância de uma Regressão Múltipla.

Fonte de Variação	Graus de liberdade	Soma de Quadrados	Quadrado Médio	F_0
Regressão	p	SQR	$MQR = \frac{SQR}{p}$	$\frac{MQR}{EMQ}$
Resíduo	$n-p-1$	SQE	$EMQ = \frac{SQE}{(n-p)-1}$	
Total	$n-1$	SQT		

Uma outra medida muito utilizada na análise de regressão é o percentual da variação dos dados explicado pelo modelo (coeficiente de determinação), que é dado por:

$$R^2 = \frac{SQR}{SQT}.$$

2.5.2 **Análise de Resíduo**

De acordo com Paula (2004), o exame da variável resposta não é muito útil em análise de regressão porque os valores observados devem depender das variáveis explicativas. Para resolver este problema é introduzido o conceito de resíduo do modelo. O resíduo é definido para uma observação como a diferença entre o seu valor real e o ajustado. Para facilitar a análise gráfica é recomendável trabalhar com os resíduos padronizados

$$e_i = \left(Y_i - \hat{Y}_i\right) \Big/ \sqrt{EMQ(1-h_{ii})} \ ,$$

em que h_{ii} é o i-ésimo elemento da diagonal da matriz $H = X(X'X)^{-1}X'$ e EMQ = $\frac{Y'Y - \hat{\beta}'X'Y}{n-p}$ é o erro médio quadrático. Um fato importante é que H é a transformação linear cujos valores ajustados são obtidos a partir dos valores observados. A principal vantagem de usar estes resíduos é na identificação de pontos aberrantes, porque os resíduos padronizados, se o modelo for adequado, devem se comportar aproximadamente como uma normal de média zero e variância

um. Logo, se para uma observação o valor do resíduo padronizado é maior do que 2, isto é um indicativo de que esta observação não está bem ajustada pelo modelo (MONTGOMERY et al, 2001).

Uma análise de resíduos é feita a partir de gráficos específicos. Para cada suposição é utilizado um gráfico adequado (DRAPER; SMITH , 1981).

2.5.3 **Seleção de Variáveis**

Para encontrar o modelo mais eficiente, algumas variáveis independentes poderão ser eliminadas do modelo por não contribuírem significantemente, além de que podem trazer problemas de multicolinearidade, então uma solução é a utilização dos métodos estatísticos para a seleção de variáveis (MONTGOMERY et al, 2001).

Existem vários métodos estatísticos para a seleção de variáveis. No entanto, a comparação dos métodos de seleção de variáveis univariado e multivariado é muito interessante, pois pode aperfeiçoar o modelo com o ganho razoável de tempo, no caso em que a quantidade de variáveis for elevada.

Segundo Draper e Smith (1981) o método estatístico univariado de Stepwise, é provavelmente o mais utilizado, pois não requer o cálculo de todas as regressões possíveis. No entanto, não é tão simples quando o número de variáveis é muito grande.

Araújo (2005) utilizou análise de componentes principais e o método de Stepwise na seleção de variáveis independentes na construção de tabelas volumétricas para *leucaena leucocephaia* (Lam) de Wit.

Claro et al. (2003) utilizou componentes principais em um estudo de identificação de fatores que afetam a cultura do feijão em Minas Gerais.

2.5.3.1 **Método Stepwise**

De acordo com Montgomery et al. (2001) pelo fato da avaliação e todas as regressões possíveis poderem ser computacionalmente pesadas, vários métodos têm sido desenvolvidos para avaliar subgrupos de modelos de regressão ou por adição ou por eliminação de regressores. Esses métodos são geralmente referidos como um procedimento de Stepwise. Eles podem ser classificados dentro de três categorias:

- Seleção Forward
- Eliminação Backward
- Regressão Stepwise

Em que a Regressão Stepwise (3) é uma combinação dos outros dois procedimentos (1) e (2).

A Regressão Stepwise é uma modificação da Seleção de Forward em que cada passo todos os regressores que entrarem no modelo, previamente, são reavaliados a partir de suas estatísticas F (parciais). Um regressor adicionado no passo anterior pode agora ser redundante por causa da relação entre eles. Se a estatística F(parcial) para a variável é menor que a F para a retirada, essa variável é mantida no modelo (DRAPER; SMITH, 1981).

Os limites de F para adição e retirada de variáveis não precisam ser iguais. Geralmente, o limite F para retirada de uma variável é especificado como sendo menor que o limite para a inclusão de variáveis (PAULA, 2004).

A Regressão Stepwise pode resultar em combinações lineares das variáveis independentes que não apresentam a menor soma de quadrados dos resíduos (SQE) e pode apresentar deficiência nas estimativas dos parâmetros quando as variáveis independentes são correlacionadas, ou seja, quando existir multicolinearidade (MONTGOMERY et al., 2001).

2.5.3.2 **Método Retenção por K componentes**

O método de seleção de variáveis de Stepwise não é indicado para variáveis correlacionadas, podem ocorrer erros nas estimativas dos parâmetros. Já o método de retenção por K componentes tende a se comportar melhor na existência da multicolinearidade.

O método consiste em considerar o autovetor correspondente ao menor autovalor e a variável com maior coeficiente em valor absoluto é removida e no passo seguinte considera o próximo autovetor correspondente ao segundo maior autovalor sem a variável retirada anteriormente e se repete o processo que é finalizado quando resulta em K componentes, em que K é arbitrário.

Segundo CADIMA(2001), o processo de exclusão das variáveis inicia em ordem inversa de importância dos componentes.

2.6 **Multicolinearidade**

Segundo Montgomery et al. (2001) é o nome dado ao problema geral que ocorre quando duas ou mais variáveis explicativas são muito correlacionadas entre si, o que torna difícil, utilizando-se apenas o modelo de regressão, distinguir suas influências separadamente. Uma das suposições do modelo de regressão, afirma que nenhuma relação linear exata pode existir entre quaisquer covariáveis ou combinações lineares destas. Quando se viola esta hipótese se têm o problema de multicolinearidade perfeita. Por outro lado, se as variáveis não estão correlacionadas entre si, denomina-se, este caso, ausência de multicolinearidade, sendo chamada de ortogonal à regressão com estas variáveis. O caso intermediário, muito comum em problemas reais, ocorre quando a correlação entre duas ou mais variáveis é alta, sendo esta situação chamada de alto grau de multicolinearidade.

A análise do grau de multicolinearidade pode ser feita através da matriz de correlações que mede a dependência linear de primeira ordem entre as variáveis explicativas (PAULA, 2004).

Outros critérios são: o valor do determinante de $X'X$, o R^2 obtido quando uma das covariáveis é eliminada e o R^2 do modelo de regressão entre cada uma das variáveis independentes sobre todas as outras remanescentes.

Multicolinearidade Perfeita: Neste caso é impossível estimar o vetor de parâmetros porque a matriz $X'X$ tem o determinante igual a zero; logo, não possui inversa (MONTGOMERY et al., 2001).

Ausência de Multicolinearidade: Quando as variáveis não estão correlacionadas entre si a matriz $X'X$ é diagonal. Portanto, a estimativa do coeficiente de uma variável independente, no modelo de regressão múltipla é obtida pelo método de mínimos quadrados (MONTGOMERY et al., 2001).

Alto Grau de Multicolinearidade: Como uma etapa preliminar de verificação de um modelo ajustado, devemos observar, na matriz de correlação, se existem pelo menos duas variáveis muito correlacionadas. Se isto ocorrer, o coeficiente de uma variável irá depender da outra, não refletindo, assim, o efeito individual da variável a qual está associado, mas, somente um efeito parcial ou marginal, condicionado a outra variável (MONTGOMERY et al., 2001).

A conseqüência mais grave deste problema é que o teste de significância do coeficiente de uma variável independente, sendo esta correlacionada com alguma

outra, pode indicar sua exclusão do modelo, mesmo que exista uma forte relação linear desta com a variável resposta (MONTGOMERY et al., 2001).

A multicolinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

Segundo Montgomery et al. (2001) algumas indicações da presença de multicolinearidade são:

- Valores altos do coeficiente de correlação.
- Grandes alterações nas estimativas dos coeficientes de regressão, quando uma variável independente for adicionada ou retirada do modelo, ou quando uma observação for alterada ou eliminada.
- A rejeição da hipótese H_o: $\beta_1 = \beta_2 = ... = \beta_k = 0$ por meio da realização do teste *F*, mas nenhuma rejeição das hipóteses H_o: $\beta_i = 0$, $i = 1, 2,..., k$, por meio da realização dos testes *t* sobre os coeficientes individuais de regresssão.
- Obtenção de estimativas para os coeficientes de regressão com sinais algébricos contrários àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis ou de experiências anteriores sobre o fenômeno estudado.
- Obtenção de intervalos de confiança com elevadas amplitudes para os coeficientes de regressão, associados a variáveis independentes importantes.

Para evitar os problemas provocados pela multicolinearidade o método mais simples é a eliminação, do modelo completo, das variáveis com os coeficientes estatisticamente não significativos, os quais são identificados a partir do teste individual para β_j, ou, mais efetivamente, através dos métodos de seleção variáveis, para encontrar o melhor subconjunto de variáveis independentes (MONTGOMERY, 2001).

A segunda alternativa é o método de regressão denominado de “Ridge”, que tem o objetivo de melhorar a precisão dos parâmetros estimados do modelo.

A terceira maneira de obter um modelo adequado, quando algumas variáveis independentes são muito correlacionadas, é a partir da técnica de componentes principais, que tem a vantagem de não descartar nenhuma variável explicativa (PAULA, 1997).

2.7 **Modelo de Regressão para as Componentes Principais**

Se as componentes principais têm um significado intuitivo, é melhor expressar a equação da regressão em termos das componentes. Nos demais casos, é mais conveniente recolocar o modelo nas variáveis originais (MONTGOMERY et al., 2001).

$$Y_i = \beta_0 + \beta_1 PC_{1i} + \beta_2 PC_{2i} + ... + \beta_p PC_{pi} + \varepsilon_i$$

Todas as suposições, estimação e análise do modelo de regressão múltipla devem ser a atendidas para o caso do modelo de regressão para as componentes principais, no entanto, as componentes principais, agora variáveis explicativas, são não correlacionados. E com isso eliminando o problema da multicolinearidade (DRAPER; SMITH, 1981).

2.8 **Regressão Ridge**

Uma alternativa ao método original dos mínimos quadrados, regressão *Ridge (regressão corrigida*), pode ser útil no combate à multicolinearidade (DRAPER; SMITH , 1981). Ele consiste em fazer uma transformação dos parâmetros estimados, através da adição de uma constante λ à equação:

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}_c = \mathbf{X}'\mathbf{y}$$

O valor da constante λ deve ser tal que a média quadrática do erro do estimador corrigido, $\hat{\boldsymbol{\beta}}_c$, seja menor do que a variância do estimador de mínimos quadrados, $\hat{\boldsymbol{\beta}}$. A escolha dessa constante não é simples; à medida que ela cresce, alguns dos parâmetros $\hat{\boldsymbol{\beta}}_c$ variarão drasticamente. Para algum valor de λ, as estimativas $\hat{\boldsymbol{\beta}}_c$ estabilizarão. O objetivo é selecionar algum valor pequeno de λ, no qual as estimativas $\hat{\boldsymbol{\beta}}_c$ sejam estáveis (MONTGOMERY et al., 2001).

Segundo Montgomery et al. (2001), o valor da constante λ é obtido através da seguinte equação:

$$\lambda = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

Em que $\hat{\sigma}^2$ e $\hat{\beta}$ são calculados pelo método de mínimos quadrados e “p” é o número de variáveis.

3. MATERIAIS E MÉTODOS

3.1 **Área de Estudo**

O experimento foi desenvolvido no Engenho Itapirema, localizada a 50 km da Cidade do Recife, sobre a BR 101 Norte. Precipitação média anual de 1900 mm. Com topografia ondulada a fortemente ondulada. Diversos tipos de solos variando do arenoso ao argiloso, e de profundidade variável. A vegetação é caracterizada por plantios de cana-de-açúcar nas áreas mais planas e bambu nas mais onduladas.

3.2 **Coleta de Dados**

Primeiramente a coleta de dados foi realizada identificando o número de touceira de cada parcela, seguidamente em cada touceira se contou o número de hastes. Para cada haste da touceira se mediu a altura, circunferência na altura da base (0,10m), circunferência na altura do peito (1,30m). Posteriormente procedeu-se o corte e pesagem das hastes principais sem galhos secundários, folhas e outros.

Foram selecionadas 3 parcelas ao acaso, cada uma medindo 15x15 m². Uma parcela com 22 touceiras, a segunda com 33 touceiras e a terceira com 25 touceiras, totalizando 80 touceiras e 450 hastes.

3.3 **Descrição das Variáveis**

Observa-se na Tabela 2 que a variável combinada, CAB^2H, é o produto entre a circunferência da base ao quadrado e a altura, nas ciências florestais é muito utilizada como estimativa do volume, mas a principal finalidade dessa variável combinada é a normalização dos resíduos nos modelos de estimativa da biomassa verde.

Tabela 2. Descrição das variáveis.

H	Altura de Bambu (m)
CAB	Circunferência na base da haste (cm)
CAP	Circunferência a 1,30 m do solo(cm)
CAB^2H	Variável combinada (m^3)
P	Peso (biomassa verde) da haste bambu (kg)

3.4 **Métodos Estatísticos**

Inicialmente, apresentaram-se os histogramas das variáveis para verificar como se comporta os dados e calcularam-se algumas estatísticas descritivas de cada variável, mínimo, máximo, média, mediana, desvio padrão, variância, coeficiente de variação, matriz de covariância e se analisou, dando ênfase a matriz de correlação para verificar se poderia existir problema de colinearidade ou multicolinearidade.

É aplicado o teste de multicolinearidade, fator de inflação da variância, para verificar se existirá erro nas estimativas do modelo. Os parâmetros da equação de regressão serão estimados a partir das inter-relações da variável dependente (resposta) e das variáveis independentes (preditoras). Todas as estimativas feitas através do método de mínimos quadrados consistem no procedimento matemático para minimizar os erros quadráticos.

De acordo com Montgomery et al. (2001) os efeitos de multicolinearidade podem ser facilmente demonstrados. Os elementos da diagonal da matriz C = $(X'X)^{-1}$ podem ser escritos como:

$$C_{jj} = \frac{1}{(1-R_j^2)} \qquad j = 1, 2, ..., p$$

Sendo R_j^2 o coeficiente de determinação múltipla, resultante da regressão de x_j nos outros p - 1 regressores. Claramente, quanto mais forte for a dependência linear de x_j nos regressores restantes, e por conseguinte mais forte a colinearidade, maior será o valor de R_j^2. Lembre-se de que $V(\hat{\beta}_j) = \sigma^2 C_{jj}$. Logo, diz-se que a variância de $\hat{\beta}_j$ é "inflacionada" pela quantidade $(1-R_j^2)^{-1}$.

Dessa maneira, define-se o fator de inflação da variância para $\hat{\beta}_j$ como:

$$FIV(\hat{\beta}_j) = \frac{1}{(1-R_j^2)} \qquad j = 1, 2, ..., p$$

Esses fatores são uma importante medida da extensão da presença de multicolinearidade. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Alguns autores sugeriram que se qualquer fator de

inflação da variância exceder 10, então a multicolinearidade será um problema.

Outros autores consideram esse valor muito liberal e sugerem que os fatores de inflação da variância não devem exceder 4 ou 5 (MONTGOMERY et al., 2001).

Depois de realizado o teste de multicolinearidade e verificada a existência da mesma, analisaram-se os diagramas de dispersão entre a variável resposta (Peso da haste do bambu, P) contra as variáveis explicativas (H, CAB, CAP, CAB^2H).

Para utilização do método de seleção de variáveis de retenção por K componentes foi utilizada a análise de componentes principais na matriz de covariância e na matriz de correlação e interpretado todos os componentes.

Aplicaram-se as técnicas de seleção de variáveis, Stepwise e retenção por K componentes, para a construção dos modelos de regressão para a estimativa da biomassa verde do bambu, para constatar qual modelo se comportaria da melhor forma e a sua eficiência. O método de seleção de variáveis de retenção por K componente, foi aplicado de duas formas através da matriz de covariância e da matriz correlação para os K=2 e K=3, pois a variabilidade quando o k>1 para qualquer caso tratado é superior a 80%.

Para tentar suprir essas dificuldades outras duas técnicas foram utilizadas para construção do modelo mais eficiente, Regressão com componentes principais e Regressão Ridge (corrigida).

A regressão com componentes principais foi obtida através do método de seleção de variável de Stepwise, onde os componentes serão as variáveis explicativas, pois a variabilidade dos componentes não é a maior importância, mas sim o seu grau de contribuição para o modelo.

Na regressão Ridge, primeiramente, foi calculado o valor do λ através das estimativas do modelo proposto pelo método de Stepwise e aplicado o método baseado nesse modelo.

A constatação do modelo mais eficiente foi feita na validação das suposições da regressão linear múltipla e no coeficiente de determinação (R^2).

As análises foram feitas no programa estatístico R, versão 2.6.1, o qual pode ser obtido gratuitamente no site <http://www.r-project.org>.

4. RESULTADOS E DISCUSSÃO

4.1 **Estatística descritiva**

Na Figura 2 está ilustrada a distribuição das variáveis explicativas (independentes) e variável resposta (dependente). Observa-se que a altura, o CAB e o CAP têm tendência de distribuições similares, já a variável combinada CAB²H tem distribuição similar a variável resposta (dependente), Peso (P).

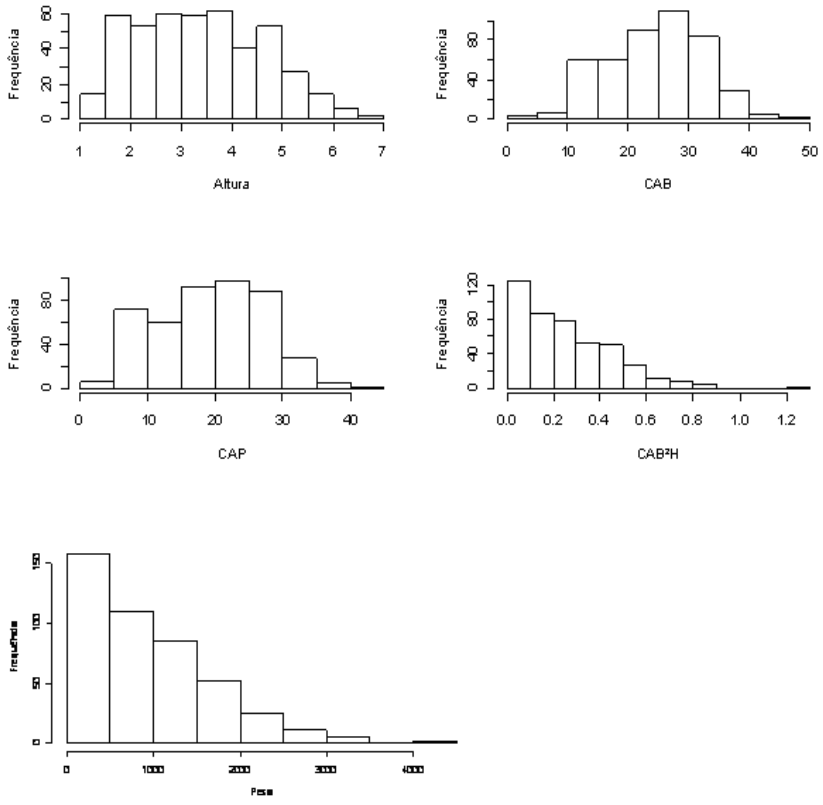

Figura 2. Histogramas das variáveis independentes altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB²H) e a variável dependente peso da haste (P).

Os resultados das medidas descritivas para todas as variáveis são dados na Tabela 3, a seguir, em que é possível notar uma discrepância muito acentuada entre as variâncias das variáveis. O número total de observações é de 450 unidades.

Tabela 3. Estatísticas descritivas das variáveis independentes altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB^2H) e a variável dependente peso da haste (P).

Estatítica	H	CAB	CAP	CAB^2H	P
Mínimo	1,32	2,1	3,2	0,0006	80
Máximo	6,95	48,7	42,1	1,2318	4200
Mediana	3,375	25,4	19,8	0,2153	800
Média	3,43	24,31	19,05	0,2588	975,2
Desvio padrão	1,2253	7,9126	7,8551	0,2103	769,5825
Variância	1,5013	62,6087	61,7026	0,0442	592257,3
Coeficiente de variação	0,3572	0,3255	0,4123	0,8126	0,7892

As matrizes de covariância e de correlação entre as variáveis explicativas estão nas Tabelas 4 e 5.

Tabela 4. Matriz de covariância entre as variáveis independentes altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB^2H).

Variável	H	CAB	CAP	CAB^2H
H	1,5013	7,4286	8,2425	0,2267
CAB	7,4286	62,6088	57,4752	1,4889
CAP	8,2425	57,4752	61,7021	1,4878
CAB^2H	0,2267	1,4889	1,4878	0,0442

Segundo Silva e Silva (1982, apud BONILLA, 1991, p.59) pode-se considerar como correlação nula entre duas variáveis quando os valores dos índices de correlação estiverem entre -0.400 e 0.400. Assim sendo, observa-se que as correlações entre todas as variáveis são não nulas, ou seja, existe de um alto grau de correlação entre as variáveis explicativas podendo resultar no problema de multicolinearidade.

Tabela 5. Matriz de correlação entre as variáveis independentes altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB^2H).

Variável	H	CAB	CAP	CAB^2H
H	1,0000	0,7662	0,8564	0,8801
CAB	0,7662	1,0000	0,9247	0,8949
CAP	0,8564	0,9247	1,0000	0,9008
CAB^2H	0,8801	0,8949	0,9008	1,0000

4.2 **Teste de multicolinearidade**

Por meio do teste de multicolinearidade identificou que poderão existir problemas nas estimativas dos coeficientes de regressão, pois os fatores de inflação da variância para os estimadores foram todos acima de cinco, como observado na Tabela 6.

Tabela 6. Fatores de inflação da variância para os estimadores das variáveis independentes altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB²H).

Variável Retirada	FIV ($\hat{\beta}_j$)
H	7,8308
CAB	7,5988
CAP	7,1022
CAB²H	5,2576

De acordo com Dias et al. (2003) Esse procedimento foi realizado porque, quando ocorre multicolinearidade, os testes estatísticos podem falhar em detectar diferenças significativas entre os fatores.

4.3 **Diagramas de dispersão**

Nota-se na Figura 3 que só a variável explicativa CAB²H tem uma relação linear com a variável resposta, P.

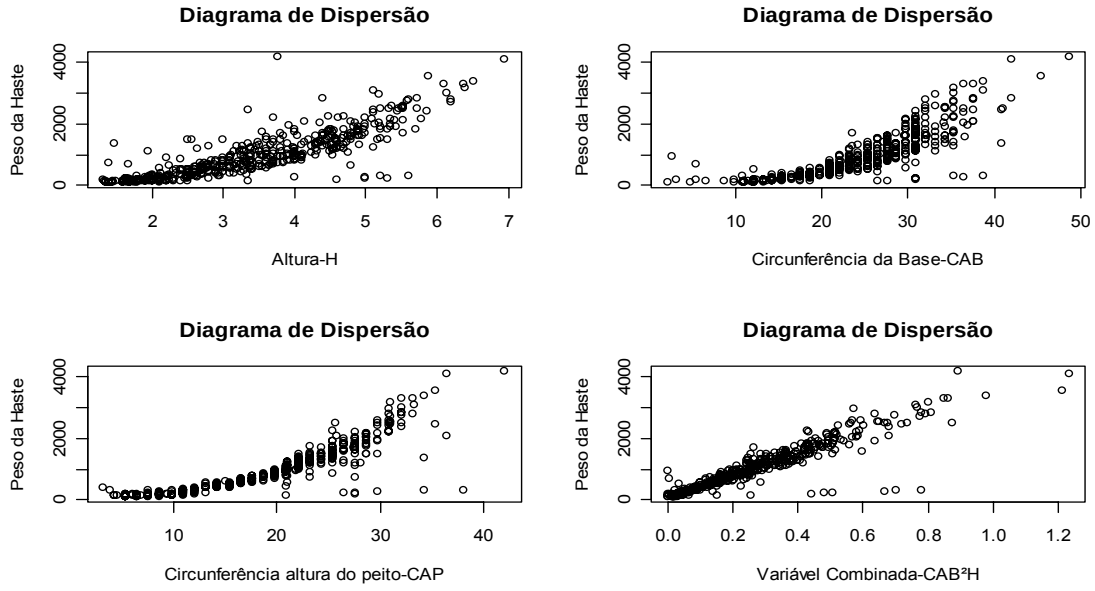

Figura 3. Diagramas de dispersão das variáveis altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB²H) versus variável resposta (P).

4.4 **Análise Componente Principais**

As Tabelas 7 e 8 informam a importância dos componentes através da variabilidade por duas maneiras, matriz de covariância e matriz de correlação, respectivamente, e a correlação entre as variáveis explicativas e os componentes .

Pela matriz de covariância a primeira componente (PC1) que é, basicamente, um índice do peso da haste, explica quase 80% da variabilidade total e em que o maior coeficiente de grandeza se refere a variável CAP. Já a segunda componente (PC2) é uma comparação entre as variáveis H,CAB²H e CAB com a variável CAP. A terceira componente (PC3) é representada pelas variáveis H e CAB²H, sendo dominada pela variável H, que tem o maior coeficiente numérico em valor absoluto. A quarta componente (PC4) representa a variável CAB²H, que é a de menor variância amostral, como visto na Tabela 3, sendo, portanto, uma componente de pouca importância prática. A Tabela 7 mostra uma correlação maior da componente, em valor absoluto, com a variável CAB²H e correlação próxima de zero com as outras variáveis.

Tabela 7. Componentes principais pela matriz de covariância das variáveis independentes, altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB²H).

Variáveis	PC1	PC2	PC3	PC4
H	0,8299	-0,2520	-0,4976	0,0044
CAB	0,9808	0,1950	-0,0027	0,0001
CAP	0,9812	-0,1926	0,0131	$7,9708e^{-6}$
CAB²H	0,9165	-0,0236	-0,2308	-0,3258
Estatística	-	-	-	-
Desvio Padrão	10,9863	2,1833	0,6206	0,0687
Proporção da variância (%)	79,27%	15,75%	4,49%	0,49%
Proporção acumulada (%)	79,27%	95,02%	99,51%	100%

Pela matriz de correlação a primeira componente (PC1) ou índice do peso da haste, explica um pouco mais de 65% da variabilidade total e sendo a variável CAP a de maior coeficiente grandeza. A segunda componente (PC2) é uma comparação entre as variáveis H e CAB, sendo dominada pela variável H, que tem o maior coeficiente numérico em valor absoluto. A terceira componente (PC3) é uma comparação entre as variáveis CAP e CAB²H, sendo dominada pela variável CAB²H, que tem o maior coeficiente numérico em valor absoluto. A quarta componente (PC4) obtida pela matriz de correlação tem uma importância maior do que a obtida

pela matriz de covariância e representa uma comparação entre as variáveis CAB e CAP, sendo dominada pela variável CAB, que tem o maior coeficiente numérico em valor absoluto.

Tabela 8. Componentes principais pela matriz de correlação entre as variáveis independentes, altura, circunferência na base da haste (CAB), circunferência a 1,30m do solo (CAP), combinada (CAB^2H).

Variáveis	PC1	PC2	PC3	PC4
H	0,9202	-0,3779	0,0593	0,0828
CAB	0,9441	0,3000	-0,0083	0,1365
CAP	0,9692	0,0956	0,1884	-0,1269
CAB^2H	0,9673	-0,0290	-0,2371	-0,0828
Estatística	-	-	-	-
Desvio Padrão	1,9008	0,4927	0,3089	0,2210
Proporção da variância (%)	65,03%	16,85%	10,56%	7,56%
Proporção acumulada (%)	65,03%	81,88%	92,44%	100%

Segundo Johnson e Wichern (1982) quando existe uma discrepância muito acentuada entre as variâncias das variáveis originais, cada componente passa a ser extremamente dominada por uma variável em particular, algo que torna as componentes sem muita utilidade na prática. Este problema pode ser amenizado se uma transformação for efetuada nos dados originais, de modo a equilibrar os valores de variância ou colocar os dados na mesma escala de medida. Uma das transformações mais comuns é aquela em que cada variável é padronizada pela sua média e desvio padrão, sendo a técnica de componente principal aplicada à matriz de covariância das variáveis padronizadas. Este procedimento é equivalente a obterem-se as componentes principais através da matriz de correlação das variáveis originais. No entanto, apesar das variâncias das variáveis originais serem bastante distintas, pode-se perceber adiante que os resultados através da matriz de covariância são mais satisfatórios que os obtidos com a matriz de correlação.

4.5 **Seleção de variáveis**

4.5.1 **Stepwise**

Utilizando o método de Stepwise é proposta a seguinte equação:

$$\hat{P}_i = 9.503 - 18.58 CAB_i + 36.674\ CAP_i + 2777.018\ CAB^2H_i$$

A Tabela 9 mostra as estimativas dos parâmetros e se percebe que apenas o intercepto não é significativo para o modelo, no entanto, por existir uma correlação entre as variáveis explicativas, multicolinearidade, todos os parâmetros podem estar mal estimados.

Tabela 9. Estimativa dos parâmetros do modelo proposto pelo método de Stepwise.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	9,503	61,315	0,877
β_1	-18,58	4,662	$7,87e^{-05}$ ***
β_2	36,674	4,827	$1,78e^{-13}$ ***
β_3	2777,018	153,780	$< 2e^{-16}$ ***

*** significativo a 1% ,** significativo a 5% , * significativo a 10%.

As contribuições das variáveis são vistas na Tabela 10 e verifica-se que são significativas para o modelo. A estimativa da variância (σ^2) é dada por S² = 76110, enquanto adequacidade do modelo foi obtido pelo coeficiente de determinação, R²= 0,8723.

Tabela 10. Análise de variância da equação proposta pelo método de Stepwise.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
CAB	1	185015684	185015684	$< 2,2e^{-16}$ ***
CAP	1	22142707	22142707	$< 2,2e^{-16}$ ***
CAB²H	1	24819867	24819867	$< 2,2e^{-16}$ ***
Resíduo	446	33945264	76110	
Total	449	265923522		

*** significativo a 1% ,** significativo a 5% , * significativo a 10%.

Na Figura 4, observa-se um comportamento razoável dos resíduos, ou seja, todas as suposições de regressão em relação ao erro são atendidas. Apesar disto, as variáveis explicativas têm um alto grau de correlação, ou seja, existe o problema de multicolinearidade. Logo, o método de seleção de variáveis de Stepwise não se comporta adequadamente quando o número de variáveis explicativas for maior que um, pois as estimativas dos parâmetros e o coeficiente de determinação (R²) não são confiáveis.

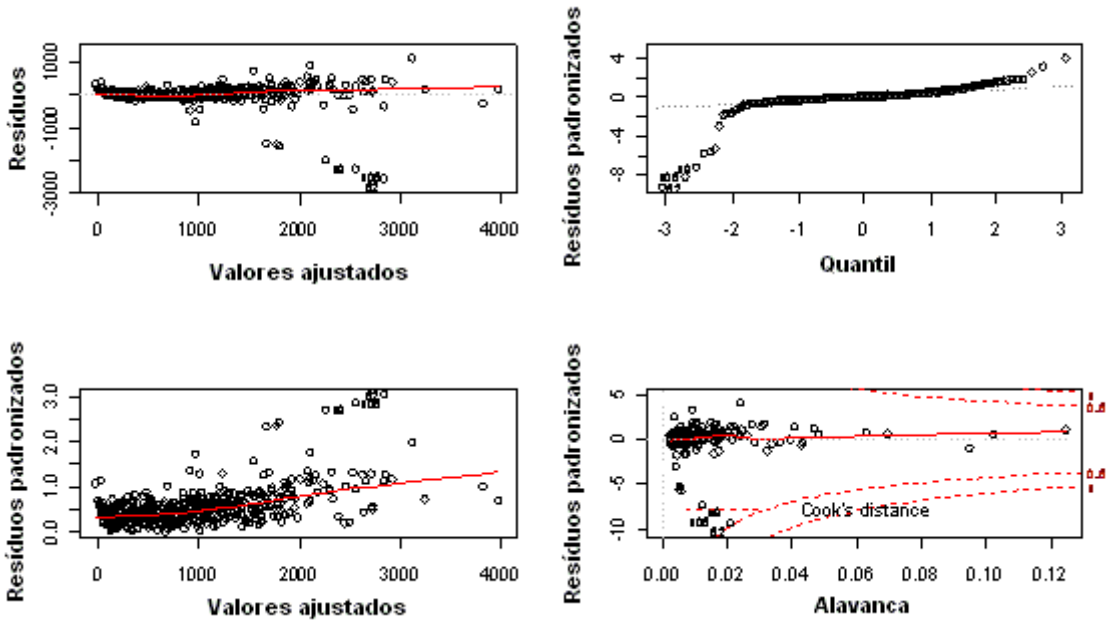

Figura 4. Gráfico dos Resíduos da equação obtida pelo método de Stepwise.

4.5.2 **Retenção por K componentes**

4.5.2.1 **Retenção por K=2 componentes pela matriz de covariância**

Juntando as duas primeiras componentes pela matriz de covariância, o percentual chega a 95,02%, resultando em um K=2. Portanto, por este método se deve excluir o terceiro e quarto componentes (PC3 e PC4) de menor importância. Os resultados apresentados pelo método de seleção de variáveis por retenção de K=2 componentes estão na Tabela 11. No segundo componente (PC2) rejeita-se a variável circunferência na altura da Base (CAB), pois tem o maior coeficiente em valor absoluto. No primeiro componente (PC1) rejeita-se a variável CAP que tem o maior valor absoluto.

Tabela 11. Matriz de autovetores obtidos com utilização da matriz de covariância para retenção por K=2 componentes.

Variável	PC1	PC2	PC3	PC4
H	0,0926	-0,1414	-0,9824	0,0790
CAB	0,7064	**0,7069**	-0,0341	0,0134
CAP	**0,7015**	-0,6930	0,1659	0,0009
CAB^2H	0,0175	-0,0023	-0,0782	-0,9968

Logo, a equação proposta pelo método de retenção de K=2 componentes é:

$$\hat{P}_i = -58.79 + 74.97\ H + 3001.00\ CAB^2H_i \text{ ; i=1,...,n}$$

Na Tabela 12 são mostradas as estimativas dos parâmetros e se percebe que nem todos são significativos para o modelo, pois o intercepto (β_0 = 0.2766) não é significativo. Vale ressaltar que os outros parâmetros são significativos até mesmo ao nível de significância de 1%.

Tabela 12. Estimativa dos parâmetros do modelo proposto pelo método de retenção de K=2 componentes pela matriz de covariância.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	-58,79	53,97	0,2766
β_1	74,97	23,50	0,00152***
β_2	3001,00	136,96	< 2e^{-16} ***

A contribuição de cada variável é vista na Tabela 13. Verifica-se que as variáveis são significativas para o modelo. A estimativa da variância (σ^2) é dada por S² = 83979, enquanto que o coeficiente de determinação foi de R²= 0,8588.

Tabela 13. Análise de variância da equação proposta pelo método de retenção de K=2 componentes pela matriz de covariância.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
H	1	188067423	2239,46	< 2,2e^{-16} ***
CAB²H	1	40317501	480,09	< 2,2e^{-16} ***
Resíduo	447	37538599	83979	
Total	449	265923523		

Na Figura 5, observa-se um comportamento razoável dos resíduos, ou seja, todas as suposições de regressão em relação ao erro são atendidas. Apesar dos resíduos obedecerem todas as suposições do modelo de regressão, o intercepto não é significativo e as variáveis explicativas têm um alto grau de correlação, ou seja, existe o problema de multicolinearidade. Logo, o método de seleção de variáveis de retenção por k=2 componentes não se comporta adequadamente, pois

as estimativas dos parâmetros e o coeficiente de determinação (R^2) não são confiáveis.

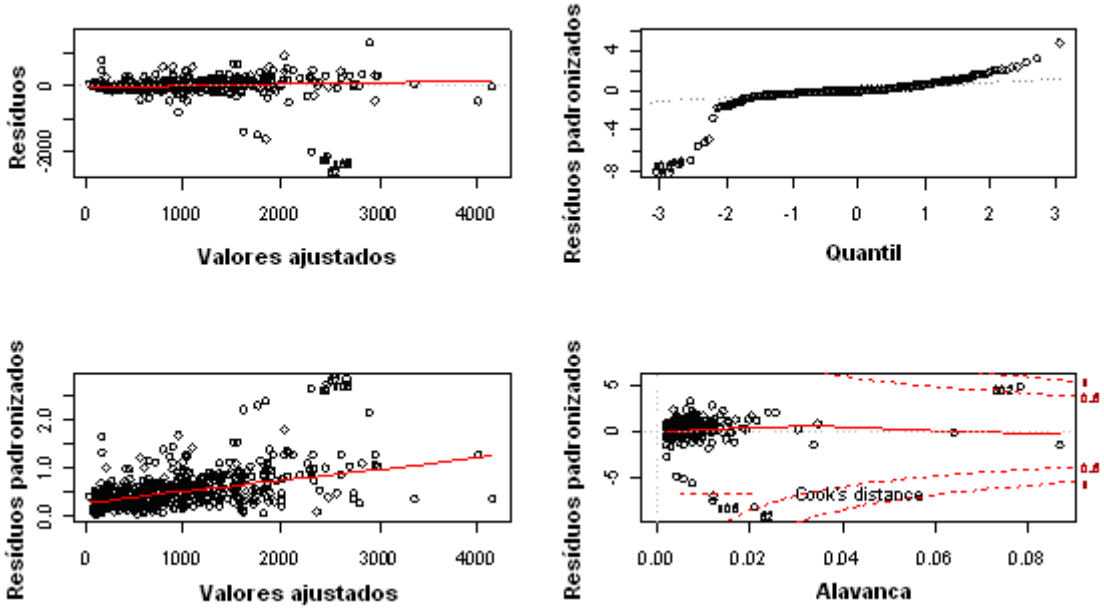

Figura 5. Gráficos dos resíduos para equação obtida pelo método de retenção de K=2 componentes pela matriz de covariância das variáveis explicativas.

4.5.2.2 **Retenção por K=3 componentes pela matriz de covariância**

Juntando os três primeiros componentes pela matriz de covariância, esse percentual chega a 99,51%, resultando em um K=3. Portanto, por este método deve-se excluir o quarto componente (PC4) de menor importância. Os resultados apresentados pelo método de seleção de variáveis por retenção de K=3 componentes estão na Tabela 14. No terceiro componente (PC3) rejeita-se a variável altura (H), pois tem o maior coeficiente em valor absoluto. No segundo componente (PC2) rejeita-se a variável CAB e na primeira componente (PC1) rejeita-se a variável CAP.

Tabela 14. Matriz de autovetores pela matriz de covariância para retenção por K=3 componentes.

Variável	PC1	PC2	PC3	PC4
H	0,0926	-0,1414	**-0,9824**	0,0790
CAB	0,7064	**0,7069**	-0,0341	0,0134
CAP	**0,7015**	-0,6930	0,1659	0,0009
CAB^2H	0,0175	-0,0023	-0,0782	-0,9968

Logo, a equação proposta pelo método de retenção de K=3 componentes é exatamente o modelo de Spurr ou da variável combinada (SPURR, 1592).

$$\hat{P}_i = 98.85 + 3385.47\ CAB^2H_i$$

Na Tabela 15 são mostradas as estimativas dos parâmetros e percebe-se que são todos altamente significativos para o modelo.

Tabela 15. Estimativas dos parâmetros do modelo proposto pelo método de retenção de K=3 componentes pela matriz de covariância.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	98,85	21,90	$8,18e^{-06}$ ***
β_1	3385,47	65,70	$< 2e^{-16}$ ***

A contribuição da variável é vista na Tabela 16 e se verifica que são significativas para o modelo. A estimativa de variância (σ^2) é dada por S² = 85699. A adequacidade do modelo foi de 85,56% (R²= 0,8556).

Tabela 16. ANOVA do modelo proposto pelo método de retenção de K=3 componentes pela matriz de covariância.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
CAB²H	1	227530496	227530496	$< 2,2e^{-16}$ ***
Resíduo	448	38393027	85699	
Total	449	265923523		

Na Figura 6, observa-se um comportamento razoável dos resíduos, ou seja, todas as suposições de regressão em relação ao erro são atendidas. Apesar do método de seleção de variáveis de retenção de K componentes propor um modelo linear simples, para k=3. Esse modelo atende todas as suposições de regressão, logo suas estimativas e coeficiente de determinação são confiáveis.

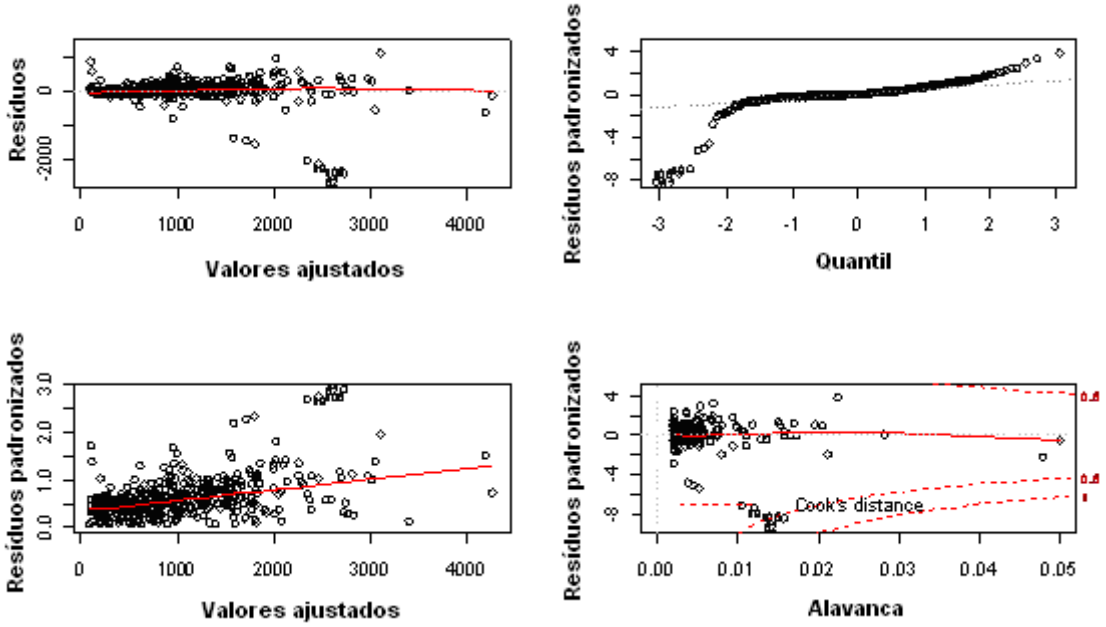

Figura 6. Gráficos dos resíduos para equação obtida pelo método de retenção de K=3 componentes pela matriz de covariância das variáveis explicativas.

4.5.2.3 **Retenção por K=2 componentes pela matriz de correlação**

Juntando as duas primeiras componentes pela matriz de correlação, o percentual chega a 81,88%, resultando em um K=2. Os resultados apresentados pelo método de seleção de variáveis por retenção de K=2 componentes estão na Tabela 17. No segundo componente (PC2) rejeita a variável circunferência de Base (CAB), pois tem o maior coeficiente em valor absoluto. No primeiro componente (PC1) rejeita a variável CAP.

Tabela 17. Matriz de autovetores pela matriz de correlação para retenção por K=2 componentes.

Variável	PC1	PC2	PC3	PC4
H	0,4841	**-0,7669**	0,1920	0,3749
CAB	0,4966	0,6089	-0,0268	0,6179
CAP	**0,5098**	0,1939	0,6103	-0,5744
CAB²H	0,5089	-0,0590	-0,7680	-0,3843

Assim, a equação proposta pelo método de retenção de K=2 componentes é:

$$P_i = 50.499 + 3.101\,CAB + 3281.054\,CAB^2 H_i$$

Na Tabela 18 são apresentadas as estimativas dos parâmetros e percebe-se que só o coeficiente da variável combinada (CAB²H), β_2, é altamente significativo para o modelo.

Tabela 18. Estimativas dos parâmetros do modelo proposto pelo método de retenção de K=2 componentes pela matriz de correlação.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	50,499	64,838	0,436
β_1	3,101	3,914	0,429
β_2	3281,054	147,277	$< 2e^{-16}$ ***

A contribuição da variável é vista na Tabela 19 e se verifica que estas são significativas para o modelo. A estimativa da variância (σ^2) é dada por S² = 85770, sendo a adequacidade do modelo foi de 85,58% (R²= 0,8558).

Tabela 19. ANOVA do modelo proposto pelo método de retenção de K=2 componentes pela matriz de correlação.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
CAB	1	185015684	185015684	$< 2{,}2e^{-16}$ ***
CAB²H	1	42568650	42568650	$< 2{,}2e^{-16}$ ***
Resíduo	447	38339189	85770	
Total	449	265923523		

Na Figura 7, observa-se um comportamento razoável dos resíduos, ou seja, todas as suposições de regressão em relação aos erros são atendidas. Apesar dos resíduos obedecerem todas as suposições do modelo de regressão, apenas um parâmetro é significativo e as variáveis explicativas têm um alto grau de correlação, ou seja, existe o problema de multicolinearidade. Assim sendo, o método de seleção de variáveis de retenção por K=2 componentes não se comporta adequadamente, pois as estimativas dos parâmetros e o coeficiente de determinação (R²) ou adequacidade do modelo não são confiáveis.

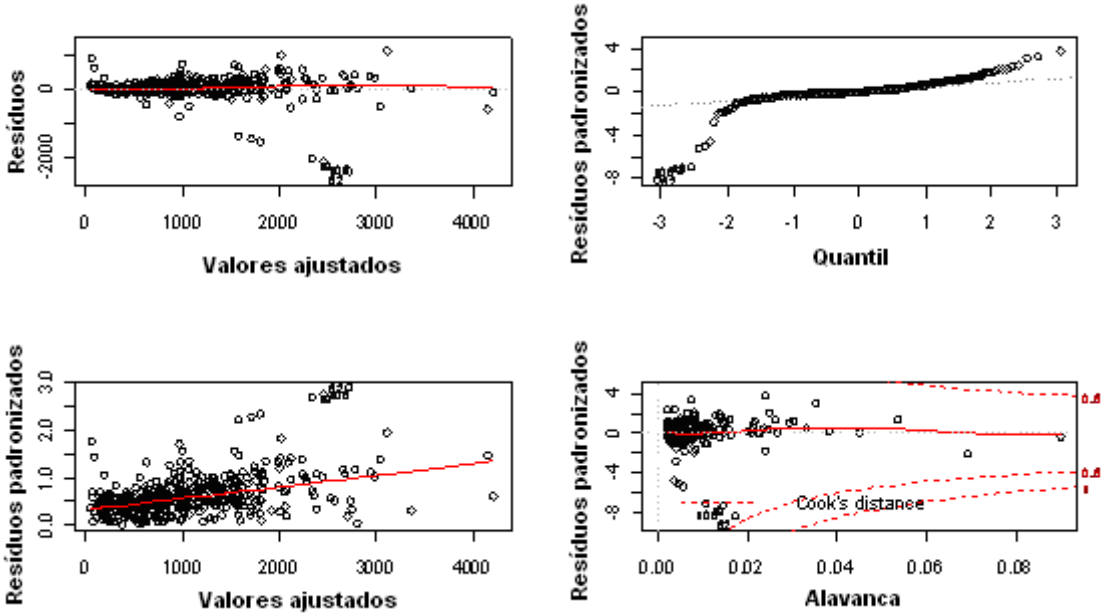

Figura 7. Gráficos dos resíduos para a equação obtida pelo método de retenção de K=2 componentes pela matriz de correlação das variáveis explicativas.

4.5.2.4 **Retenção por K=3 componentes pela matriz de correlação**

Ao juntar os três primeiros componentes pela matriz de correlação, dada na Tabela 8, o percentual chega a 92,44%, resultando em um K=3. Portanto, por este método se deve excluir o quarto componente (PC4) de menor importância. Os resultados apresentados pelo método de seleção de variáveis por retenção de K=3 componentes estão na Tabela 20. No terceiro componente (PC3) rejeita-se a variável combinada (CAB²H), por apresentar o maior coeficiente em valor absoluto. No segundo componente (PC2) rejeita-se a variável altura (H) e na primeira componente (PC1) rejeita-se a variável CAP.

Tabela 20. Matriz de autovetores pela matriz de correlação para retenção por K=2 componentes.

Variável	PC1	PC2	PC3	PC4
H	0,4841	**-0,7669**	0,1920	0,3749
CAB	0,4966	0,6089	-0,0268	0,6179
CAP	**0,5098**	0,1939	0,6103	-0,5744
CAB²H	0,5089	-0,0590	**-0,7680**	-0,3843

Logo, a equação proposta pelo método de retenção de K=3 componentes é:

$$P_i = -996.896 + 81.127\, CAB_i$$

Na Tabela 21 observa-se as estimativas dos parâmetros e percebe-se que são todos significativos para o modelo até mesmo ao nível de significância de 1%.

Tabela 21. Estimativas dos parâmetros do modelo proposto pelo método de retenção de K=3 componentes pela matriz de correlação.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	-996,896	64,78	$< 2e^{-16}$ ***
β_1	81,127	2,535	$< 2e^{-16}$ ***

A contribuição da variável é vista na Tabela 22 e verifica-se que são significativas para o modelo. A estimativa da variância (σ^2) é dada por S^2 =180598, em que o modelo está 69,57% adequado, ou seja, coeficiente de determinação foi de R^2= 0,6957.

Tabela 22. ANOVA do modelo proposto pelo método de retenção de K=3 componentes pela matriz de correlação.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
CAB	1	185015684	185015684	$< 2,2e^{-16}$ ***
Resíduo	448	80907839	180598	
Total	449	265923523		

Na Figura 8, observa-se que os resíduos não têm um comportamento razoável, ou seja, alguma das suposições de regressão em relação ao erro não foi atendida. E observa-se no gráfico dos resíduos pelos valores ajustados que não há uma relação linear, ou seja, provavelmente faz-se necessária a inclusão de um termo quadrático, como, a variável combinada, CAB^2H. Apesar do método de seleção de variáveis de retenção de K componentes propor um modelo linear simples, para K=3. Esse modelo não atende as suposições de regressão, logo suas estimativas e coeficiente de determinação não são confiáveis.

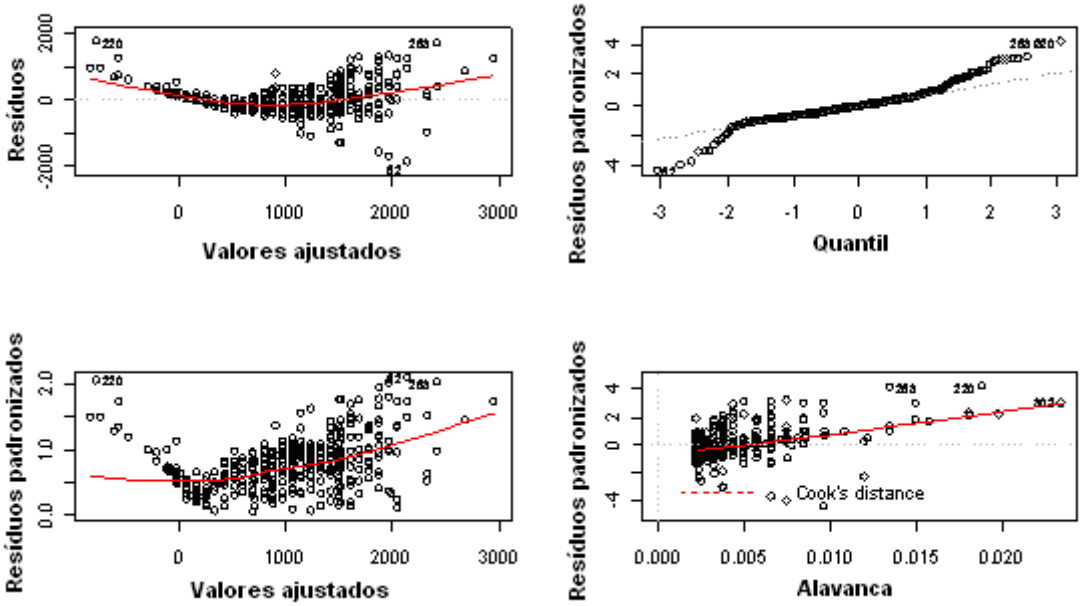

Figura 8. Gráficos dos resíduos para a equação obtida pelo método de retenção de K=3 componentes pela matriz de correlação das variáveis explicativas.

4.6 **Modelo de regressão com os componentes principais**

Para esse método os componentes principais serão as variáveis explicativas e aplicando-se o método de Stepwise para a seleção dos componentes obteve-se a seguinte equação:

$$\hat{P}_i = 975.202 - 370.920\ PC1_i - 65.711\ PC2_i - 270.211\ PC3_i - 486.493\ PC4$$

Na Tabela 23 são apresentadas às estimativas dos parâmetros e nota-se que apenas o intercepto não é significativo para o modelo ao nível de significância adotado. No entanto, por existir uma correlação entre as variáveis explicativas, multicolinearidade, todos os parâmetros podem estar mal estimados, apesar de grande parte deles ser significativo ao nível de 1%.

Tabela 23. Estimativas dos parâmetros do modelo proposto pelo método de Stepwise para a seleção dos componentes.

Parâmetros	Estimativas	Erros Padrão	p-valor
β_0	975,202	13,019	$< 2e^{-16}$ ***
β_1	370,920	6,857	$< 2e^{-16}$ ***
β_2	-65,711	26,453	0,0134*
β_3	-270,211	42,217	$3,93e^{-10}$ ***
β_4	-486,493	58,986	$1,84e^{-15}$ ***

As contribuições das variáveis são vistas na Tabela 24 e verifica-se que são significativas para o modelo. A estimativa da variância (σ^2) é dada por S² = 76278, enquanto que o coeficiente de determinação é dado por R²= 0,8724.

Tabela 24. ANOVA do modelo proposto pelo método de Stepwise para a seleção dos componentes.

Fontes de Variação	Graus de liberdade	Somas de Quadrados	Quadrados Médios	p-valor
PC1	1	223195419	223195419	$< 2,2e^{-16}$ ***
PC2	1	470659	470659	0,01336*
PC3	1	3124928	3124928	$3,932e^{-10}$ ***
PC4	1	5188719	5188719	$1,839e^{-15}$ ***
Resíduo	445	33943797	76278	
Total	449	265923522		

Na Figura 9, observa-se que um comportamento razoável dos resíduos, ou seja, todas as suposições de regressão em relação ao erro são atendidas. Os resíduos obedeceram todas as suposições do modelo de regressão. As variáveis explicativas ou componentes, não são correlacionados, ou seja, não existe problema de multicolinearidade. Aplicou-se o método de seleção de variáveis de Stepwise nos componentes porque não é a variabilidade do componente que vai obter o melhor modelo e sim a contribuição do componente. O quarto componente (PC4) é o de menor variabilidade, porem é ele que contribui para a suposição do resíduo ser válida. A Figura 10 mostra o gráfico de resíduos sem o quarto componente (PC4) e

observa-se que uma das suposições residuais não é atendida. Logo, precisa-se de um termo quadrático no modelo. Como visto na Tabela 7, o PC4 representa a variável combinada, CAB²H, pois as correlações com as outras variáveis explicativas são próximas de zero.

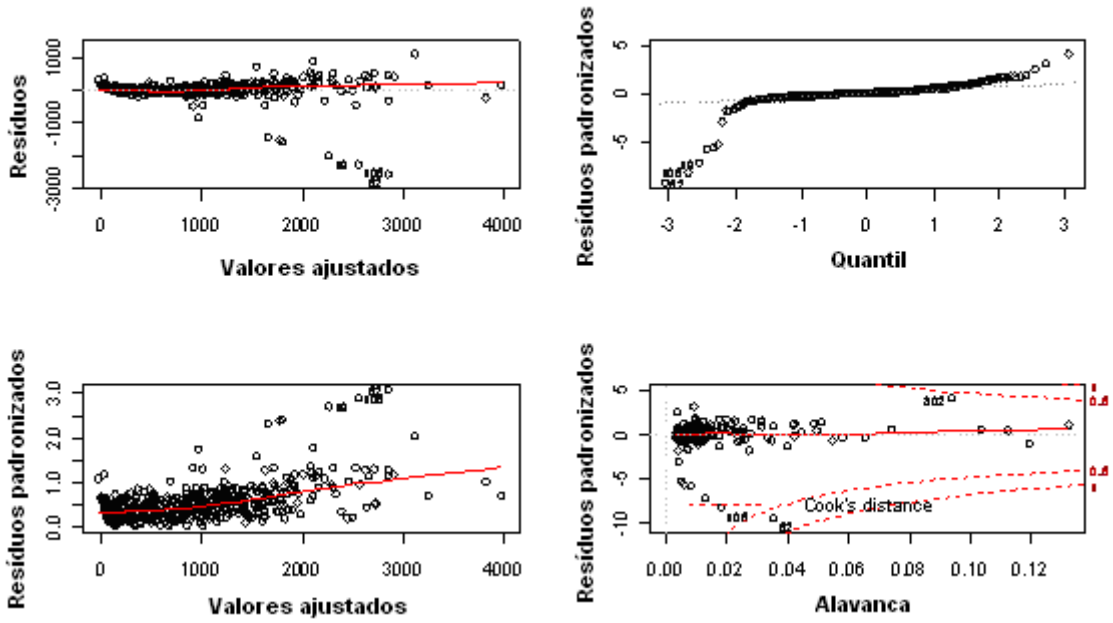

Figura 9. Gráficos dos resíduos para a equação obtida pelo método de Stepwise para a seleção dos componentes.

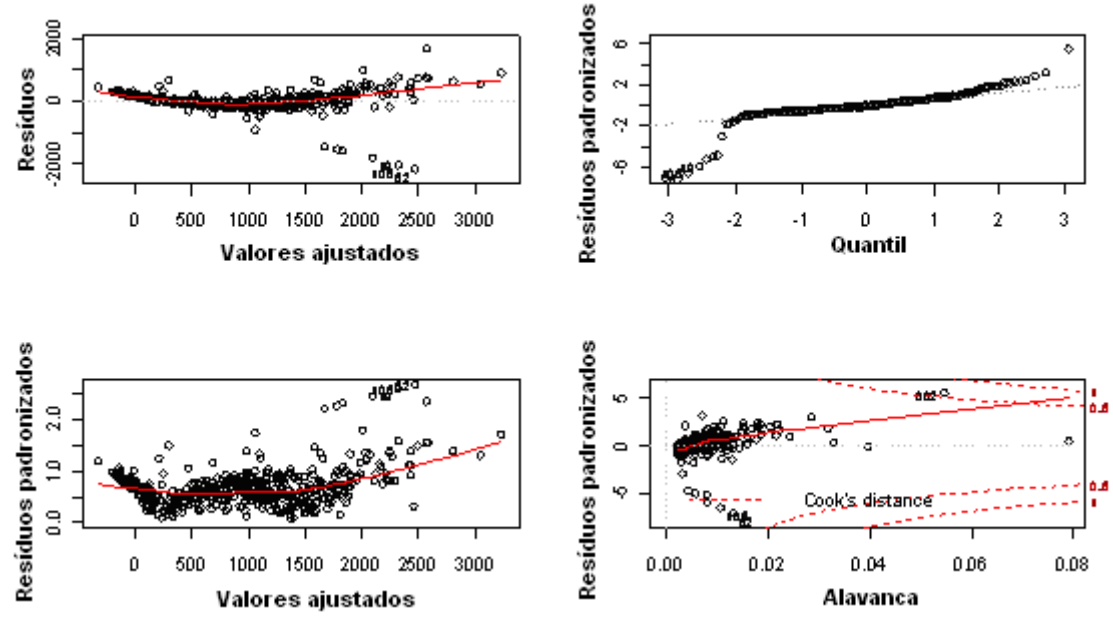

Figura 10. Gráficos dos resíduos para a equação obtida pelo método de Stepwise para a seleção dos componentes sem o quarto componente (PC4).

4.7 **Regressão Ridge**

Utilizou-se as estimativas do modelo proposto pelo método de Stepwise e chegou ao $\lambda = 0{,}0296$ que estabilizar a estimativa dos parâmetros e a equação proposta é a seguinte:

$$\hat{P}_i = 2.5396 - 18.2920\, CAB_i + 37.0162\, CAP_i + 2751.7057\, CAB^2H_i; i = 1 \ldots n$$

A estimativa de σ^2 é dada por S² = 76115,08, enquanto que o coeficiente de determinação foi de R²= 0,8723.

Na Figura 11, observa-se que uma pequena tendência dos resíduos, mas, pode-se considerar para nosso estudo que todas as suposições de regressão em relação ao erro são atendidas.

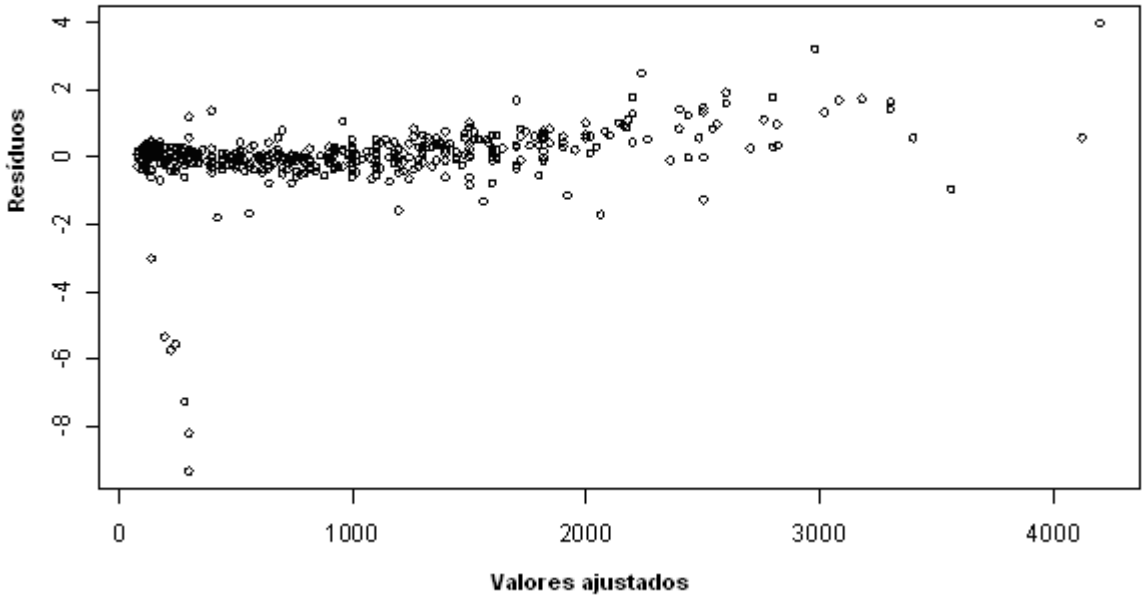

Figura 11. Gráfico dos resíduos para equação obtida pelo método de regressão Ridge.

5. CONSIDERAÇÕES FINAIS

Conclui-se, inicialmente, através da matriz de correlação e fator de inflação da variância que existe um alto grau de multicolinearidade entre as variáveis explicativas (ALT, CAB, CAP, CAB²H).

O modelo proposto pelo método de seleção de variáveis de Stepwise não obtém resultados satisfatórios para a estimativa do peso da haste do bambu, pois as variáveis explicativas que fazem parte da equação (CAB, CAP, CAB²H) são correlacionadas, ou seja, o problema de multicolinearidade faz com que as estimativas não sejam confiáveis.

A mesma situação do método de Stepwise acontece com o método de seleção de retenção de K componentes pela matriz de correlação. Para o K=2 existe o problema de colinearidade. Já para o K=3, não existe problema de colinearidade ou multicolinearidade, pois só tem uma variável explicativa (CAB). Porém esta variável não tem uma forte relação linear com a variável resposta (P), logo, as suposições dos erros de regressão linear simples não são atendidas.

O modelo proposto pelo método de seleção de variáveis de retenção de K componentes pela matriz de covariância obteve um bom resultado para o K=3 e problemas de colinearidade para K=2. Quando o K=3, modelo de Spurr (SPURR, 1952), todas as suposições de análise de regressão linear simples são válidas. O coeficiente de determinação (R^2 = 85,56%), a estimativa da variância (S^2= 85699) e as estimativas dos parâmetros são confiáveis.

Tanto o método de regressão com os componentes principais como a regressão Ridge chegaram a bons resultados e estimativas confiáveis. Porém a regressão Ridge é mais viável na prática, pela facilidade de obtenção das variáveis explicativas.

No geral, os resíduos não se comportam perfeitamente para todos os modelos, ou seja, apresenta uma pequena tendência, pela presença de outliers e podendo ter alguma influência nos resultados das estimativas.

Em termos práticos para estimativa de biomassa verde da haste do bambu para economizar tempo, custo e mão de obra, o modelo de Spurr, seria o mais indicado para os casos lineares, pois se trabalha com variáveis de fácil medição.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, M. A. **Estabilidade em análise de agrupamento**. Recife, 2005. 53f. Dissertação (Mestrado em Biometria) – UFRPE, Recife, 2005.

AZZINI, A. ; BERALDO, A. L. Determinação de fibras celulósicas e amido em cavacos laminados de três espécies de bambu gigante. **Scientia Florestalis**, Piracicaba, n. 57, p. 45-51, 2000.

AZZINI, A. ; TOMAZELLO FILHO, M. Estrutura anatômica, dimensões das fibras e densidade básica de colmos de bambusa vulgaris SCHRAD. **Revista IPEF**, Piracicaba (SP), n. 36, p. 43-50, ago. 1987. Disponível em: <http://www.ipef.br/publicações/scientia/nr36.asp. Acesso em: 25 jan. 2007.

ARAUJO, A. G. **Comparação entre métodos univariados e multivariados na seleção de variáveis independentes, na construção de tabelas volumétricas para Leucaena Leucocephala (LAM) de Wit**. Recife, 2005. Mestrado (Dissertação em Biometria) – UFRPE, Recife, 2005.

BONILLA, O. H. **Análises quantitativas da produção de bambusa vulgares Scharder ex Wendland for vulgares no estado da Paraíba**. Recife, 1991. 89 f. Dissertação (Mestrado em Biometria) – UFRPE, Recife, 1991.

BRITTO, J. O.; TOMAZELLO FILHO, M.; SALGADO, A. L. B. Produção e caracterizaçãodo carvão vegetal de espécies e variedades de bambu. **Instituto de Pesquisas Florestais – IPEF,** Piracicaba, v. 36, p. 13-17, 1997.

CADIMA, J. F. C. J. Redução de dimensionalidade através duma análise de componentes principais: um critério para o número de componentes principais a reter. **Revista de Estatística INE**, Lisboa, n. 23, p. 37-49, 2001.

CALDEIRA, M. V. W. **Determinação da biomassa e nutrientes em uma floresta ombrófila mista montana em General Carneiro, Paraná**. Curitiba, 2003. Tese (Doutorado em Engenharia Florestal) – UFPR, Curitiba, 2003.

CLARO, D. P. Uma utilização da análise multivariada, na identificação de fatores que afetam a cultura de feijão em Minas Gerais, período de 1983/93. **Cad. Adm. Rural**, Lavras, v. 10, n. 1, p. 35-44, jan./jun. 1998.

CORDEIRO, G. M. **Modelos paramétricos**. São Paulo: Associação Brasileira de Estatística, 2004.

CYSNEIROS, F. J. A.; PAULA, G. A.; GALEA, M. **Modelos simétricos aplicados**. São Paulo: ABE, 2005.

DIAS, L. T.; FARO, L.; ALBUQUERQUE, L. G. Estimativas de herdabilidade para perímetro escrotal de animais da raça nelore. **Rev. Bras. de Zootecnia**, Viçosa, v. 32, n. 6, p. 1878-1883, nov./dez. 2003. (Suple. 2).
Disponível em: <http://www.scielo.br/pdf/rbz/v32n6s2/20959.pdf>. Acesso em: 30 jan. 2007.

DRAPER, N.; SMITH, H. **Applied regression analysis**. 2. ed. New York: John Wiley & Sons, 1981.

FRANCO, E. J. **Estudo dos métodos estimativos de volume, biomassa e níveis de produtividade para eucalytptus camaldulensis**. Lavras, 1996. 100 f. Dissertação (Mestrado em Engenharia Florestal) – UFLA, Lavras, 1996.

HUSCH, B. ; MILLER, C. I. ; BEERS, T. W. **Forest mensuration**. 3. ed. New York: John Wiley & Sons, 1982.

ITAPAGÉ S/A - CELULOSE, PAPÉIS E ARTEFATOS, Informações disponíveis no site: <http://www.itpage.com/html/a_fabrica_p.htm> Acesso em: 10 fev. 2007.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 3. ed. New Jersey: Prentice Hall, 1982.

KALEY, V. **Venu Bharati, a comprehensive volume on bamboo**. Maharashtra, Índia. 2000. 189 p.

KENDALL. M. G. Factor analysis. **Journal of the Royal Statistical Society**, London, v. 12, p. 60-94, 1950.

KUMAR, M.; **Field identification key to native bamboos of Kerala**.; Kerala Forest Research institute, 38p, 2002.

LONDOÑO, X. **La Subtribu Guaduinae de América**, SIMPOSIO INTERNACIONAL GUADUA; Pereira, Colômbia, 2004.

MENDES, S. C. **Distribuição de biomassa e nutrientes em plantios comerciais de bambu (bambusa vulgaris Schard ex Wendl) no Nordeste do Brasil**. Recife, 2005. Dissertação (Mestrado em Biometria) – UFRPE, Recife, 2005.

MONTAVÃO FILHO, A.; GOMIDE, J. L.; CONDÊ, A. R. Variabilidade da constituição química e das características dimensionais das fibras de Bambusa vulgaris. **Revista Árvore**, Viçosa, v. 8, n.1, p. 12-27, 1984.

MONTGOMERY, D. C. ; PECK, E. A. ; VINING, G. G. **Introduction to linear regression analysis**. 3. ed. New York: John Wiley & Sons, 2001.

PARDÉ, J. Forest Biomass. In: **Forestry Abstracts Review Article**, France, ago 1980. Station de Sylviculture et de Producion, Centre Nacional de Recherches Forestières, v.41, n.8, p. 349; 350; 352.

PAULA, G. A. **Estimação e testes em modelos de regressão com parâmetros restritos**. São Paulo: ABE, 1997.

PAULA, G. A. **Modelos de regressão**: com apoio computacional. São Paulo: IME / USP, 2004.

SALGADO, A. L. B. ; BRITO, J. O. ; TOMAZELLO FILHO, M. Produção e caracterização do carvão vegetal de espécies e variedades de bambu. **Revista IPEF**, Piracicaba (SP), n. 36, p. 13-17, ago. 1987. Disponível em: <http://www.ipef.br/publicações/scientia/nr36.asp. Acesso em: 25 jan. 2007.

SANTOS JÚNIOR, R. C. B. **Modelagem matemática na estimativa de crescimento em altura de leucena (leucaena leucocephata LAM de Wit.), no agreste de Pernambuco**. Recife, 2005. Dissertação (Mestrado em Ciências Florestais) – UFRPE, Recife, 2005.

SILVA, H. D. **Modelos matemáticos para estimativa da biomassa e do conteúdo de nutrientes em plantações de eucalyptus grandis Hill (ex. Maiden) em diferentes idades**. Curitiba, 1996. 101 p. Tese (Doutorado em Engenharia Florestal) – UFPR, Curitiba, 1996.

SILVA, R. M. de C. e. **O bambu no Brasil e no mundo**. [Goiânia]: Embambu, 2005. Disponível em:<http://www.embambu.com.br/imagens/bambu_brasil.pdf> Acesso em: 25 jan. 2007.

SILVEY, S. D. **Statistical inference**. Penguin: Harmondsworth, 1970.

SOARES, R. V. ; HOSOKAWA, R. T. Estimativa da biomassa energética de árvores de bracatinga. **Boletim IBDF**, Brasília, n. 8, p.37-48, 1984.

SPURR, S.H. **Forest inventory**. New York, Ronald Press, 1952, 476 p.

TEIXEIRA, A. A. **Painéis de bambu para habitações econômicas: avaliação do desempenho de painéis revestidos com argamassa**. Brasília, 2006. 204 f. Dissertação (Mestrado em Arquitetura e Hurbanismo) – Universidade de Brasília. Brasília, 2006.

TEIXEIRA, L. M. **Influência da intensidade de exploração seletiva de madeira no crescimento e respiração do tecido lenhoso das árvores em uma floresta tropical de terra-firme na região de Manaus**. Manaus, 2003. 61 f. Dissertação (Mestrado) – INPA/UFAM, Manaus, 2003.

VASCONCELLOS, R. M. Bambu Brasileiro. Informações disponíveis no site <http://www.bambubrasileiro.com/info/> Acesso em: 10 fev. 2007.