ADALBERTO GOMES DE ARAÚJO

COMPARAÇÃO ENTRE MÉTODOS UNIVARIADOS E MULTIVARIADOS NA SELEÇÃO DE VARIÁVEIS INDEPENDENTES, NA CONSTRUÇÃO DE TABELAS VOLUMÉTRICAS PARA *Leucaena leucocephala* (Lam) de Wit.

RECIFE
2005

ADALBERTO GOMES DE ARAÚJO

COMPARAÇÃO ENTRE MÉTODOS UNIVARIADOS E MULTIVARIADOS NA SELEÇÃO DE VARIÁVEIS INDEPENDENTES, NA CONSTRUÇÃO DE TABELAS VOLUMÉTRICAS PARA *Leucaena leucocephala* (Lam) de Wit.

Dissertação apresentada ao Programa de Pós-Graduação em Biometria da Universidade Federal Rural de Pernambuco, como parte dos requisitos para obtenção do grau em Mestre em Biometria. Área de concentração: Métodos Estatísticos aplicados a Ciências Agrárias.

Orientador: Prof. José Antonio Aleixo da Silva, Phd
Co-Orientador: Prof. Rinaldo Luiz Caraciolo Ferreira, Dr.

RECIFE
Estado de Pernambuco – Brasil

Ficha catalográfica
Setor de Processos Técnicos da Biblioteca Central – UFRPE

A663c Araújo, Adalberto Gomes de
Comparação entre métodos univariados e multivariados na seleção de variáveis independentes, na construção de tabelas volumétricas para *Leucaena leucocephala* (Lam) de Wit. / Adalberto Gomes de Araújo – 2005.
83 f. : il., tabs.

Orientador: José Antônio Aleixo da Silva
Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernambuco. Departamento de Física e Matemática.
Inclui referências e apêndice.

CDD 311.2

1. Estatística aplicada
2. Seleção de Variáveis
3. Análise Multivariada
4. Análise Univariada
5. Leucaena leucocephala
6. Análise de Regressão
7. Métodos de Agrupamento
8. Tabela de volume

I. Araújo, Adalberto Gomes de
II. Título

Maio – 2005

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA (MESTRADO)

COMPARAÇÃO ENTRE MÉTODOS UNIVARIADOS E MULTIVARIADOS NA SELEÇÃO DE VARIÁVEIS INDEPENDENTES, NA CONSTRUÇÃO DE TABELAS VOLUMÉTRICAS PARA *Leucaena leucocephala* (Lam) de Wit.

ADALBERTO GOMES DE ARAÚJO

Banca Examinadora:

Orientador: _____
Prof. Dr. José Antonio Aleixo da Silva, Phd

Examinadores: _____
Prof[a]. Dra. Jacira Guiro Marino

_____
Prof. Dr. Paulo de Paula Mendes

_____
Prof. Dr. Rinaldo Luiz Caraciolo Ferreira

RECIFE
2005

SUMÁRIO

LISTA DE TABELAS

ARAÚJO, ADALBERTO GOMES DE. Comparação entre métodos univariados e multivariados na seleção de variáveis independentes, na construção de tabelas volumétricas para *Leucaena leucocephala* (Lam) de Wit. 2005. Orientador: Prof. Dr. José Antônio Aleixo da Silva. Co-orientador: Prof. Dr. Rinaldo Luiz Caraciolo Ferreira.

RESUMO

O objetivo deste trabalho foi utilizar métodos estatísticos univariados e multivariados na seleção de variáveis independentes, em modelos matemáticos, para a construção de tabelas de volumes para *Leucaena leucocephala,* visando reduzir tempo e custos sem perda de precisão. Os dados foram provenientes de um experimento conduzido na Estação Experimental da Empresa Pernambucana de Pesquisa Agropecuária (IPA), Caruaru-PE. Foram utilizadas 201 árvores de leucena, que tiveram seus volumes cubados pelo método de Smalian, e 20 variáveis independentes medidas nas mesmas árvores. Para a seleção das variáveis independentes foram utilizados os seguintes métodos: Componentes Principais, Análise de Agrupamento, R^2 Máximo e Mínimo, Stepwise, Forward, Backward e Critério de Akaike. No geral, os métodos univariados e multivariados empregados no descarte de variáveis independentes para modelos volumétricos, conduzem a respostas semelhantes, mesmo que possuam estruturas diferentes em relação às variáveis independentes, desde que o número dessas variáveis seja elevado. Além dos testes estatísticos aplicados, o julgamento do pesquisador sobre a relevância das variáveis selecionadas nas equações resultantes, é de grande importância, principalmente, na redução de custos e do erro de amostragem.

ARAÚJO, ADALBERTO GOMES DE. Comparison among univariate and multivariate methods in the selection of independent variables, in the construction of volume tables for *Leucaena leucocephala* (Lam) de Wit. 2005. Adviser: Prof. Dr. José Antônio Aleixo da Silva. Co-adviser: Prof. Dr. Rinaldo Luiz Caraciolo Ferreira.

ABSTRACT

The objective of this work was to use multivariate and univariate statistical methods, in the selection of independent variables, in mathematical models, in the construction of volume tables for *Leucaena leucocephala*, looking for reduction in time and costs, without loss of precision. The data came from an experiment carried out at the Experimental Station of the Institute of Agriculture Research (IPA), Caruaru-PE. It was used 201 trees of leucena that had their volumes (dependent variable) measured by the method of Smalian, and 20 variables independent measured in the same trees. For the selection of the independent variables the following methods were used: Principal Components, Cluster Analysis, Maximum and Minimum R^2, Stepwise, Forward, Backward and Criterion of Akaike. In the general, the univariate and multivariate methods used in the selection of independent variables for volume models, showed similar responses, even though they had different structures in relation to the independent variables, since the number of those variables is high. Besides the applied statistical tests, the researcher's judgment about the relevance of the selected independent variables in the final equations has a great importance, mainly, in the reduction of costs and sampling errors.

1. INTRODUÇÃO

Nenhum país pode prescindir dos recursos oriundos de suas florestas para seu desenvolvimento econômico e social. As regiões das Caatingas e do Cerrado precisam ser consideradas potencialmente produtivas em termos florestais. Péllico Netto e Brena (1997) complementam afirmando que uma grande parte dessas áreas é de vocação florestal e sua utilização através da implantação de reflorestamentos, abrirá nova perspectiva regional.

Na região Nordeste, a vegetação natural de caatinga tem sido explorada de forma desordenada. Há muito vem se alertando para a necessidade de proteção desse patrimônio e conscientização da população da região para esse fato, com finalidade de assegurar sua própria condição de sobrevivência. A lenha e o carvão vegetal continuam sendo as formas mais importantes de utilização dos recursos florestais nessa região.

Campello et al. (1999) afirmam que há uma grande dependência da população e dos demais setores da economia em relação ao produto florestal como fonte de energia, e que este representa 30% a 50% da energia primária da região Nordeste.

O reflorestamento em pequenas e médias propriedades rurais é de interesse público por ser: uma fonte de renda, contribuir para evitar o êxodo rural, o desemprego e, simultaneamente, possibilitar inúmeros e imprescindíveis benefícios ambientais (GALVÃO, 2000).

Uma silvicultura diversificada e consorciada, direcionada mais para o social e ecológico do que para o econômico, é o que sugerem Drumond e Couto (1994) para a região semi-árida. Nessas áreas, os sistemas agroflorestais constituem importante alternativa para contribuir com o aumento da capacidade produtiva e reabilitação de áreas degradadas, bem como para incluir áreas mais frágeis no sistema produtivo rural (MEDRADO, 2000).

A estrutura dos sistemas agroflorestais viabiliza os princípios do manejo sustentado dos ecossistemas, conforme alegação de Macedo e Camargo (1994), principalmente, através da utilização de espécies de usos múltiplos. Lima (1986) também indica essa solução, como uma contribuição na oferta de alimentos e na geração de energia, em resposta à demanda de madeira e à necessidade de aumento de forragem para os animais nessa região.

Numerosos estudos indicam as leguminosas florestais como espécies potenciais, e dentre elas, a *Leucaena leucocephala* (Lam.) de Wit, vem se destacando por apresentar persistência produtiva, capacidade de rebrota, tolerância a baixas precipitações, ausência de pragas e doenças (DIAS FILHO e SERRÃO, 1982; SOUZA, 1999), boa sobrevivência e ótimo crescimento em altura (NÓBREGA, 1978; CARVALHO, 1978; LIMA, 1978; EMBRAPA, 1981; PIRES e FERREIRA, 1982; LIMA, 1982; FRANCO e SOUTO, 1986).

Informações sobre o crescimento e a produção de espécies e povoamentos florestais são imprescindíveis no planejamento e manejo florestal. Para tanto, os inventários constituem instrumentos fundamentais, assim como as tabelas de volume e de produção, que têm a função de estimar volumes de árvores individuais e/ou povoamentos através da mensuração de variáveis das árvores fáceis de serem mensuradas. Portanto, essas tabelas constituem um instrumento indispensável ao profissional florestal na execução de trabalhos relativos ao crescimento e produtividade de povoamentos florestais.

Ao se analisar um experimento ou dados provenientes de uma amostra, o número de variáveis mensuradas, às vezes, é superior ao que se pode representar ou modelar o fenômeno sem perda de precisão. O inverso também pode ocorrer e nesses casos se perdem valiosas informações. Portanto, cabe ao pesquisador usar técnicas univariadas e multivariadas para fazer inferências por meio de um número ideal de variáveis, sem perda de precisão e a baixos custos, uma vez que a melhor equação é aquela que produz estimativas precisas envolvendo um número mínimo de variáveis independentes, facilmente mensuráveis.

Na prática, a seleção de variáveis, freqüentemente, combina o conhecimento sobre relevância delas, o julgamento subjetivo do pesquisador e, ainda, procedimentos estatísticos (SILVA, 2001).

Em quaisquer áreas de estudo, variáveis que assumem, praticamente, o mesmo valor para todos os objetos são pouco discriminatórias, e suas inclusões pouco contribuirão para o conhecimento e solução do problema em questão. Por outro lado, a inclusão de variáveis com grande poder de discriminação, porém irrelevante ao problema pode mascarar o estudo e levar a resultados equivocados. (JOLLIFFE, 1972).

O que se procura no estudo de seleção de variáveis é determinar uma estrutura de correlação em um subconjunto de variáveis que capture a maior variação do conjunto original sem comprometer a interpretação do problema estudado.

Segundo Jolliffe (1972), em análises multivariadas quando um grande número de variáveis (>10) é avaliado, os resultados mudam muito pouco quando um subconjunto de variáveis é utilizado, pois, algumas variáveis são redundantes e podem ser descartadas da análise. Com menos variáveis a serem analisadas não se desperdiça tempo, tanto na tomada de medidas como nas análises computacionais, e, conseqüentemente, sobre os custos em análises futuras.

Um dos primeiros trabalhos a comparar métodos multivariados na seleção de variáveis independentes foi o de Beale et al. (1967). Nesse trabalho, os autores citam como potenciais as análises de:

- REGRESSÃO: Consiste em descartar qualquer variável que adicione pouco a acurácia com qual a equação de regressão se correlaciona com a variável dependente.

- INTERDEPENDÊNCIA: Dentre um conjunto de p dimensões possíveis que se colapsam, exatamente ou aproximadamente, procura-se reduzir o conjunto em poucas dimensões, através da seleção de variáveis.

- DISCRIMINANTE: Descarta-se dentre um conjunto de p variáveis, aquelas que não comprometem, seriamente, o poder de discriminação do conjunto.

- AGRUPAMENTO: Identificam-se grupos de objetos similares.

Jolliffe (1972, 1973) investigou vários métodos de seleção de variáveis a partir da análise de componentes principais. McCabe (1984) também, baseado no mesmo processo, propôs critérios para seleção de variáveis, que são conhecidos como “Critério de Seleção de Variáveis“ (Variable Selection Criteria - VSC).

Beale et al (1967) e Jolliffe (1972), obtiveram a redução do número de variáveis independentes trabalhando com dados artificiais e reais.

Segundo La Chenbruc (1975), a análise discriminante também pode ser utilizada como teste de suficiência na seleção de variáveis. Exemplos de utilização dessa técnica podem ser observados em Mckay e Campbell (1982) e Le Roux et al (1997).

Concomitantemente, a análise de dados através do emprego de regressão linear é uma das técnicas de estimação mais usadas.

Segundo Cordeiro e Neto (2004) o modelo clássico de regressão teve origem nos trabalhos de Astronomia elaborados por Gauss, no período de 1809 a 1821. É a técnica mais adequada quando se deseja estudar o comportamento de uma variável dependente (variável resposta) em relação a outras variáveis independentes (variáveis explicativas) que são responsáveis pela variabilidade da variável resposta.

Depois do ajustamento preliminar de um modelo de regressão, devem-se selecionar as variáveis explicativas que podem ser eliminadas do modelo, objetivando obter uma equação para explicar os dados em questão com poucas variáveis e boa precisão.

Para tal, o teste F da análise de variância permite apenas inferir que algumas das variáveis explicativas são realmente importantes para explicar a variabilidade da variável resposta. Para selecionarmos as variáveis independentes que são significativas, precisa-se determinar a distribuição das estimativas dos parâmetros através de outros procedimentos estatísticos.

Desta forma, o presente trabalho objetiva comparar métodos multivariados e univariados, na seleção de variáveis independentes em modelos matemáticos, na construção de tabelas de volumes para *Leucaena leucocephala,* visando reduzir tempo e custos sem perda de precisão, através da seleção de variáveis independentes.

2. REVISÃO DE LITERATURA

2.1 Leucaena leucocephala (Lam.) de Wit.

O gênero *Leucaena* Bentham é classificado como pertencente à sub-família *Mimosoidae* da família Leguminosas (BARROSO, 1984; NFTA, 1985).

Nativo das Américas, dispersando-se naturalmente do Peru ao Texas, o local de sua origem é desconhecido, mas se presume que o mais antigo centro de dispersão seja a península de Yucatan, no México (NFTA, 1985; BREWBAKER, 1989; ALCÂNTARA, 1993).

Supõe-se que, inicialmente, a leucena tenha sido levada para a região do Pacífico, após a conquista do México, nos séculos 16 e 17 e, posteriormente, com a ocupação espanhola das Filipinas e da Indonésia (FREITAS et al., 1991). Amplamente, cultivada no México, Filipinas, Hawai, África e Austrália (CUNHA, 1979), hoje pode ser encontrada em quase todas as regiões tropicais (NFTA, 1985).

Brewbaker e Ito (1980) consideravam apenas 10 espécies de Leucaena como de validade inquestionável. Mais tarde, Brewbaker (1985), num trabalho de revisão na sistemática do gênero, incluiu as espécies *L. greggii, L. Watson, L. pallida* Britton e Rose, e se referiu a uma nova espécie *L. cuspidata* Standley descrita em 1984, mas não amplamente reconhecida e por esse motivo não incluída na chave sistemática que propôs.

Atualmente, há 13 espécies do gênero *Leucaena* (BREWBAKER, 1989) e embora muitas espécies apresentem valor forrageiro nas regiões tropicais, somente a *Leucaena leucocephala* conhecida popularmente como leucena, tem sido mais explorada (FREITAS et al., 1991).

O gênero *Leucaena* forma simbiose eficiente com bactérias fixadoras de nitrogênio atmosférico, atingindo valores de N_2 fixado de 598 Kg/ha/ano (Saginga et al., apud FRANCO E SOUTO, 1986). Segundo Seiffert (1984), mobiliza no primeiro ano, na parte aérea de planta, cerca de 370 Kg/ha de N, representando uma excelente fonte de proteína para rebanhos, ao acumular, aproximadamente, 2 t/ha/ano de proteína bruta.

Além da simbiose com bactérias do gênero *Rhizobium*, as quais fixam até 400 kg/ha/ano de N, Kluthcouski (1982) ressalta a associação com fungos do

gênero *Mycorrhizae* que viabilizaram a utilização do fósforo não disponível para a maioria das culturas.

Golfari e Caser (1977), em seu zoneamento ecológico para a região Nordeste, recomendam a leucena como apta à experimentação nas regiões correspondentes aos tipos climáticos: sub-úmido seco, semi-árido e árido. A Superintendência do Desenvolvimento do Nordeste (SUDENE, 1972) e Pires e Ferreira (1982), indicam-na com grande potencial para reflorestamento.

A viabilidade do cultivo da *Leucaena leucocephala* na região de Petrolina, ao se comparar seu desenvolvimento com *Prosopis juliflora* e *Eucalyptus alba*, espécies recomendadas para regiões semi-áridas, é atestada por Lima (1982), que encontrou para essa espécie incremento médio anual em volume superior às várias procedências de *Eucalyptus alba* testadas.

A importância do uso da leucena como forrageira também é evidenciada por diversos pesquisadores, cujos trabalhos tratam do estabelecimento e manejo da cultura, produtividade, manejo do banco de proteínas, palatabilidade, valor nutritivo, toxidez da planta pela mimosina, entre outras características (RIBEIRO, 2001).

2.2 Tabela de Volume

Para se cubar um povoamento, seria necessário a determinação do volume individual de cada árvore, em função do diâmetro, da altura e da forma, o que é impraticável. A conveniência do uso da tabela de volume está no fato de existir grande dificuldade em se obter o volume de uma árvore, e como as variáveis requeridas para entrada são de fácil obtenção, a estimativa do volume é simplificada (ASSOCIAÇÃO PARANAENSE DE ENGENHEIROS FLORESTAIS, 1987).

Uma tabela de volume para árvores individuais pode ser definida como uma relação gráfica ou numérica expressa por modelos lineares e não lineares, capaz de exprimir o volume total ou parcial de uma árvore em função de variáveis independentes tais como diâmetro, altura, fator de forma ou outras. Também pode ser definida como a representação tabular do volume individual de árvores inteiras ou partes delas, através de variáveis fáceis de mensurar (FINGER, 1992).

Presume-se que a mais antiga tabela de volume tenha sido construída na segunda metade do século XVIII, enquanto caberia a Henrich Cotta a publicação, em 1804 e 1817, das primeiras tabelas modernas (SPURR, 1952).

São citadas outras tabelas antigas, de modo geral construídas a partir de elevado número de árvores, algumas com mais de 40.000. A construção era realizada a partir de métodos gráficos, facilitada pela amplitude da amostra (ASSOCIAÇÃO PARANAENSE DE ENGENHEIROS FLORESTAIS, 1987). Seu uso no Brasil, conforme relata Machado (1979), teria iniciado em meados da década de 1960.

As tabelas de volume podem ser construídas através de métodos gráficos ou através de métodos analíticos. O método gráfico por envolver subjetividade é menos recomendado.

Veiga (1973), ao analisar a literatura relacionada ao assunto, comenta que os estudos destinados à elaboração de tabelas volumétricas devem ser dirigidos, preferencialmente, a processos analíticos.

No método analítico, necessita-se usar um modelo matemático, previamente escolhido entre os muitos já conhecidos, ou então testar vários deles e escolher a equação que apresentou melhor ajuste e precisão.

Com o advento do conhecimento da análise de regressão e, posteriormente, com o aparecimento de computadores, o método analítico passou a ser o único utilizado para a construção de tabelas de volume. Este método apresenta, além da maior precisão e facilidade de cálculo, a vantagem de não ser subjetivo, visto que se utiliza a análise de regressão para ajuste dos modelos matemáticos (FINGER, 1992).

Os coeficientes das equações volumétricas resultantes, que são estimados através de análise de regressão linear, baseiam-se no princípio dos mínimos quadrados ou método da máxima verossimilhança (ASSOCIAÇÃO PARANAENSE DE ENGENHEIROS FLORESTAIS, 1987; FINGER, 1992; SCOLFORO, 1993). Nos caso dos modelos intrinsecamente não lineares, usam-se técnicas de análise numérica (SILVA e SILVA, 1982).

Após a amostragem e cubagem rigorosa de um número representativo de árvores, do ajuste de vários modelos volumétricos e da seleção da equação mais adequada, constrói-se a tabela de volume para a amplitude dos dados observados.

Os volumes estimados não são exatos, pois as variáveis independentes são obtidas em uma série de indivíduos medidos no povoamento, que estão sujeitos às variações naturais não controladas, bem como erros de mensuração. Desta forma, deve-se admitir que as relações volumétricas possibilitem a estimativa de volumes médios em torno dos quais devem se distribuir os volumes verdadeiros (FINGER, 1992; SILVA et al., 1993).

2.3 Trabalhos realizados com o mesmo experimento de leucena

O primeiro trabalho realizado no experimento que deu origem a esta dissertação recebeu o título de "Crescimento de mudas de *Leucaena leucocephala* (Lam) de Wit, em função do uso de composto de resíduo urbano, adubação fosfatada e inoculação com *Rhizobium loti* (MEUNIER, 1991). Avaliou-se o comportamento das mudas de leucena em relação a diferentes fontes de adubação, sendo que o composto orgânico de resíduo urbano foi o mais eficiente.

Os dados do presente trabalho foram coletados para a dissertação de mestrado de Ribeiro (2001), tendo como título "Seleção de Modelos volumétricos para leucena no Agreste do Estado de Pernambuco". Modelos matemáticos lineares e não-lineares foram avaliados para selecionar uma equação de volume a ser utilizada na construção de tabela de volume total (fuste e galhos).

Os modelos da variável combinada de Spurr e o de Schumacher-Hall foram usados como padrão de comparação com os modelos propostos, nos quais foram usadas combinações das variáveis independentes: CAP (circunferência à altura do peito: 1,30m), H (altura total), NG (número de galhos), CG (circunferência na base do galho maior) e volumes de segmentos do tronco, com comprimentos e áreas seccionais variáveis. Com base nos critérios adotados para avaliação comparativa, foi considerado, como mais adequado para construção de tabelas volumétricas para leucena, o modelo de regressão linear simples, cuja variável independente é o volume da secção V1 (0,30m a 0,90m), cuja equação resultante apresentou um coeficiente de determinação (R^2) de 0,9477 (RIBEIRO, 2001).

Souza (2003) analisou o crescimento da leucena em sua dissertação intitulada: "Avaliação do crescimento em altura de leucena *Leucaena leucocephala* (Lam.) de Wit., no Agreste de Pernambuco", por meio da análise multivariada de medidas repetidas, e encontrou que as árvores atingiram o

crescimento máximo aos 4 anos de idade, verificando também que, a partir dessa idade, os efeitos dos tratamentos aplicados em 1989 não diferiam entre si

Júnior (2005) aplicou diferentes modelos matemáticos de crescimento na análise do crescimento de leucena e encontrou que os melhores resultados obtidos foram com os modelos de Brody e Silva-Bailey.

2.4. Análise de regressão linear

Cordeiro e Paula (1989) inferem que a análise de dados utilizando técnicas de regressão linear é a mais usada nos processos estimativos.

O uso de técnicas de regressão em inventários florestais possui um vasto espectro de aplicação, tais como: industriais, comerciais, exploratórios, fixação de carbono, índice de sítio, produção de biomassa, estudos de crescimento e mortalidade, dinâmica de populações, etc (SILVA, 1986; CLUTTER, 1989; FIERROS-GONZALEZ et al., 1992; RIBEIRO, 2001; SCHNEIDER e TONINI, 2003; JUNIOR, 2005).

Os parâmetros da equação de regressão são estimados a partir das inter-relações da variável dependente (resposta) e das variáveis independentes (preditoras).

Entretanto, para se realizar análises de regressão, existem alguns requisitos que devem ser verificados (LEWIS-BECK, 1980):

- A variável dependente e as independentes devem ser mensuradas com precisão;
- Para cada observação o valor esperado do erro seja zero, $E(\xi_i)= 0$;
- A variância do erro é constante, $Var(\xi_i) = E(\xi_i^2)= \sigma^2$;
- As variáveis independentes X_i não possuam erros correlacionados $E(\xi_t,\xi_s)=0$;
- Que os erros ξ_i sejam normalmente distribuídos.

Este último requisito é importante para a realização de testes de hipóteses e obtenção de intervalos de confiança, pois a confiabilidade da equação estimada dependerá da validade desse conjunto de restrições.

Segundo Neter et al. (1989), na construção de modelos de regressão, deve-se evitar usar variável(is) que:

- Não seja(m) fundamental(is) para o problema;
- Estar(em) sujeita(s) a grandes erros de medidas;
- Estar(em), efetivamente, duplicada(s) por uma outra na lista de variáveis possíveis.

A priori esses critérios são julgados, arbitrariamente, pelo pesquisador. Após essa primeira seleção, restam outras variáveis que são potencias para participarem da equação selecionada. Neste momento, processos de busca do melhor conjunto de variáveis, baseados em critérios estatísticos, são de grande ajuda ao pesquisador.

Por outro lado, quando se usam muitas variáveis independentes, geralmente, estas podem apresentar erros correlacionados, gerando o problema da multicolinearidade, que significa que existem relações de dependência linear aproximada entre as variáveis independentes, implicando que o det (X'X) seja igual ou muito próximo de zero.

Sintoma típico dessa situação é a ocorrência de valores não significativos para a estatística de Student associados a cada um dos coeficientes da regressão, apesar de mostrar alto valor para o coeficiente de determinação (SOUZA, 2001).

2.5. Métodos de seleção de variáveis usados.

- **Análise de Componentes Principais**
  a) Método de seleção de variáveis por retenção de k componentes
  b) Método de Beale
  c) Método de Beale interativo
  d) Método de seleção por coeficiente de correlação múltipla
- **Analise de Agrupamento**
  a) Medida de Dissimilaridade
     a.1) Distância Euclidiana
     a.2) Distância Generalizada de Mahalanobis

- **Métodos de agrupamento**
  a) Métodos hierárquicos
     a.1) Método do vizinho mais próximo
     a.2) Método de otimização
  b) Métodos de seleção de variáveis
     b.1) Método de ligação máxima
     b.2) Método de ligação média
- **R^2 múltiplo**
- **R^2 ajustado**
- **Erro médio quadrático**
- **Critério Cp de Mallows**
- **Busca direta de t**
- **Incremento de R^2 máximo e mínimo**
- **Stepwise**
- **Forward**
- **Backward**
- **Critério de Akaike**
- **Todas as possibilidades**

2.5.1 Análise de Componentes Principais

A análise dos componentes principais fornece um método multivariado para a redução de variáveis em casos em que ocorrem multicolinearidade nas variáveis independentes e pode revelar relações que não são observadas, previamente nos subconjuntos de variáveis, permitindo interpretações que, ordinariamente, não apareceriam (JOHNSON & WINCHERN, 1982; SOUZA, 2001).

Os componentes principais são combinações lineares de variáveis, construídas de maneiras a captar o máximo da variância, em que o primeiro componente explica a maior variação existente, o segundo componente explica a segunda maior variação e assim, sucessivamente. A técnica consiste na transformação de um conjunto de n variáveis padronizadas, X_{i1}, X_{i2}, ..., X_{in} em um

novo conjunto Y_{i1}, Y_{i2}, ..., Y_{in}, em que os Y_{is} são funções lineares dos X_{is} e independentes entre si.

Claro et al. (1998) utilizou a análise de componentes principais em um estudo da identificação de fatores que afetam a cultura do feijão em Minas Gerais. Ferreira et al. (2003) utilizou essa técnica na redução de variáveis, na avaliação da divergência genética entre clones de palma (*Opuntia fícus indica,* Mill).

Nos componentes principais as seguintes propriedades são verificadas:

a) Se Y_{ij} é um componente principal então:

$Y_{ij} = a_1X_{i1} + a_2X_{i2} + ... + a_nX_{in}$

b) Se $Y_{ij'}$ é outro componente principal então:

$Y_{ij'} = b_1X_{i1} + b_2X_{i2} + ... + b_nX_{in}$

$\sum_{j=1}^{n} a^2_j = \sum_{j=1}^{n} b^2_j = 1$; $\sum_{j=1}^{n} a_jb_j = 0$, ou seja, os componentes são independentes.

c) Dentre todos os componentes, Y_{i1} apresenta a maior variância, Y_{i2} a segunda maior e assim sucessivamente.

Os componentes principais são obtidos pela solução do sistema:

$\det (R-\lambda_1 I) a = 0,$

em que: R = matriz de correlação entre as médias estimadas;

λ_1 = raízes características ou (autovalores) de R;

I = matriz identidade de dimensão pxp;

a = vetor característico (ou autovetor) associado aos autovalores (HOFFMANN, 1999).

Desta forma, os autovalores de R correspondem às variâncias de cada componente e ao autovetores normalizados correspondem aos coeficientes de ponderação dos caracteres padronizados.

A importância relativa de um componente, que é avaliada pela porcentagem da variação total que ele explica, é expressa por:

Importância de $Y_j = \frac{\lambda_j}{\text{traço}(R)}$

a) **Método de seleção de variáveis por retenção de k componentes**

Esse método considera o autovetor correspondente ao menor autovalor e rejeita a variável com maior coeficiente (em valor absoluto) e em seguida considera o próximo autovalor e repete o processo que termina quando resulta em k componentes, em que k é arbitrário. Nesse método se deve excluir todas os p-k componentes de menor importância, sendo que a variável, cuja participação no componente é mais elevada em valor absoluto, deverá ser excluída, caso já não tenha sido anteriormente. Quando isso ocorre, a variável que ainda não foi eliminada em nenhum dos componentes anteriores e que tiver maior participação deverá ser excluída. O processo de exclusão inicia em ordem inversa de importância dos componentes (CADIMA, 2001)

b) Método de Beale

Nesse método a determinação de k não é subjetiva. Sendo λ_0 um valor limite para a inclusão de variáveis, a exclusão, semelhante ao método anterior, se dará para todas componentes tais que $\lambda_i < \lambda_0$, que segundo Jolliffe (1972) o valor de λ_0 deve ser 0,70.

Após a exclusão das componentes que satisfazem o requisito acima, o restante do procedimento é igual ao método de retenção de k componentes.

Também pode se considerar que ao invés de usar λ_0, utilize-se k como sendo o número de variáveis cuja proporção da variabilidade acumulada seja α_0=0,8, embora o método que usa λ_0 tenha mostrado melhor desempenho (JOLLIFFE, 1972).

c) Método de Beale interativo

Agora a seleção de variáveis independentes que se baseia na inspeção dos autovalores e dos componentes dos autovetores, como se segue:

a) Uma análise de componentes principais é feita sobre todas as P variáveis originais, e os autovalores são inspecionados;

b) Então, se p_1 autovalores são menores que algum valor λ_0, o autovetor correspondente, ou seja, os componentes em si são considerados a entrar na análise, começando com o componente correspondendo ao menor autovalor, então o componente correspondendo ao segundo menor autovalor e assim por diante;

c) Uma variável é então associada com cada um destes p_1 componentes, isto é, a variável que tem o maior coeficiente nos componentes sob consideração e que ainda não tenha sido associado com um componente previamente considerado. As p_1 variáveis associadas com os p_1 componentes considerados são então rejeitadas;

d) No próximo passo, outra análise de componentes é feita sobre as $P–p_1$ variáveis. Novamente, se qualquer um dos autovalores é menor que λ_0 uma variável é associada com cada um dos componentes correspondentes, e estas p_2 variáveis são rejeitadas;

e) Uma análise de componentes é então feita nas $Pp_1 – p_2$ variáveis, e esse procedimento continua até que todos os autovalores na última análise de componente serem maiores que λ_0. Neste estágio o procedimento para, tendo reduzido o número de variáveis de P para
$P – p_1 –p_2 –... – p_i = p$;

f) O valor de p será determinado pela escolha de λ_0.

d) Método de seleção por coeficiente de correlação múltipla

Esse método é baseado na matriz de correlação em que se deve fixar o número k de variáveis a serem consideradas e obter os coeficientes de correlações múltiplas (R^2) para cada um dos C_p^k modelos, em que k é o número total de variáveis independentes e p o número de parâmetros testados no modelo. O escolhido será aquele que proporcionar o maior R^2.

2.5.2. Análise de Agrupamento

A análise de agrupamento tem por finalidade reunir, por algum critério de classificação, os indivíduos (ou qualquer outro tipo de unidade amostral) em vários grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos. Alternativamente, as técnicas de análise de agrupamento têm por objetivo, ainda, dividir um grupo original de observações em vários grupos, segundo algum critério de similaridade ou dissimilaridade (CRUZ e REGAZZI, 2001).

Na análise de agrupamento, várias questões surgem. Assim, questiona-se o número final de grupos desejados, a adequação da partição obtida e o tipo de medida de similaridade a ser utilizada. Com relação ao número de grupos desejados, o que se faz mais comumente, é utilizar vários números de grupos e, por algum critério de otimização, selecionar o mais conveniente. Para avaliação da adequação da partição, é comum a utilização da análise discriminante e, com relação às medidas de similaridade, várias são citadas, entretanto, as mais empregadas são as distâncias Euclidianas e a de Mahalanobis (CRUZ e REGAZZI, 2001).

O processo de agrupamento envolve, basicamente, duas etapas. A primeira se relaciona com a estimação de uma medida de similaridade (ou dissimilaridade) entre os indivíduos e a segunda, com a adoção de uma técnica de agrupamento para a formação dos grupos.

a) Medida de Dissimilaridade

Dentro das medidas dissimilaridades o uso mais rotineiro é a distância Euclidiana média ou a distância generalizada de Mahalanobis, sendo esta última a preferida. Entretanto, só é possível de ser estimada quando se dispõe da matriz de covariâncias residuais a partir de ensaios experimentais.

a.1) Distância Euclidiana

Sendo X_{ij} a observação no i-ésimo Individuo (i = 1,2,...,p), em referência ao j-ésimo caráter (j = 1,2,...,n) estudado, define-se a distância Euclidiana entre dois indivíduos i e i' por meio da expressão:

$$D_{ii'} = \sqrt{\sum_{j=1}^{n} (x_{ij} - x_{i'j})^2}$$

Apesar da distância Euclidiana média padronizada contornar os problemas inerentes ao número e à escala dos caracteres avaliados, ela apresenta o inconveniente de não levar em considerações as correlações residuais entre os caracteres disponíveis.

a.2) Distância Generalizada de Mahalanobis

A distância generalizada de Mahalanobis, denominada D_{ij}^2, além de ponderar cada uma das componentes, também considera o grau de correlação entre elas.

Segundo Reis, citado por Albuquerque (2005), a Distância Generalizada de Mahalanobis também pode ser usada como técnica de comparação na separação entre diversos grupos, permitindo avaliar a extensão e a direção dos afastamentos entre os valores médios das variáveis usadas na discriminação.

A Distância Generalizada de Mahalanobis é expressa por:

$$D_{ij}^2 = (\underset{\sim}{X}_i - \underset{\sim}{X}_j)' \sum^{-1} (\underset{\sim}{X}_i - \underset{\sim}{X}_j)$$

Em que:

$\underset{\sim}{X}_i$ = vetor de médias do i'ésimo grupo;

$\underset{\sim}{X}_j$ = vetor de médias do j'ésimo grupo;

$\sum^{-1}$ = estimativa combinada da matriz da covariância/variância dentro dos grupos.

2.5.3. Métodos de Agrupamentos

Como no processo de agrupamento é desejável ter informações relativas a cada par de indivíduos, o número de estimativas de medidas de dissimilaridade é relativamente grande, o que torna impraticável o reconhecimento de grupos homogêneos pelo simples visual daquelas estimativas. Para realizar esta tarefa, faz-se uso dos métodos de agrupamento.

Existe grande número de métodos de agrupamento disponível, dos quais o pesquisador tem que decidir qual o mais adequado ao seu trabalho, uma vez que as variedades técnicas podem levar a diferentes padrões de agrupamentos (CRUZ e REGAZZI, 2001).

Dentre os métodos de agrupamento mais comumente utilizados, citam-se os hierárquicos e os de otimização. A descrição destes métodos é apresentada a seguir:

a) Métodos hierárquicos

Nos métodos hierárquicos, os indivíduos são agrupados por um processo que se repete em vários níveis até que seja estabelecido o dendrograma ou o diagrama de árvore. Nesse caso, não há preocupação com o número ótimo de grupos, uma vez que o interesse maior está na "árvore" e nas ramificações que são obtidas. As delimitações podem ser estabelecidas por um exame visual do

dendrograma em que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de indivíduos para determinado grupo.

Os métodos hierárquicos são também divididos em métodos aglomerativos e divisivos. Dentre os métodos aglomerados, citam-se o do vizinho mais próximo ("Sinlge Linkage Method"); o ponderado ou não; o do centróide, também ponderado ou não; e o proposto por Ward (1963). Dentre os métodos divisivos, o mais conhecido é o de Edwards e Cavalli-Sforza (1965).

Serão apresentadas considerações apenas a respeito do método do vizinho mais próximo; as informações sobre os demais são apresentadas por Sneath e Sokal (1973).

a.1) Método do Vizinho mais Próximo

Método em que se identificam na matriz de dissimilaridade, indivíduos mais similares, os quais são reunidos, formando o grupo inicial. Calculam-se as distâncias daquele grupo em relação aos demais indivíduos e, nos estádios mais avançados, em relação a outros grupos já formados.

O processo de identificação das entidades (grupos) mais similares se repete sobre a nova matriz de dissimilaridade, cuja dimensão é reduzida a cada passo e finaliza quando todos os indivíduos são reunidos em um único grupo.

A distância entre um indivíduo K e um grupo formado pelos indivíduos i e j é dada por:

$$d_{(ij)k} = \min \{d_{ij}, d_{jk}\}$$

isto é, $d_{(ij)k}$ é dada pelo menor elemento do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k).

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \min\{d_{ik}, d_{il}, d_{jk}, d_{jl}\}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre pares de indivíduos (j e k), (i e l), (j e k) e (j e l).

a.2) Métodos de otimização

Nos métodos de otimização se realiza a partição do conjunto de indivíduos em subgrupos não-vazios e, mutuamente exclusivos, por meio da maximização ou minimização de alguma média preestabelecida. Um dos métodos de otimização mais comumente empregados é o proposto por Tocher (RAO, 1952).

Esse é um método de otimização que adota o critério de que a média das medidas de dissimilaridade (distância euclidiana, distância euclidiana média. D^2 de Mahalanobis, etc.) intragrupo deve ser menor do que as distâncias médias intergrupos.

Inicialmente, a partir de uma matriz de distância é identificado o par de indivíduos mais similar, os quais constituíram o grupo inicial. Em seguida, é avaliada a possibilidade de inclusão de novos indivíduos no grupo inicial, adotando-se o critério anteriormente citado.

A inclusão de um novo indivíduo provoca aumento no valor médio da distância intragrupo. Assim, decide-se pela inclusão de um determinado indivíduo num grupo se o acréscimo resultante no valor médio da distância intragrupo não ultrapassar um nível máximo predefinido, que pode ser estabelecido de forma arbitrária, ou se adota o valor máximo da medida de dissimilaridade encontrado no conjunto das menores distâncias que envolvem cada indivíduo ((CRUZ e REGAZZI, 2001).

b) Seleção de variáveis

Segundo Jolliffe (1972), os métodos de análise de agrupamento poderiam ser usados para reduzir o número de variáveis independentes. As K variáveis seriam colocadas em p grupos ou conjuntos, e uma variável seria selecionada de cada grupo. As K – p variáveis restantes seriam então rejeitadas.

Jolliffe (1972) sugeriu dois métodos hierárquicos de seleção de variáveis independentes baseados em análise de agrupamento, ambas utilizando a matriz de correlação entre as variáveis. Ambos os métodos seguiram o mesmo conjunto de passos:

a) Definir uma medida de similaridade, r_{xy}, entre quaisquer dois grupos de variáveis x e y, uma única variável é um caso especial de um grupo;

b) Calcular r_{xy} para cada um dos $\frac{1}{2}K\ (K - 1)$ pares de grupos de variáveis únicas;

c) Se A e B são os dois grupos para os quais r_{xy} é um máximo, substituir A e B pelo conjunto único C = AUB;

d) Para cada grupo X não envolvido nas junções prévias de grupos calcular txc, e retornar ao passo c. O processo, então, percorre os passos c e d até que, se p variáveis são retiradas, restam p grupos de variáveis.

b.1) Método da ligação máxima

É um método de ligação máxima, em que a medida de similaridade entre os grupos X e Y é dada por:

$$r_{xy} = \max_{\substack{i \in X \\ j \in Y}} r_{ij}$$

em que r_{ij} é o coeficiente de correlação entre as variáveis i e j.

b.2) Método da ligação média

É um método de ligação média, sendo a medida de similaridade

$$r_{xy} = \frac{\Sigma i \in x\ \Sigma j \in y\ r_{ij}}{n_1 n_2}$$

em que: n_1 e n_2 são os números de variáveis em X e Y, respectivamente.

Uma maneira óbvia de decidir quantas variáveis serão retidas quando esses métodos são usados, é continuar os passos c e d acima, até que todos os r_{xy} entre aqueles grupos restantes fiquem abaixo de algum $r_{\mathbf{0}}$. O processo então pára e o número de grupos formados naquele estágio é o número requerido de variáveis.

Quando utilizando métodos de agrupamento seria necessário um procedimento para selecionar uma variável de cada grupo. Supondo, por

exemplo, que um grupo consiste das variáveis X_1, X_2, X_3 e X_4 e que X_1 e X_2 formam os primeiros dois grupos destas variáveis a ficarem juntas. Estas seriam então unidas à X_3 e finalmente X_4. Entre as maneiras possíveis de selecionar uma destas variáveis seriam:

a) Escolher a última variável a se unir ao grupo, aqui X_4, referida como agrupamento externo;
b) Escolher uma das varáveis originais ou internas aos membros da maioria dos grupos, aqui, X_1 e X_2, nomeada como agrupamento interno.
c) Escolher uma das quatro variáveis aleatoriamente, aqui X_1, X_2, X_3 e X_4.

Dentre estas três possibilidades, o agrupamento interno seria um tanto mais rápido que os outros dois. Os dois métodos de agrupamentos seriam mais rápidos que qualquer dos outros métodos já citados para a seleção de variáveis independentes. Se a matriz de correlação for dada, ambos os métodos poderiam ser aplicados manualmente, o método de ligação simples para até 30-10 variáveis e o método de ligação média para 10-15 variáveis (JOLLIFFE, 1972).

2.5.4. R^2 múltiplo

Este critério examina o coeficiente de determinação múltiplo (R^2) na seleção do melhor conjunto de variáveis independentes (BEALE et al. 1967). O número de parâmetros na regressão será representado pelo subscrito p em R^2. Assim, R_p^2 indica que há p parâmetros, ou p – 1 variáveis independentes.

R_p^2 é a razão entre as somas de quadrados da regressão com p parâmetros (SQR_p) e a soma de quadrados total (SQTot).

$$R_p^2 = \frac{SQR_p}{SQTot}$$

A SQTot é constante para todas as equações, e R_p^2 varia inversamente proporcional a soma de quadrados dos resíduos (SQRes). Mas, é conhecido que a SQRes pode sempre diminuir à medida que variáveis independentes são

incluídas no modelo. Desta forma R_p^2 tem um máximo quando todas as k–1 variáveis independentes estão incluídas no modelo de regressão.

Mas na realidade o ideal é buscar o ponto em que a adição de mais uma variável independente não valeria a pena porque levaria a um pequeno aumento em R_p^2. Esse ponto, realmente, é encontrado somente quando um número limitado de variáveis independentes está incluído no modelo.

A determinação de onde a diminuição não retornaria benefício ao conjunto deve ser feita ao arbítrio do pesquisador (JOLLIFFE, 1972; NETER et al., 1989).

2.5.5. R_a^2 múltiplo ajustado

A diferença deste método com relação ao anterior é que considera o número de variáveis no modelo. É expresso por (NETER et al., 1989)

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQRes}{SQTot}$$

2.5.6. Quadrado Médio do Resíduo (QMRes.)

Devido a R_p^2 não considerar o número de parâmetros no modelo e que max(R_p^2) pode não diminuir nunca com o aumento de p, o uso do $QMRes_p$ seria um método alternativo. O $QMRes_p$ levaria em conta o número de parâmetros no modelo através dos graus de liberdade. Assim, min($QMRes_p$) pode diminuir com o aumento de p, se a redução em $QMRes_p$ se tornar tão pequena que não é suficiente para permitir a perda de um grau de liberdade adicional. Os usuários deste critério ou procuram o conjunto de variáveis independentes que minimizam $QMRes_p$ ou um conjunto para o qual $QMRes_p$ está tão próximo ao mínimo que a adição de mais uma variável independente não é desejável.

2.5.7. Critério C_p de Mallows

O critério C_p de Mallows (MALLOWS, 1973; DRAPER e SMITH, 1981) está relacionado com o total do quadrado médio do resíduo (QMRTot) das n observações ajustadas para qualquer modelo de regressão. O QMRTot teria um componente de viés e um componente de erro aleatório. A medida Γ_p proposta seria uma função padronizada do QMRTot.

Sendo

$E(\hat{Y}_i) - \mu_i =$ viés de i-ésima observação

$\sigma^2(\hat{Y}_i) =$ sua variância

tem-se que:

$$QMR = [E(\hat{Y}_i) - \mu_i] + \sigma^2(\hat{Y}_i)$$

e

$$QMRTot = \sum_{i=1}^{n}[E(\hat{Y}_i - \mu_i)^2 + \sum_{i=1}^{n}\sigma^2(\hat{Y}_i)]$$

$$\Gamma_p = \frac{1}{\sigma^2}\ \{\sum_{i=1}^{n}[E(\hat{Y}_i - \mu_i)^2 + \sum_{i=1}^{n}\sigma^2(\hat{Y}_i)]\}$$

em que:

σ^2 = variância do erro real.

O modelo que inclui todas as P–1 variáveis independentes é escolhido, cuidadosamente, tal que produz estimativas não viesadas de σ^2. Esta estimativa será denotada por $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = QM\ Res\ (X_1, X_2, \ldots, X_{p-1})$$

Pode ser demonstrado que (Neter et al., 1989)

$$\sum_{i=1}^{n}\sigma^2(\hat{Y}_i) = p\sigma^2$$

Assim, o erro aleatório total de n valores estimados ($\hat{Y}_i$) cresce com o aumento do número de variáveis no modelo de regressão. Sendo

$$E(SQ\ Res_p) = \sum_{i=1}^{n}[E(\hat{Y}_i - \mu_i)^2] + (n - p)\sigma^2$$

Desta forma Γ_p pode ser expresso como

$$\Gamma_p = \frac{1}{\sigma^2}[E(SQ\ Res_p) - (n - p)\ \sigma^2 + p\sigma^2] = \frac{E(SQRes_p)}{\sigma^2} - (n - 2p)$$

Substituindo $E(SQ\ Res_p)$ pelo estimador $SQ\ Res_p$ e usando $\hat{\sigma}^2$ = QMRes $(X_1, X_2, \ldots, X_{p-1})$ como um estimador de σ^2, foi demonstrado que um estimador de Γ_p é C_p (NETER et al., 1989):

$$C_p = \frac{SQRes_p}{\hat{\sigma}^2}(n - 2p)$$

Quando não há viés em uma equação de regressão com p–1 variáveis independentes, C_p tem o valor esperado de p:

$$E(C_p|vi \equiv n_i) = p$$

Assim, quando o valor de C_p para todas as regressões possíveis são colocados contra p, aquelas regressões com pouco viés tendem a ficar próximas da linha $C_p = p$. Aquelas regressões com grande viés tendem a ficar acima desta linha.

Usando o critério C_p, se busca identificar o conjunto de variáveis independentes que levam ao menor valor de C_p. Se um dado valor de C_p pode conter um substancial componente de viés, pode-se preferir um conjunto de variáveis independentes que leva a um C_p ligeiramente maior que não contém um grande componente de viés.

2.5.8. Método de busca t direta

Este procedimento de busca, primeiramente, ajustar um modelo com todas as P variáveis independentes. Todas as variáveis para as quais o valor absoluto da estatística t, expressa por:

$$t = \frac{b_k}{s(b_k)}$$

excederem um nível pré-determinado, essas são, automaticamente, retidas. O subconjunto de todas as possíveis regressões consistindo daquelas equações as quais contêm as variáveis automaticamente retidas, entre outras, é então obtido a partir de um critério especificado. Se, por exemplo, quatro variáveis estão sendo consideradas e resulta que X_1 e X_4 são automaticamente retidas, as regressões (X_1, X_4), (X_1, X_4, X_2), (X_1, X_4, X_3) e (X_1, X_4, X_2, X_3) poderiam ser estudadas com esta abordagem.

Este procedimento pode levar a sérias dificuldades. Suponha-se que se X_2 e X_3 são altamente correlacionados e cada um está proximamente relacionado com a variável dependente. Devido à colinearidade existente entre X_2 e X_3, ambos t_2 e t_3 podem ser absolutamente pequenos, levando a retirada de ambas variáveis com este procedimento. Um bom método de busca precisa estar habilitado para tratar variáveis independentes intercorrelacionadas de tal forma que nem todas elas seriam excluídas (NETER et al. 1989).

2.5.9. Método do incremento R^2 máximo e mínimo

São métodos que se fundamentam no incremento máximo ou no mínimo causado pela substituição de uma variável no modelo.

O método do R^2 máximo inicia procurando o modelo de uma variável que maximiza o R^2. Então uma outra variável que produz o maior aumento no R^2 é incluída. Uma vez com o modelo de duas variáveis, cada variável no modelo é comparada a cada variável fora dele. Para cada comparação o método determina se removendo uma variável e substituindo com outra aumenta o R^2. Depois da comparação de todas as trocas possíveis o método faz a troca que aumenta mais o R^2. As comparações começam novamente até não haver mais trocas que

aumentem o R^2. Assim, o modelo com duas variáveis é o “melhor” possível com duas variáveis e assim por diante.

A diferença com o método Stepwise é que todas as trocas são avaliadas, assim, este método consome mais tempo que Stepwise.

O método R^2 mínimo é o mesmo que o anterior, exceto pelo julgamento de busca de variáveis, pois inicia a busca do melhor modelo pela variável que produz o menor aumento do R^2 e não o maior. O método procura por mais modelos que o anterior.

2.5.10. Método Stepwise

O método Stepwise seria, provavelmente, o mais utilizado, pois não requer o cálculo de todas as regressões possíveis. Ele foi desenvolvido para evitar esforços computacionais, quando comparado com a abordagem de todas as regressões possíveis, a partir do momento que chega, razoavelmente, ao “melhor” conjunto de variáveis independentes.

Essencialmente, esse método calcula uma seqüência de equações de regressão, adicionando ou retirando a cada passo uma variável independente. O critério para adicionar ou retirar uma variável independente pode ser estabelecido, equivalentemente, em termos da redução da SQRes, coeficiente de correlação parcial ou estatística F.

O método segue os seguintes passos:

1. Para começar, a rotina calcula todas as regressões simples para cada uma das P-1 variáveis independentes. Para cada equação de regressão a estatística F é utilizada para testar se a inclinação zero é obtida.

$$F = \frac{\text{QM Reg}\,(X_k)}{\text{QM Res}\,(X_k)}$$

em que: QM Reg = Quadrado médio da regressão

QM Res = Quadrado médio do resíduo

Lembrar que QMReg (X_k) mede a redução na variação total de Y associado com o uso da variável independente X_k. A variável independente com o maior F significativo a um determinado nível de probabilidade α, é a candidata para a primeira inclusão. Caso contrário, o programa termina não considerando mais nenhuma variável independente, suficientemente, útil para entrar no modelo de regressão.

2. Seja que X_i é uma variável independente a entrar no passo 1, a rotina calcula agora todas as regressões com duas variáveis independentes, onde X_i é uma do par. Para cada regressão é obtida a estatística F

$$F = \frac{QM\,Reg\,(X_k\ /\ X_i)}{QM\,Res\,(X_i)} = \left[\frac{b_k}{S\,(b_k)}\right]^2$$

Esta será a estatística para testar se $\beta_k = 0$ quando X_i e X_k são as variáveis independentes no modelo. A variável independente com maior valor F é candidata para inclusão no segundo estágio. Se este valor F excede um nível predeterminado a segunda variável independente é adicionada, caso contrário o programa termina.

3. Seja X_j a variável adicionada no segundo estágio. Agora a rotina examina se qualquer uma das outras variáveis independentes já no modelo pode ser eliminada. Supondo que neste estágio há somente uma outra variável independente $X_{i,}$ tal que somente uma estatística F é obtida.

$$F = \frac{QM\,Reg\,(X_j\ /\ X_i)}{QM\,Res\,(X_i\ .\ X_j)}$$

Nos próximos estágios, poderia haver um número de estatísticas F, para cada uma das variáveis no modelo ao lado daquela última adicionada. A variável para a qual este valor de F é menor é a candidata à eliminação. Se F ficar abaixo de um limite predeterminado, a variável independente é eliminada do modelo, caso contrário é retida.

4. Considere que ambas X_j e X_i estão agora no modelo. A partir deste momento a rotina examina quais variáveis são as próximas candidatas à inclusão, examina, então, se quaisquer variáveis já no modelo poderiam ser eliminadas e assim por diante até não haver mais variáveis independentes que possam ser incluídas ou retiradas, neste ponto a busca termina.

Nota-se que o método permite que uma variável independente trazida ao modelo em um estágio anterior, seja eliminada em um estágio posterior, se ela não for mais útil no conjunto com as variáveis adicionadas em estágios posteriores.

Os limites de F para adição e retirada de variáveis não precisam ser, necessariamente, o mesmo. Freqüentemente, o limite F para retirada de uma variável é especificado como sendo menor que o limite para inclusão de variáveis (NETER et al., 1989).

Deve-se observar que o Stepwise pode resultar em combinações lineares das variáveis independentes que não apresentem a menor SQRes. Isso só é possível através do ajuste de todas as possíveis combinações (método de todas as possibilidades) e, então, comparando seus resultados (FREESE, 1964)

2.5.11. Método Forward

O procedimento de seleção Forward tem o objetivo de chegar a conclusões trabalhando a partir da inclusão de variáveis no modelo de regressão até que ele esteja satisfatório (DRAPER e SMITH, 1981).

A ordem de inclusão é determinada pelo uso do coeficiente de correlação parcial como uma medida da importância da variável ainda não incluída no modelo. O procedimento básico é como segue:

1. Selecionar a variável X_i mais correlacionada com Y, suponha-se ser X_1, e encontrar a regressão linear $\hat{Y} = f(X_1)$;

2. O próximo passo é encontrar o coeficiente de correlação parcial de X_j ($j \neq 1$) e Y, após a alocação de X_1. O X_j com o maior coeficiente de

correlação parcial com Y é agora selecionado, suponha-se ser X_2, e uma segunda equação de regressão $Y = f(X_1, X_2)$ é ajustada. O processo continua.

3. Depois de $X_1, X_2, ..., X_q$ estarem na regressão os coeficientes de correlação parcial são as correlações entre:

   (a) os resíduos da regressão $Y = f(X_1, X_2, ..., X_q)$
   (b) os resíduos da regressão $X_j = f(X_1, X_2, ..., X_p)$ $(j > q)$.

Conforme cada variável é incluída no modelo, os seguintes valores são examinados:

(a) R^2: o coeficiente de correlação múltiplo;

(b) O teste do valor do F parcial para a última variável incluída, o qual mostraria se alcançou uma quantidade significativa de variação daquela removida pelas variáveis previamente na regressão.

O método seria uma simplificação do Stepwise, omitindo o teste em que uma variável já incluída no modelo poderia ser retirada. É mais econômico sob o ponto de vista computacional e permite trabalhar com mais variáveis independentes.

Uma de suas desvantagens é que ele não faz esforços no sentido de explorar os efeitos que a introdução de uma nova variável pode ter sobre o papel desempenhado por uma variável que foi incluída em um estágio anterior.

2.5.12. Método Backward

O método seria um melhoramento da abordagem de todas as regressões possíveis, no que ele objetiva permitir o exame não de todas as regressões, mas de somente a "melhor" regressão contendo um certo número de variáveis.

O método consiste em:

1. Ajustar uma regressão contendo todas as variáveis;

2. O valor do teste F parcial é calculado para toda variável considerada como se “ela fosse a última variável a entrar na equação de regressão”;

3. O menor valor do F parcial, suponha-se F_L, é comparado com um nível de significância pré-selecionado, suponha-se F_0:

   (a) Se $F_L < F_0$ a variável X_L é removida. Faz-se o reajustamento da equação de regressão das variáveis restantes e retorna ao estágio 2;

   (b) Quando $F_L > F_0$, adotar a equação como selecionada.

Esse é um procedimento satisfatório a partir do momento em que não desconsidera, inicialmente, nenhuma variável. Também requer menor esforço computacional que o método de todas as regressões. Entretanto, se os dados de entrada produzem uma matriz X’X mal condicionada; isto é, aproximadamente singular, então esse procedimento pode produzir distorções devido a erros de arredondamento. O método seria ligeiramente inferior ao Forward (DRAPER e SMITH, 1981).

2.5.13. Critério de Akaike

As técnicas de seleção de variáveis consideradas, anteriormente, Stepwise, Backward e Forward, também podem ser empregadas usando o Critério de Akaike (AIC) (AKAIKE, 1974) que é expresso por:

$$AIC = -2L(\beta) + 2K$$

sendo K o número de variáveis independentes (ou, explicativas) e L(β) a log-verossimilhança do modelo avaliado em $\hat{\beta}_i$ parâmetros estimados.

As duas primeiras técnicas são iniciadas no modelo completo (isto é, com todas as variáveis independentes), mas, diferentemente, do que acontece no

Backward, em cada passo do Stepwise, após a exclusão de uma variável, é verificado se alguma das variáveis que estão fora do modelo podem entrar no mesmo.

A terceira técnica é iniciada no modelo nulo (contém apenas o intercepto) e a cada passo é verificado se a inclusão de uma determinada variável faz com que o AIC diminua. Os três procedimentos terminam quando nenhuma das variáveis que entrarão (ou sairão) do modelo fizerem com que o AIC seja diminuído e estabilizado.

Segundo Cordeiro e Paula (1989) tal critério quando aplicado a modelos lineares produz soluções razoáveis em casos em que a teoria tradicional da máxima verossimilhança não pode ser aplicada.

2.5.14. Método de todas as possibilidades

Nesse método que exige intensos procedimentos computacionais, faz-se o ajuste de todos os possíveis modelos, o que o torna inviável quando existe um grande número de variáveis independentes.

Portanto, para um modelo linear envolvendo k variáveis independentes, existirá 2^k modelos, devendo-se assumir que o número de observações tem que exceder o número de potenciais parâmetros (n>k) (CLUTTER et al., 1983, NETER et al., 1989).

Portanto, o número de modelos incluindo o parâmetro β_0 é expresso por:

$$\sum_{p=0}^{k} C_p^k = \sum_{p=0}^{k} \binom{k}{p} = 2^k$$

Prova

Seja

$$(X+a)^k = \binom{k}{0} X^k a^0 + \binom{k}{1} X^{k-1} a^1 + \ldots + \binom{k}{k} X^0 a^k$$

$$(X+a)^k = \sum_{p=0}^{k} C_p^k X^{k-p} a^p$$

Fazendo X=1 e a=1, tem-se:

$$\sum_{p=0}^{k} C_p^k 1^{k-p} 1^p = (1+1)^k = 2^k$$

Então, no processo de todas as possibilidades, o número de modelos a ser testado cresce, exponencialmente, em relação ao número de variáveis independentes testadas.

2. 4.15. Outros procedimentos

Na literatura especializada se podem encontrar outros métodos de seleção de variáveis, destacando-se entre eles a Regressão Ridge e o PRESS (Predição da Soma de Quadrados) (DRAPER e SMITH, 1981; NETER et al., 1989).

3. MATERIAL E MÉTODOS

No presente estudo, foram utilizadas 201 árvores de *Leucaena leucocephala* (Lam.) de Wit., provenientes de um experimento conduzido na Estação Experimental da Empresa Pernambucana de Pesquisa Agropecuária (IPA) em Caruaru – PE, cujas coordenadas geográficas de posição: lat. 08°14'18''S; 38°00'00''WGr. e altitude de 537 m. Pela classificação climática de Thornthwaite esta área está enquadrada no tipo Dd'a' (semi-árido megatérmico). Na área considerada existe uma associação de solos litólicos com características de solos planassólicos e podzólicos vermelho-amarelos (RIBEIRO, 2001).

As mudas de leucena foram produzidas a partir de sementes procedentes de Ibimirim e o plantio no campo foi feito em 20 / 12 / 89, com espaçamento 2m x 1 m.

Cada árvore amostrada foi derrubada para as medições de campo. Elas foram cortadas a 0,30 m do solo. Os dados para a cubagem das árvores foram coletados em dezembro de 1998. Tomou-se para cada árvore-amostra, as medidas de circunferência a 0,30; 0,50; 0,70; 0,90; 1,10; 1,30; 1,50 e 1,70 m de altura, e a partir de 2,30 m, foram tomadas as circunferências necessárias a cada 1,0 m(RIBEIRO, 2001).

Para os galhos foram medidos o comprimento e as circunferências na base e no topo. As medidas de circunferência foram feitas com fitas métricas, graduadas em centímetros com aproximação de 0,5 centímetro e as alturas foram tomadas nas árvores derrubadas, em metros, com aproximação de centímetros, usando-se trena.

Os volumes (fuste e galhos) foram calculados, empregando-se a fórmula de Smalian (HUSCH et al., 1972) assim expressa:

$$V = \frac{h}{2}\left(A_b + A_u\right)$$

Em que: V = volume em m^3;

h = altura ou comprimento em m;

A_b = área transversal da base em m^2;

A_u = área transversal do topo em m^2;

Para o uso dos métodos de seleção, as seguintes variáveis foram consideradas (Tabela 1):

Tabela 1. Descrição de todas as variáveis usadas.

Variáveis	Descrição
Vtot	Volume total da árvore (m^3)
Cap	Circunferência a altura do peito (1,30 m) (m)
H	Altura total da árvore (m)
NG	Número de galhos
CG	Circunferência na base do galho maior (m)
C30	Circunferência a 0,30 m de altura (m)
V1	Volume da tora de 0,30 m a 0,90 m (m^3)
V2	Volume de 0,50 m a 1,10 m (m^3)
V3	Volume de 0,70 m a 1,30 m (m^3)
V4	Volume de 0,90 m a 1,50 m (m^3)
V5	Volume de 1,10 m a 1,70 m (m^3)
V6	Volume de 0,30 m a 1,10 m (m^3)
V7	Volume de 0,50 m a 1,30 m (m^3)
V8	Volume de 0,70 m a 1,50 m (m^3)
V9	Volume de 0,90 m a 1,70 m (m^3)
V10	Volume de 0,30 m a 1,30 m (m^3)
V11	Volume de 0,50 m a 1,50 m (m^3)
V12	Volume de 0,70 m a 1,70 m (m^3)
V13	Volume de 0,30 m a 1,50 m (m^3)
V14	Volume de 0,50 m a 1,70 m (m^3)
V15	Volume de 0,30 m a 1,70 m (m^3)

Primeiramente, calcularam-se as estatísticas descritivas de cada variável considerada, o que permitiu que fosse também calculado o número mínimo de unidades de amostra (NMUA) representativo para cada variável, considerando o processo de amostragem inteiramente aleatório (MEUNIER et al., 2002).

$$NMUA = \frac{t_{\alpha}^2 . CV^2}{E^2 \%}$$

Em que:

t_{α} = valor tabelado do t de Student no nível de 5% de probabilidades;

CV = coeficiente de variação;

E = erro estipulado (10%).

Depois de realizados os cálculos de NMUA, os erros de amostragem (EA%) para cada variável também foram calculados para que fosse comparado com o erro estipulado.

O erro de amostragem é expresso por:

$$EA\% = \frac{t_{\alpha} s_{\bar{x}}}{\bar{X}}$$

Em que:

$s_{\bar{x}}$ = erro padrão da média;

$\bar{X}$ = média da variável independente considerada.

As técnicas de seleção de variáveis aplicadas nos foram os seguintes:

- **Análise de Componentes Principais**
  - e) Método de seleção de variáveis por retenção de k componentes
  - f) Método de Beale
  - g) Método de Beale interativo
  - h) Método de seleção por coeficiente de correlação múltipla
- **Métodos de agrupamento**
  - a) Método de ligação máxima
  - b) Método de ligação média
- **R^2 múltiplo**
- **Critério Cp de Mallows**
- **Busca direta de t**
- **Incremento de R^2 máximo e mínimo**

- **Stepwise**
- **Forward**
- **Backward**
- **Critério de Akaike**

O método de todas as possibilidades mesmo sendo o mais preciso porque estuda todos os possíveis modelos envolvendo todas as variáveis independentes, não foi considerado no presente estudo visto que seria necessário avaliar 1048576 (2^{20}) equações.

A descrição dos métodos citados já foi apresentada na revisão de literatura desta dissertação.

Após a seleção de variáveis pelos métodos acima, foram geradas equações lineares para comparação de qual técnica explicaria melhor o volume total da árvore. Entretanto, devido a dificuldade de mensuração no campo de algumas variáveis, principalmente, a altura total, usamos critérios pessoais para selecionar a melhor equação, em função da precisão e custos (tempo para obter as medidas no campo).

Os programas estatísticos utilizados para análise foram: R, versão 1.8.0, (VENABLES & RIPLEY 2002), o qual poder ser obtido gratuitamente em http://www.r-project.org, SYSTAT (System for Statistics) versão 10 DEMO e SAS (Statistical Analysis System) (1982).

4. RESULTADOS E DISCUSSÃO

4.1. Estatísticas Descritivas

Os resultados das estatísticas descritivas para todas as variáveis são os que constam na Tabela 01.

Tabela 01. Estatísticas descritivas e suficiência amostral para todas variáveis.

	VTOTAL	CAP	H	NG	CG	C30
Média	0.0117	0.1657	5.3795	2.2736	0.1181	0.2003
Desvio padrão	0.0081	0.0458	1.0044	1.6022	0.0357	0.0502
C.V.	0.6933	0.2762	0.1867	0.7047	0.3022	0.2509
NMUA	183.0000	29.0000	14.0000	189.0000	35.0000	24.0000
EA%	9.5693	3.8205	2.5807	9.7405	4.1782	3.4642

	V1	V2	V3	V4	V5	V6
Média	0.0018	0.0017	0.0016	0.0015	0.0014	0.0024
Desvio padrão	0.0010	0.0009	0.0009	0.0008	0.0008	0.0013
C.V.	0.5400	0.5368	0.5525	0.5525	0.5588	0.5312
NMUA	111.0000	110.0000	116.0000	116.0000	119.0000	120.0000
EA%	7.6790	7.3176	7.7736	6.9111	7.8984	7.4870

	V7	V8	V9	V10	V11	V12
Média	0.0022	0.0020	0.0019	0.0029	0.0026	0.0025
Desvio padrão	0.0012	0.0011	0.0011	0.0015	0.0014	0.0014
C.V.	0.5407	0.5465	0.5639	0.5356	0.5352	0.5569
NMUA	112.0000	114.0000	121.0000	110.0000	109.0000	119.0000
EA%	7.5394	7.6022	8.0024	7.1494	7.4427	7.7440

	V13	V14	V15
Média	0.0034	0.0031	0.0038
Desvio padrão	0.0018	0.0017	0.0021
C.V.	0.5293	0.5450	0.5383
NMUA	107.0000	113.0000	111.0000
EA%	7.3176	7.5800	7.6386

Em que:
C.V.= Coeficiente de Variação

NMUA= Número Mínimo de Unidades Amostrais para um erro estipulado de 10%

EA%= Erro de amostragem

Observa-se que os erros de amostragem para todas variáveis foram inferiores a 10%, significando que o tamanho da amostra utilizada é representativa para toda população.

A matriz de correlação entre todas as variáveis está na Tabela 02.

Tabela 02. Matriz de correlação para todas as variáveis

	VTOT	CAP	H	NG	CG	C030	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	V15
VTOT	1.000	0.904	0.668	0.574	0.830	0.952	0.974	0.967	0.963	0.956	0.944	0.970	0.964	0.964	0.958	0.966	0.963	0.966	0.967	0.966	0.969
CAP	0.904	1.000	0.665	0.501	0.755	0.946	0.932	0.952	0.958	0.958	0.965	0.952	0.962	0.953	0.948	0.960	0.957	0.945	0.957	0.949	0.949
H	0.668	0.665	1.000	0.253	0.511	0.658	0.625	0.628	5.628	0.638	0.645	0.630	0.633	0.637	0.644	0.634	0.644	0.643	0.645	0.650	0.651
NG	0.574	0.501	0.253	1.000	0.461	0.530	0.523	0.513	0.528	0.492	0.492	0.536	0.509	0.508	0.484	0.532	0.489	0.501	0.514	0.482	0.507
CG	0.830	0.755	0.511	0.461	1.000	0.823	0.808	0.808	0.792	0.777	0.763	0.807	0.799	0.790	0.772	0.799	0.796	0.786	0.797	0.793	0.793
C030	0.952	0.946	0.658	0.530	0.823	1.000	0.975	0.970	0.966	0.957	0.947	0.978	0.970	0.968	0.953	0.977	0.971	0.965	0.980	0.968	0.977
V01	0.974	0.932	0.625	0.523	0.808	0.975	1.000	0.989	0.986	0.986	0.965	0.992	0.987	0.983	0.984	0.990	0.984	0.982	0.989	0.983	0.988
V02	0.967	0.952	0.628	0.513	0.808	0.970	0.989	1.000	0.991	0.987	0.983	0.996	0.996	0.989	0.983	0.991	0.944	0.986	0.991	0.991	0.988
V03	0.963	0.958	0.625	0.528	0.792	0.966	0.986	0.991	1.000	0.984	0.946	0.991	0.995	0.995	0.979	0.994	0.989	0.991	0.990	0.985	0.986
V04	0.956	0.958	0.638	0.492	0.777	0.957	0.986	0.987	0.984	1.000	0.985	0.985	0.989	0.988	0.995	0.985	0.991	0.984	0.989	0.987	0.985
V05	0.944	0.965	0.645	0.492	0.763	0.947	0.965	0.983	0.976	0.985	1.000	0.981	0.982	0.976	0.988	0.978	0.983	0.982	0.981	0.987	0.985
V06	0.970	0.952	0.630	0.536	0.807	0.978	0.992	0.996	0.991	0.985	0.981	1.000	0.993	0.988	0.981	0.996	0.989	0.985	0.994	0.987	0.992
V07	0.964	0.962	0.633	0.509	0.799	0.970	0.987	0.996	0.995	0.989	0.982	0.993	1.000	0.991	0.983	0.996	0.995	0.987	0.993	0.991	0.989
V08	0.964	0.953	0.637	0.508	0.790	0.968	0.983	0.989	0.995	0.988	0.978	0.988	0.991	1.000	0.982	0.989	0.994	0.995	0.994	0.990	0.989
V09	0.958	0.948	0.644	0.484	0.772	0.953	0.984	0.983	0.979	0.995	0.988	0.971	0.983	0.982	1.000	0.980	0.985	0.987	0.983	0.991	0.989
V10	0.966	0.960	0.634	0.532	0.799	0.977	0.990	0.991	0.994	0.985	0.978	0.996	0.996	0.989	0.980	1.000	0.990	0.985	0.996	0.986	0.992
V11	0.963	0.957	0.644	0.489	0.796	0.971	0.984	0.994	0.989	0.991	0.983	0.989	0.995	0.994	0.985	0.990	1.000	0.989	0.996	0.995	0.991
V12	0.966	0.645	0.643	0.501	0.786	0.965	0.982	0.986	0.991	0.984	0.982	0.985	0.987	0.995	0.987	0.985	0.989	1.000	0.989	0.994	0.993
V13	0.967	0.957	0.645	0.514	0.797	0.980	0.989	0.991	0.990	0.989	0.981	0.994	0.993	0.994	0.983	0.996	0.996	0.989	1.000	0.992	0.996
V14	0.966	0.949	0.650	0.482	0.793	0.968	0.983	0.991	0.985	0.987	0.987	0.987	0.991	0.990	0.991	0.986	0.995	0.994	0.992	1.000	0.996
V15	0.969	0.949	0.651	0.507	0.793	0.977	0.988	0.988	0.986	0.985	0.985	0.992	0.989	0.989	0.989	0.992	0.991	0.993	0.996	0.996	1.000

Nota-se que quase todas as variáveis apresentam altos graus de correlação entre si, principalmente, aquelas referentes a volumes de secções no tronco da árvore, o que de certa forma vai de encontro ao princípio da análise de regressão correspondente à independência entre as variáveis independentes. Entretanto, é extremamente difícil atingir esse requisito quando se trabalha com populações biológicas em que as variáveis que fazem parte dos indivíduos compõem sistemas complexos e inter-relacionados. Também como o tronco de uma árvore pode ser resultado da combinação de vários sólidos geométricos (cilindro, neilóide, parabolóide, tronco de cone, etc), procurou-se estudar uma variedade de volumes de secções no tronco da árvore com o objetivo de selecionar a(s) mais correlacionada(s) com o volume total da árvore.

4.2. Análise de Componentes Principais

A importância dos Componentes Principais se encontra na Tabela 03.

Tabela 03. Importância dos Componentes Principais

	Comp. 1	**Comp.2**	**Comp. 3**	**Comp. 4**	**Comp.5**
Desvio padrão	4.2502602	0.88290216	0.73258341	0.59591578	0.291840971
Proporção da Variância	0.9032356	0.03897581	0.02683392	0.01775578	0.004258558
Proporção cumulativa	0.9032356	0.94221142	0.96904534	0.98680112	0.991059680
	Comp. 6	**Comp.7**	**Comp. 8**	**Comp. 9**	**Comp.10**
Desvio padrão	0.247820437	0.181185057	0.171003663	0.142870982	0.1240983584
Proporção da Variância	0.003070748	0.001641401	0.001462113	0.001020606	0.0007700201
Proporção cumulativa	0.994130429	0.995771830	0.997233943	0.998254548	0.9990245686
	Comp. 11	**Comp.12**	**Comp. 13**	**Comp. 14**	**Comp.15**
Desvio padrão	0,1014294931	0.0894053360	3.444768e-02	3.153191e-03	2.914181e-03
Proporção da Variância	0.0005143971	0.0003996657	5.933214e-05	4.971306e-07	4.246227e-07
Proporção cumulativa	0.9995389657	0.9999386314	9.999980e-01	9.999985e-01	9.999989e-01
	Comp. 16	**Comp.17**	**Comp. 18**	**Comp. 19**	**Comp.20**
Desvio padrão	2.625467e-03	2.341253e-03	2.005869e-03	1.868368e-03	1.550878e-03
Proporção da Variância	3.446537e-07	2.740732e-07	2.011755e-07	1.745400e-07	1.202612e-07
Proporção cumulative	9.999992e-01	9.999995e-01	9.999997e-01	9.999999e-01	1.000000e+00

Baseando-se nas importâncias de cada componente principal, aplicaram-se os métodos de seleção de variáveis descritos abaixo:

i) Método de seleção de variáveis por retenção de k componentes

Os resultados apresentados pelo método de seleção de variáveis por retenção de K componentes estão na Tabela 04.

Tabela 04. Coeficientes das variáveis em cada componente principal

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
CAP	-0.227				0.872				0.173			0.113	-0.337							
H	-0.156	-0.416	0.887																	
NG	-0.125	0.895	0.339	0.124																
CG	-0.192	0.101		-0.956		-0.170														
C030	-0.231					0.684	-0.427	0.131	0.129	0.103	-0.267	-0.109	0.401							
V01	-0.233				-0.302	0.034		-0.470	0.192	-0.164	-0.132		-0.368	0.203	0.569		0.120			
V02	-0.234						0.142	-0.149	-0.402	0.193	-0.488	-0.110	-0.168	-0.115	-0.323	-0.121	0.480			
V03	-0.234						0.493		0.115	-0.351	-0.104		0.305	0.432	-0.125	-0.404	-0.203		-0.184	
V04	-0.233					0.240		-0.394	0.338	0.276	0.131	-0.235	0.219	0.327	-0.301	0.395	0.146	0.148		
V05	-0.232			0.126	0.195	-0.397	-0.277	0.159	-0.406	-0.149		-0.336	0.306	0.161	0.445					
V06	-0.234					0.137		-0.177	-0.367	-0.195		-0.342	-0.252		-0.209	0.284	-0.618			
V07	-0.234						0.272	-0.130	-0.208	0.121		0.457	0.297	-0.222	0.234	0.421		-0.248	0.347	-0.131
V08	-0.234						0.357	0.279	0.305	0.109	-0.374			-0.473	0.266			0.390		-0.166
V09	-0.232					-0.371	-0.347	-0.221	0.282	-0.117		0.156	0.151	-0.485	-0.167	-0.349	-0.243	-0.121		
V10	-0.234					0.192	0.128	-0.160	-0.180	-0.271	0.412	0.137	0.193	-0.218	-0.108		0.321	0.168	-0.558	0.112
V11	-0.234						0.126	0.103		0.595				0.151		0.279	-0.300	-0.274	-0.469	-0.210
V12	-0.233				-0.173	-0.126		0.459	0.250	-0.274	-0.217		-0.127		-0.140	0.366	0.143	-0.468	-0.239	0.146
V13	-0.234		0.151							0.165	0.529	-0.195	-0.151			-0.161	-0.102	-0.289	0.418	0.497
V14	-0.234				-0.125	-0.136	-0.160	0.276		0.222	-0.123	0.441	-0.134	0.162				0.564		0.401
V15	-0.234				-0.155		-0.287	0.229		-0.193	0.324	0.165	-0.230	0.106	-0.174				0.232	-0.668

Observa-se que o primeiro componente explica cerca de 90% de toda variabilidade dos dados. Junto ao segundo componente, esse percentual chega a 94%, resultando em um K=2. Portanto, por este método se devem excluir os 18 componentes de menor importância, restando apenas os componentes C1 e C2, correspondentes as variáveis NG e V2.

A distribuição dos resíduos para o Método dos Componentes Principais com Retenção de K Componentes se encontra no Gráfico 01.

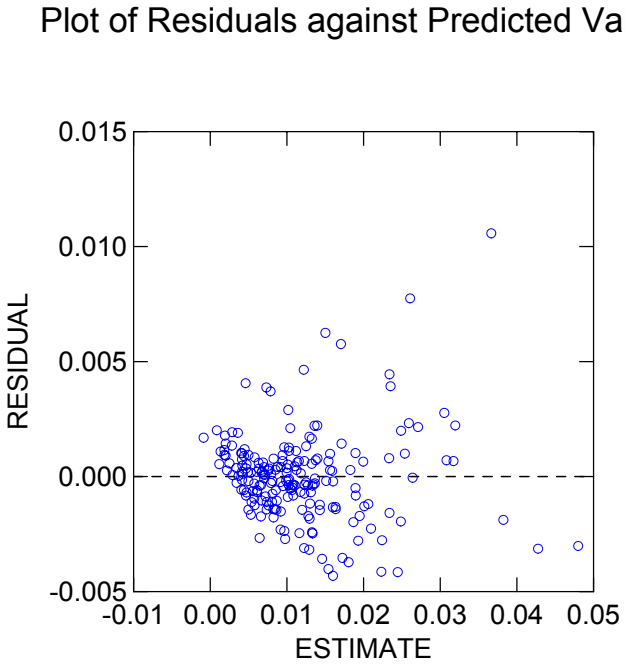

Gráfico 01. Distribuição residual para o Método dos Componentes Principais com Retenção de K Componentes.

j) Método de Beale

Os autovalores associados a cada um dos componentes principais então mostrados na Tabela 05.

Tabela 05. Autovalores associados a cada um dos componentes principais

Comp.1	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**	**Comp. 7**	**Comp. 8**	**Comp. 9**	**Comp. 10**
18.065	0.780	0.537	0.355	0.085	0.061	0.033	0.029	0.020	0.015
Comp. 11	**Comp. 12**	**Comp. 13**	**Comp. 14**	**Comp. 15**	**Comp. 16**	**Comp. 17**	**Comp. 18**	**Comp. 19**	**Comp. 20**
0.010	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Observando a tabela acima, devem-se escolher os dois primeiros componentes principais como sugere Jollife (1972), ao mesmo tempo em que se excluí todos os componentes principais em que os autovalores são menores que 0.70.

Assim sendo, foram excluídas todas as variáveis com maiores valores absolutos em ordem decrescentes dos componentes principais, restando apenas as variáveis NG e V2, resultado semelhando ao procedimento anterior.

O gráfico dos resíduos para este método foi igual ao anterior, pois as variáveis independentes selecionadas foram as mesmas, e, consequentemente, apresentaram a mesma equação resultante.

k) Método de Beale interativo

Os resultados obtidos pelo método de Beale interativo estão nas Tabelas 06 e 07.

Tabela 06. Resultados do método de Beale interativo para o primeiro passo.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
CAP	-0.227				0.872				0.173			0.113	-0.337							
H	-0.156	-0.416	0.887																	
NG	-0.125	0.895	0.339	0.124																
CG	-0.192	0.101		-0.956		-0.170														
C030	-0.231					0.684	-0.427	0.131	0.129	0.103	-0.267	-0.109	0.401							
V01	-0.233				-0.302	0.034		-0.470	0.192	-0.164	-0.132		-0.368	0.203	0.569		0.120			
V02	-0.234						0.142	-0.149	-0.402	0.193	-0.488	-0.110	-0.168	-0.115	-0.323	-0.121	0.480			
V03	-0.234						0.493		0.115	-0.351	-0.104		0.305	0.432	-0.125	-0.404	-0.203		-0.184	
V04	-0.233					0.240		-0.394	0.338	0.276	0.131	-0.235	0.219	0.327	-0.301	0.395	0.146	0.148		
V05	-0.232			0.126	0.195	-0.397	-0.277	0.159	-0.406	-0.149		-0.336	0.306	0.161	0.445					
V06	-0.234					0.137		-0.177	-0.367	-0.195		-0.342	-0.252		-0.209	0.284	-0.618			
V07	-0.234						0.272	-0.130	-0.208	0.121		0.457	0.297	-0.222	0.234	0.421		-0.248	0.347	-0.131
V08	-0.234						0.357	0.279	0.305	0.109	-0.374			-0.473	0.266			0.390		-0.166
V09	-0.232					-0.371	-0.347	-0.221	0.282	-0.117		0.156	0.151	-0.485	-0.167	-0.349	-0.243	-0.121		
V10	-0.234					0.192	0.128	-0.160	-0.180	-0.271	0.412	0.137	0.193	-0.218	-0.108		0.321	0.168	-0.558	0.112
V11	-0.234						0.126	0.103		0.595				0.151		0.279	-0.300	-0.274	-0.469	-0.210
V12	-0.233				-0.173	-0.126		0.459	0.250	-0.274	-0.217		-0.127		-0.140	0.366	0.143	-0.468	-0.239	0.146
V13	-0.234		0.151							0.165	0.529	-0.195	-0.151			-0.161	-0.102	-0.289	0.418	0.497
V14	-0.234				-0.125	-0.136	-0.160	0.276		0.222	-0.123	0.441	-0.134	0.162				0.564		0.401
V15	-0.234				-0.155		-0.287	0.229		-0.193	0.324	0.165	-0.230	0.106	-0.174				0.232	-0.668

No processo interativo, foram necessárias 19 interações para que se chegasse ao resultado final (Tabela 07)

Tabela 07. Resultado final do método de Beale interativo.

CP1	CP2
1.826	0.747

As importâncias dos componentes principais selecionados estão na Tabela 08.

Tabela 08. Importância dos componentes principais selecionados

	Comp. 1	Comp.2
Desvio padrão	1.1193244	0.8643569
Proporção da Variância	0.6264435	0.3735565
Proporção cumulativa	0.6264435	1.0000000

Os resultados da matriz de covariância das variáveis selecionadas estão na tabela 09.

Tabela 09. Matriz de covariância para as variáveis selecionadas.

	Comp. 1	Comp.2
H	0.707	0.707
NG	0.707	-0.707

Este método é uma variante do anterior, sendo que em cada estágio do procedimento se rejeitou a variável associada com o menor componente e com maior coeficiente em valor absoluto. Após a exclusão, fez-se uma nova análise sem a variável excluída e assim por diante, até que o processo parou quando restaram apenas auto-valores maiores que o limiar λ_0 = 0.70, resultando na seleção de H e NG.

Observa-se que pelos dois métodos de Beale, os resultados não foram iguais, uma vez que mesmo tendo igual número de variáveis independentes selecionadas, apresentaram-se dois subconjuntos diferentes, no primeiro selecionou-se NG e V2, e no segundo NG e H.

A distribuição dos resíduos para o Método dos Componentes Principais pelo Método de Beale Interativo se encontra no gráfico 02.

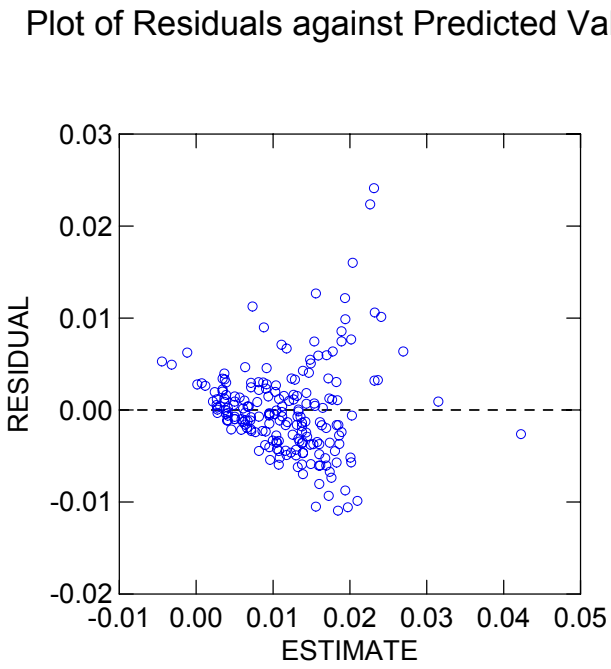

Gráfico 02. Distribuição residual para os Componentes Principais, Método de Beale Interativo.

Observa-se que a distribuição residual do Método de Beale Interativo quando comparado com os métodos de Retenção de K Componentes e de Beale, apresentou-se maiores valores residuais, principalmente, para resíduos positivos, o que se justifica também através dos valores de R^2 encontrados (ver Tabela 23).

Caso os únicos métodos utilizados fossem os três citados anteriormente, por questões práticas, o método da Retenção de K Componentes ou o de Beale deveria ser o indicado, pois envolve a variável V2 que é mais fácil de ser mensurada que a variável H.

I) Método de seleção por coeficiente de correlação múltipla

Os resultados para o método acima citado estão na Tabela 10.

Tabela 10. Matriz de coeficientes de correlação múltipla

	CAP	H	NG	CG	CO30	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	V15
CAP	1.000																			
H	0.8251	1.000																		
NG	0.8368	0.6214	1.000																	
CG	0.8674	0.7692	0.7351	1.000																
C030	0.9068	0.9097	0.9134	0.9134	1.000															
V01	0.9478	0.9536	0.9535	0.9532	0.9480	1.000														
V02	0.9375	0.0408	0.9429	0.9414	0.9380	0.9485	1.000													
V03	0.9322	0.9352	0.9340	0.9400	0.9351	0.9482	0.9363	1.000												
V04	0.9164	0.9202	0.9287	0.9334	0.9309	0.9482	0.9347	0.9301	1.000											
V05	0.8908	0.8963	0.9061	0.9192	0.9235	0.9480	0.9360	0.9282	0.9146	1.000										
V06	0.9438	0.9456	0.9443	0.9465	0.9404	0.9485	0.9402	0.9404	0.9402	0.9416	1.000									
V07	0.9377	0.9356	0.9393	0.9397	0.9449	0.9482	0.9349	0.9316	0.9304	0.9304	0.9402	1.000								
V08	0.9307	0.9334	0.9379	0.9410	0.9345	0.9489	0.9370	0.9304	0.9292	0.9284	0.9414	0.9334	1.000							
V09	0.9177	0.9220	0.9335	0.9376	0.9346	0.9477	0.9363	0.9333	0.9186	0.9180	0.9413	0.9330	0.9324	1.000						
V10	0.9409	0.9388	0.9387	0.9430	0.9351	0.9480	0.9383	0.9343	0.9344	0.9338	0.9401	0.9342	0.9366	0.9367	1.000					
V11	0.9319	0.9320	0.9421	0.9388	0.9331	0.9488	0.9352	0.9333	0.9283	0.9284	0.9410	0.9314	0.9309	0.9310	0.9661	1.000				
V12	0.9340	0.9370	0.9441	0.9463	0.9391	0.9504	0.9405	0.9353	0.9344	0.9342	0.0440	0.9381	0.9338	0.9339	0.9406	0.9361	1.000			
V13	0.9409	0.9387	0.9434	0.9449	0.9359	0.9487	0.9394	0.9370	0.9354	0.9362	0.9409	0.9364	0.9359	0.9369	0.9365	0.9354	0.9395	1.000		
V14	0.9351	0.9362	0.9488	0.9444	0.9379	0.9501	0.9383	0.9382	0.9337	0.9375	0.9433	0.9361	0.9361	0.9335	0.9401	0.9338	0.9361	0.9384	1.000	
V15	0.9421	0.9420	0.9488	0.9447	0.9402	0.9502	0.9433	0.9418	0.9398	0.9438	0.9438	0.9412	0.9408	0.9397	0.9410	0.9402	0.9404	0.9401	0.9398	1.000

Com um valor fixado de k=2, obtiveram-se os coeficientes de correlação múltipla para cada um dos cento e noventa modelos ajustados com duas variáveis independentes. O modelo selecionado foi aquele que proporcionou um maior R^2 do qual foi possível concluir que as variáveis H e V1 explicaram 95,36% das variações dos dados.

A distribuição residual para o Método do Coeficiente de Correlação Múltipla se encontra no gráfico 03.

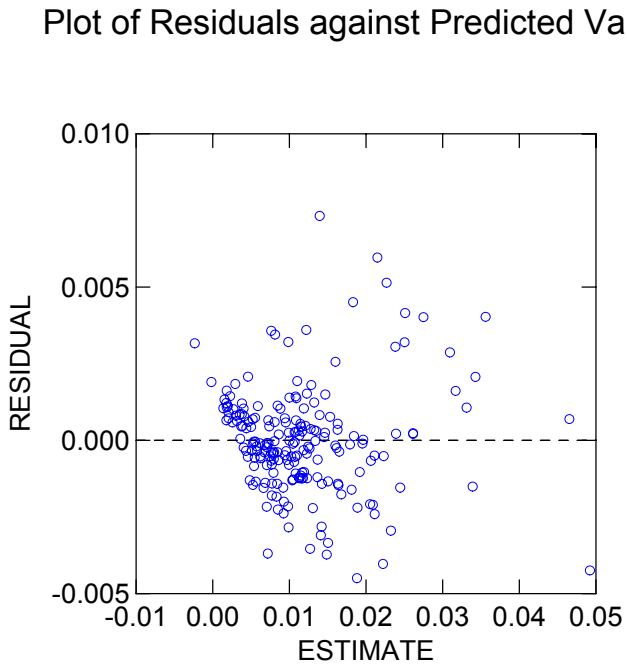

Gráfico 03. Distribuição residual para o Método do Coeficiente de Correlação Múltipla.

Observa-se que a distribuição residual deste método é a que apresenta menores valores residuais, quando comparados com os procedimentos anteriores. Isto pode ser explicado pela introdução da variável V1 na equação resultante, pois esta variável independente foi a que apresentou maior coeficiente de correlação com a variável dependente (0.974).

4.3. Análise de agrupamento

Foram consideradas duas técnicas de seleção de variáveis independentes: método da ligação máxima e ligação média.

a) Método da ligação máxima

Os resultados da formação de grupos estão na tabela 11.

Tabela 11. Formação de grupos no método da ligação máxima

Etapas	Variáveis agrupadas em cada passo
1	V06. V10
2	V02. V06. V10
3	V02. V06. V07, V10
4	V02. V06. V07, V10,V13
5	V02. V06. V07, V10,V13, V15
6	V02. V06. V07, V10,V11, V13, V15
7	V12. V14
8	V03, V08
9	V02. V03. V06. V07, V08. V10, V11, V13, V15
10	V04 . V09
11	V02. V03. V06. V07, V08. V10, V11, V12, V13, V14, V15
12	V01, V02. V03. V06. V07, V08. V10, V11, V12, V13, V14, V15
13	V01, V02. V03. V04, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
14	V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
15	C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
16	CAP, C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
17	CG, CAP, C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
18	H, CG, CAP, C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15

Como a variável NG não fez parte do agrupamento acima, ela deve constituir outro grupo, resultando em:

Grupo 1) NG

Grupo 2) CAP, H, CG, C030, V01, V02, V03, V04, V05, V06, V07, V08, V09, V10, V11, V12, V13, V14, V15

Depois de formados todos os grupos, deve-se decidir quantas variáveis irão representá-los. Existem três formas de análise:

1º) Agrupamento interno (AI). Selecionou-se uma das variáveis que, originalmente, criaram o grupo. Por esse método se escolheu as variáveis NG e V6.

A distribuição dos resíduos para o método da Ligação Máxima com Agrupamento Interno se encontra no Gráfico 04.

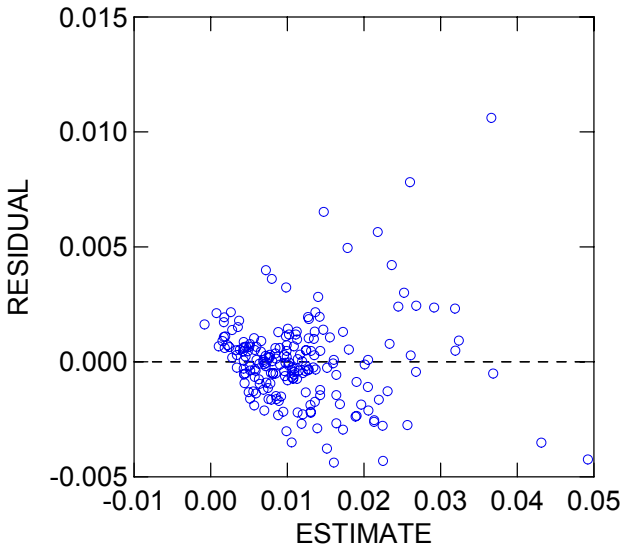

Gráfico 04. Distribuição residual para o Método de Ligação Máxima com Agrupamento Interno.

2º) Agrupamento externo (AE). Selecionou-se a última variável a entrar no grupo. Por esse método se escolheu as variáveis NG e H.

A distribuição dos resíduos para o método da Ligação Máxima com Agrupamento Externo se encontra no Gráfico 05.

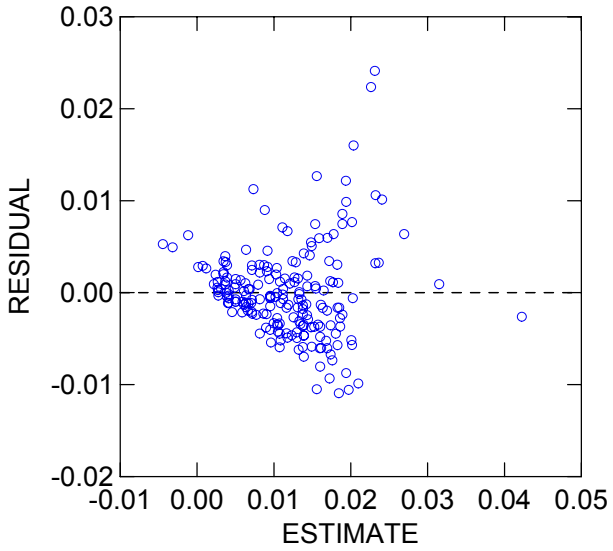

Gráfico 05. Distribuição residual para o Método de Ligação Máxima com Agrupamento Externo.

3º) Agrupamento aleatório (AA). Seleciona-se qualquer uma das variáveis que pertence ao grupo. Neste caso foi escolhida a variável NG que deve ser associada a uma outra do grupo 2.

A distribuição dos resíduos para o método da Ligação Máxima com Agrupamento Aleatório se encontra no Gráfico 06.

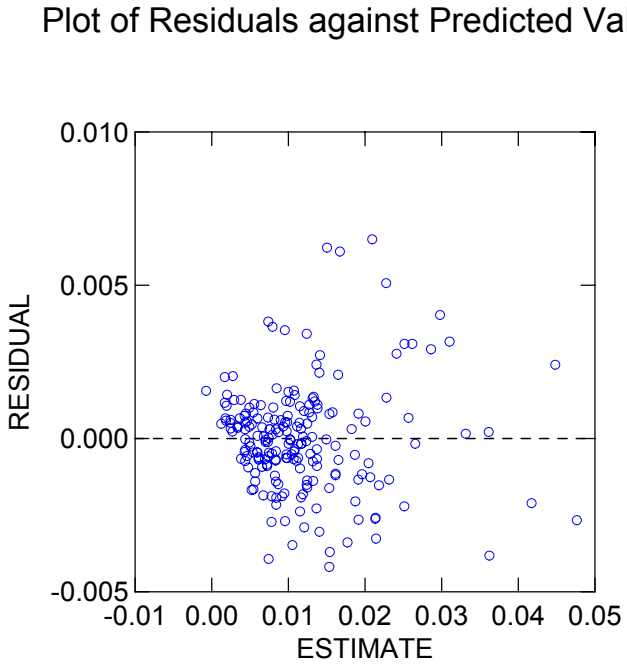

Gráfico 06. Distribuição residual para o Método de Ligação Máxima com Agrupamento Aleatório.

Semelhante aos resultados dos Componentes Principais, o Método da Ligação Máxima com Agrupamento Aleatório apresentou menores valores residuais por conter a variável V1.

b) Método da ligação média

Os resultados da formação de grupos estão na tabela 12.

Tabela 12. Formação de grupos no método da ligação média

Etapas	Variáveis agrupadas em cada passo
1	V06. V10
2	V02. V06. V10
3	V02. V06. V07, V10
4	V13, V15
5	V11,V13, V15
6	V02. V06. V07, V10,V11, V13, V15
7	V03, V08
8	V04, V09
9	V02. V03. V06. V07, V08. V10, V11, V13, V15
10	V12 . V14
11	V02. V03. V04,V06. V07, V08. V09. V10, V11, V13, V15
12	V01, V02. V03. V04. V06. V07, V08. V09. V10, V11, V13, V15
13	V05. V12, V14
14	C030. V01, V02. V03. V04, V06. V07, V08. V09, V10, V11, V13, V15
15	CAP, V05, V12, V14
16	CAP, C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15
17	CAP, CG, C030. V01, V02. V03. V04, V05, V06. V07, V08. V09, V10, V11, V12, V13, V14, V15

Os seguintes grupos foram formados:

Grupo 1) H

Grupo 2) NG

Grupo 3) CAP, CG, C30, V01, V02, V03, V04, V05, V06, V07, V08, V09, V10, V11, V12, V13, V14, V15

Depois da formação dos grupos, deve-se decidir quantas variáveis irão representá-los. Semelhante ao procedimento anterior, tem-se:

1º Agrupamento interno (AI). Selecionam-se as variáveis que, originalmente, criaram o grupo, assim sendo, escolheu-se: H, NG e V6.

A distribuição dos resíduos no Método da Ligação Média com Agrupamento Interno se encontra no Gráfico 07.

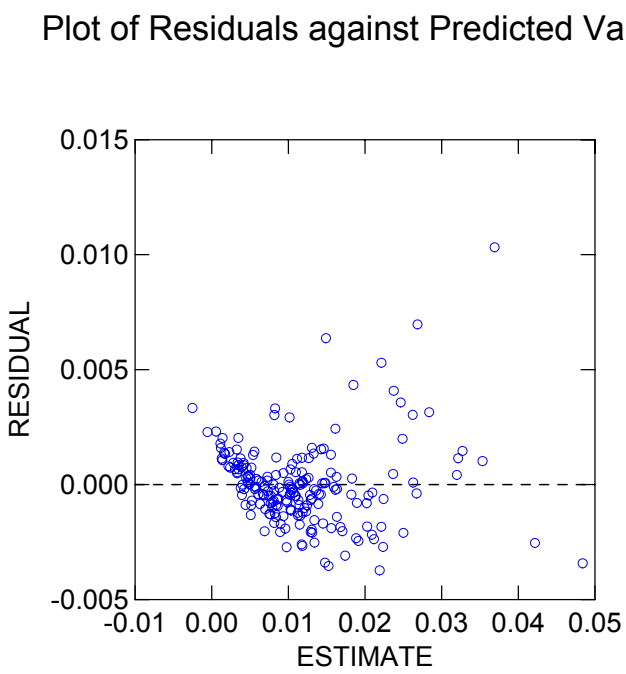

Gráfico 07. Distribuição residual do Método de Ligação Média com Agrupamento Interno.

2º Agrupamento externo (AE). Seleciona-se a última variável a entrar no grupo. Escolheram-se as variáveis H, NG e CG.

A distribuição dos resíduos no Método da Ligação Média com Agrupamento externo se encontra no Gráfico 08.

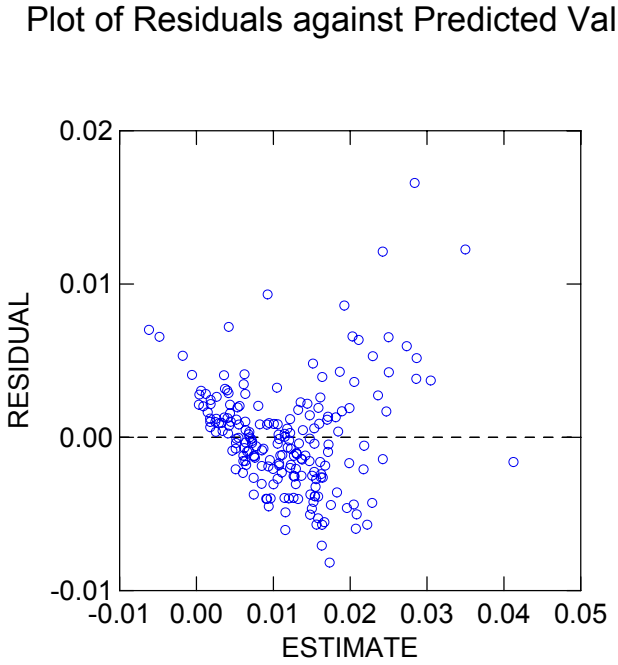

Gráfico 08. Distribuição residual do Método de Ligação Média com Agrupamento externo.

3º Agrupamento aleatório (AA). Seleciona-se qualquer uma das variáveis que pertencem ao grupo. Selecionaram-se as variáveis H, NG e que pode ser associada a qualquer outra do grupo 3.

A distribuição dos resíduos para o Método da Ligação Média com Agrupamento Aleatório se encontra no Gráfico 09.

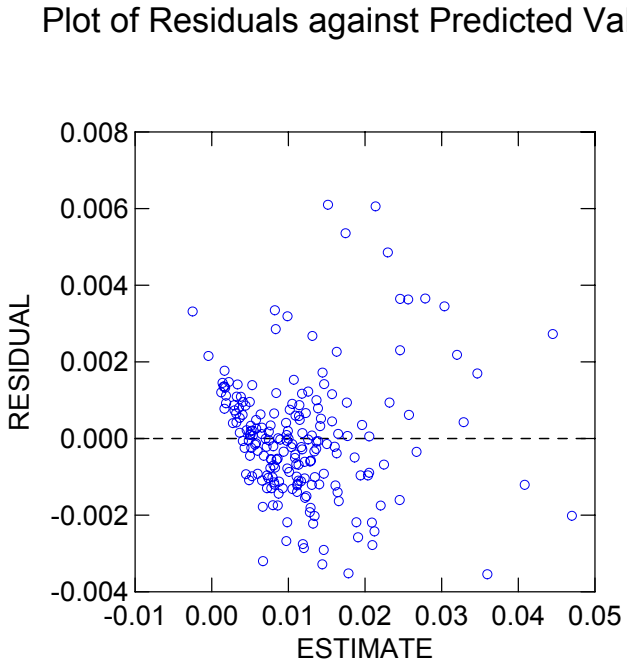

Gráfico 09. Distribuição residual do Método de Ligação Média com Agrupamento Aleatório.

4.2. R^2 Múltiplo

Neste procedimento, o resultado final sempre é a equação composta de todas as variáveis independentes. Inicia com os modelos envolvendo uma variável sendo que o primeiro modelo testado é aquele composto da variável independente com maior grau de associação com a variável dependente, neste caso, a variável V1.

Para o R^2 Múltiplo foram necessárias 20 etapas, sendo que abaixo, encontram-se as etapas 1 (inclusão de 1 variável independente); 2 (inclusão de 2 variáveis independentes); 19 (inclusão de 19 variáveis independentes) e 20 (inclusão de todas as variáveis).

Etapa	R^2	Variável na equação
1	0.9477	V1
1	0.9401	V6
1	0.9397	V15
1	0.9354	V13
1	0.9346	V2
1	0.9339	V14
1	0.9337	V10
1	0.9333	V12
1	0.9301	V7
1	0.9284	V8
1	0.9281	V11
1	0.9280	V3
1	0.9175	V9
1	0.9145	V4
1	0.9067	C30
1	0.8902	V5
1	0.8171	CAP
1	0.6886	CG
1	0.4461	H
1	0.3294	NG

Etapa	R^2	Variáveis na equação
2	0.9536	H V1
2	0.9535	NG V1
2	0.9532	CG V1
2	0.9504	V1 V12
2	0.9503	V1 V15
2	0.9502	V1 V14
2	0.9498	CG V15
2	0.9493	NG V14
2	0.9489	V1 V8
2	0.9488	NG V15
2	0.9488	V1 V11
2	0.9487	V1 V13
2	0.9485	V1 V2
2	0.9485	V1 V6
2	0.9482	V1 V3
2	0.9482	V1 V7
2	0.9482	V1 V4
2	0.9480	V1 V10
2	0.9480	V1 V5
2	0.9480	C30 V1

.
.
.

Etapa	R^2	Variáveis no modelo
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V11 V12 V13 V14 V15
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V13 V14 V15
19	0.9768	CAP H NG CG C30 V1 V2 V3 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V9 V10 V11 V12 V13 V14 V15
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14
19	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V15
19	0.9768	CAP H NG CG C30 V1 V2 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9767	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V10 V11 V12 V13 V14 V15
19	0.9767	CAP H NG CG C30 V1 V2 V3 V4 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9767	CAP H NG CG C30 V1 V2 V3 V4 V5 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9766	H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9766	CAP H NG CG C30 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9765	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V14 V15
19	0.9765	CAP H NG CG C30 V1 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9765	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V12 V13 V14 V15
19	0.9743	CAP H NG CG V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9722	CAP H NG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9701	CAP NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9682	CAP H CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15

Etapa	R^2	Variáveis na equação
20	0.9768	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15

Uma vantagem desse método é que permite ao pesquisador escolher a equação que melhor o satisfaça em termos de custos (praticidade da coleta de dados) e precisão.

4.5. Cp de MALLOWS

Processo semelhante ao anterior, só que o critério de seleção desta vez é o valor do Cp de Mallows, cuja ordem de seleção de variáveis independentes se dá para a equação que apresentar um menor valor do Cp.

Abaixo seguem 4 etapas do Cp de Mallows.

Etapa	R^2	Cp	Variável na equação
1	0.9477	208.9657	V1
1	0.9401	268.1157	V6
1	0.9397	271.1731	V15
1	0.9354	304.8147	V13
1	0.9346	310.8521	V2
1	0.9339	316.3419	V14
1	0.9337	318.0617	V10
1	0.9333	321.1671	V12
1	0.9301	345.8937	V7
1	0.9284	359.3495	V8
1	0.9281	361.1281	V11
1	0.9280	361.8800	V3
1	0.9175	443.4649	V9
1	0.9145	467.1325	V4
1	0.9067	527.8438	C30
1	0.8902	655.6679	V5
1	0.8171	1223.616	CAP
1	0.6886	2222.062	CG
1	0.4461	4105.506	H
1	0.3294	5011.604	NG

Etapa	R^2	Cp	Variáveis na equação
2	0.9536	165.4293	H V1
2	0.9535	165.8857	NG V1
2	0.9532	168.6365	CG V1
2	0.9504	190.2569	V1 V12
2	0.9503	191.4232	V1 V15
2	0.9502	192.0733	V1 V14
2	0.9498	195.3004	CG V15
2	0.9493	199.1171	NG V14
2	0.9489	201.6392	V1 V8
2	0.9488	202.5381	NG V15
2	0.9488	203.0423	V1 V11
2	0.9487	203.2287	V1 V13
2	0.9485	204.9535	V1 V2
2	0.9485	205.0153	V1 V6
2	0.9482	207.0552	V1 V3
2	0.9482	207.6376	V1 V7
2	0.9482	207.6769	V1 V4
2	0.9480	208.6333	V1 V10
2	0.9480	209.1996	V1 V5
2	0.9480	209.2066	C30 V1

Etapa	R^2	Cp	Variáveis na equação
19	0.9768	19.0087	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V11 V12 V13 V14 V15
19	0.9768	19.0150	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V13 V14 V15
19	0.9768	19.0153	CAP H NG CG C30 V1 V2 V3 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9768	19.0515	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9768	19.1487	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V9 V10 V11 V12 V13 V14 V15
19	0.9768	19.2524	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14
19	0.9768	19.2812	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V15
19	0.9768	19.3606	CAP H NG CG C30 V1 V2 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9767	19.7055	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V10 V11 V12 V13 V14 V15
19	0.9767	19.7109	CAP H NG CG C30 V1 V2 V3 V4 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9767	19.7542	CAP H NG CG C30 V1 V2 V3 V4 V5 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9766	21.1056	H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9766	21.1287	CAP H NG CG C30 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9765	21.4394	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V14 V15
19	0.9765	21.6608	CAP H NG CG C30 V1 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9765	21.6837	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V12 V13 V14 V15
19	0.9743	38.5899	CAP H NG CG V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9722	54.8327	CAP H NG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9701	71.4541	CAP NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
19	0.9682	86.2911	CAP H CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15

Etapa	R^2	Cp	Variáveis na equação
20	0.9768	21.0000	CAP H NG CG C30 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15

A distribuição dos resíduos para os Métodos de R^2 (Máximo e Mínimo) e Cp Mallows se encontra no Gráfico 10.

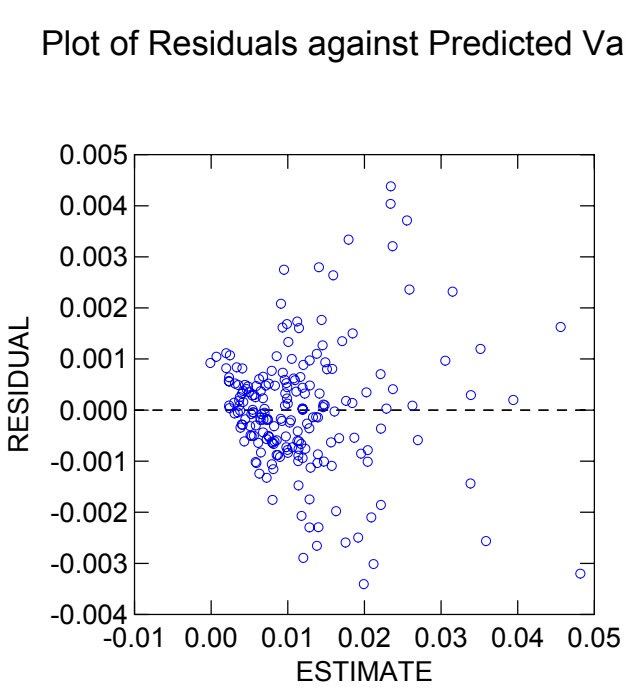

Gráfico 10. Distribuição residual para os Métodos de R^2 (Máximo e Mínimo) e Cp Mallows.

Observa-se que o Gráfico 10 é o que apresenta menores valores residuais, mas a justificativa se dá ao fato de que na equação final todas as variáveis independentes estão presentes, quer tenham coeficientes significativos ou não.

4.6. Busca direta de t

Os resultados para as estimativas dos parâmetros (coeficientes), associados aos valores de erro padrão, teste de t e níveis de probabilidades estão nas Tabela 13.

Tabela 13. Estatísticas para as variáveis independentes na equação completa

Efeitos	Coeficientes	Erro Padrão	t	Probabilidade
CONSTANTE	0.00130	0.00128	1.01558	0.31119
CAP	0.02622	0.02078	1.26162	0.20872
H	0.00097	0.00013	7.48546	0.00000
NG	0.00060	0.00007	8.45597	0.00000
CG	0.02978	0.00478	6.23194	0.00000
C30	0.09957	0.02291	4.34636	0.00002
V1	29.23912	19.76767	1.47914	0.14085
V2	27.82791	26.68653	1.04277	0.29845
V3	22.24157	26.99966	0.82377	0.41116
V4	10.78824	26.50820	0.40698	0.68451
V5	17.71794	19.85286	0.89246	0.37334
V6	7.24736	22.04189	0.32880	0.74269
V7	25.61419	23.80269	1.07610	0.28332
V8	3.12497	23.13208	0.13509	0.89269
V9	12.25611	20.93985	0.58530	0.55908
V10	10.54054	21.71652	0.48537	0.62800
V11	63.94891	23.52979	2.71778	0.00721
V12	10.82607	22.17373	0.48824	0.62598
V13	49.90933	22.55946	2.21235	0.02820
V14	53.69221	21.99185	2.44146	0.01560
V15	39.33287	23.10295	1.7025(	0.09039

O quadro da análise da variância se encontra na Tabela 14.

Tabela 14. Quadro da análise da variância para a equação com todas as variáveis

Fontes de Variação	GL	Soma de Quadrados	Quadrado Médio	F	P
Regressão	20	0.01284	0.0006400	391.60858	0.00000
Resíduo	180	0.00029	0.0000016		
TOTAL	200	0.01313			

R^2 = 0.97753

R^2 ajustado = 0.97504

Pelo teste de t, as variáveis H, NG, CG, C30 e V11 devem ser mantidas a nível de 1% de probabilidades na equação final. Observa-se que variáveis que possuem altos graus de correlações com a variável dependente, podem ser excluídas da equação resultante pelo teste de t. Por exemplo, a variável V1 que é a mais correlacionada com a variável dependente foi excluída da equação resultante.

Este problema é um dos problemas da aplicação do teste de t na seleção de variáveis independentes em modelos de regressão linear, principalmente, devido a multicolinearidade entre variáveis independentes.

A distribuição dos resíduos para a equação envolvendo as 5 variáveis independentes selecionadas (H, NG, CG, C30 e V11) se encontra no gráfico 11.

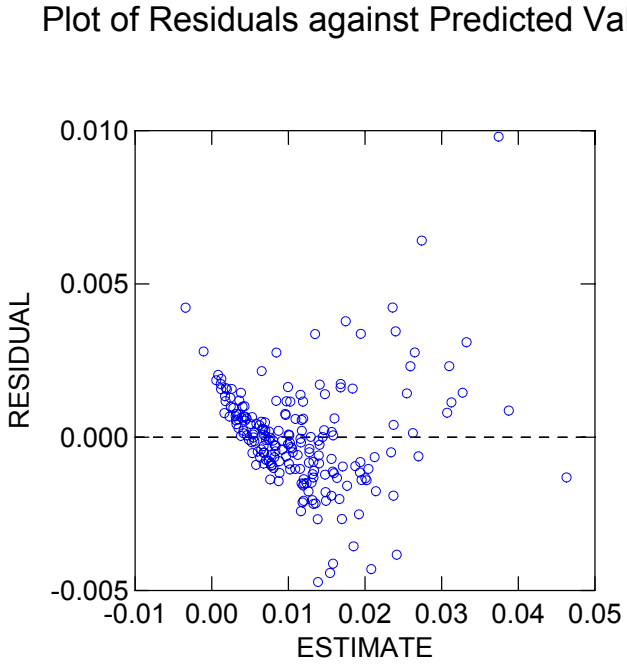

Gráfico 11. Distribuição residual para o Método da Busca de t.

4.7. Incremento de R^2 máximo e mínimo

São métodos de que visam incrementar o R^2 através do acréscimo e troca de variáveis em modelos propostos, sendo no R^2 o processo inicia com a variável que produz um maior R^2. No método do R^2 mínimo o processo inicia com a variável que produz o menor R^2. Por esta razão é um processo que envolve um maior número de etapas. O processo para quando todas as variáveis estão envolvidas na equação resultante.

Os resultados para ambos os métodos estão nas Tabelas 15 e 16.

Tabela 15. Exemplo de algumas etapas no processo de R^2 máximo

Etapa	Variável incluída	R^2	Cp de Mallow
1	V1	0.9477	208.9857
2	H	0.9536	165.4293
.	.	.	.
.	.	.	.
.	.	.	.
27	V10	0.9768	21.0000

Tabela 16. Exemplo de algumas etapas no processo de R^2 mínimo

Etapa	Variável incluída	R^2	Cp de Mallow
1	V1	0.3294	5011.604
2	H	0.4261	4105.506
.	.	.	.
.	.	.	.
.	.	.	.
243	V10	0.9768	21.0000

Ambos processos terminam quando a equação com todas as variáveis independentes estão incluídas na equação final. Existe uma grande diferença no número de total de etapas entre os dois procedimentos.

Na realidade esses procedimentos podem ajudar o pesquisador porque mostram um grande número de subconjuntos de equações com respectivos R^2, dando oportunidade da escolha de equações com boas precisões e variáveis fáceis de serem mensuradas.

A distribuição dos resíduos para os dois métodos acima se encontra no Gráfico 10, pois contém todas as variáveis independentes.

4.8. Stepwise

Este procedimento envolveu seis etapas (Tabela 17)

Tabela 17. Etapas executadas no processo Stepwise

Etapa	Variável incluída	R^2 da equação	Cp de Mallow	F	Probabilidade
1	V1	0.9477	208.966	3608.53	<.0001
2	H	0.9536	165.429	25.02	<.0001
3	NG	0.9609	111.063	36.52	<.0001
4	CG	0.9652	78.9456	24.77	<.0001
5	V14	0.9675	63.0979	13.81	0.0003
6	CAP	0.9726	25.8327	35.79	<.0001

As estatísticas para as variáveis independentes na equação resultante estão na Tabela 18.

Tabela 18. Estatísticas para as variáveis independentes na equação completa

Efeitos	Coeficientes	Erro Padrão	F	Probabilidade
CONSTANTE	-0.0053200	0.0007942	44.91	<.0001
CAP	-0.0412400	0.0068900	35.79	<.0001
H	0.0009160	0.0001302	49.47	<.0001
NG	0.0005961	0.0000715	69.43	<.0001
CG	0.0256900	0.0045900	31.32	<.0001
V1	3.8986400	0.5857000	44.31	<.0001
V14	2.3998200	0.3770800	40.50	<.0001

Observa-se que a primeira variável a ser incluída foi V1, exatamente, a que apresenta maior grau de associação com a variável dependente e que todas as variáveis independentes foram incluídas na equação final ao nível de 1% de probabilidade.

4.9. Forward

Os resultados do procedimento Forward foram iguais ao Stepwise.

A distribuição dos resíduos para os métodos Stepwise e Forward se encontra no Gráfico 12.

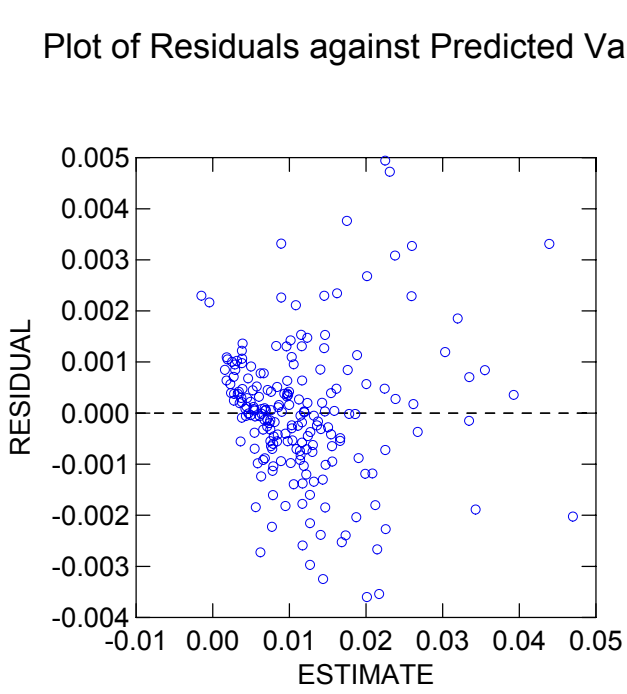

Gráfico 12. Distribuição residual para os Métodos Stepwise e Forward.

Observa-se que o gráfico 12 apresenta valores residuais menores que todos os métodos considerados e bem próximos dos que consideram todas as variáveis independentes.

4.10. Backward

Este procedimento envolveu sete etapas (Tabela 19)

Tabela 19. Etapas executadas no processo Backward

Etapa	Variável excluída	R^2 da equação	Cp de Mallow	F	Probabilidade
1	V10	0.9768	19.0087	0.01	0.9257
2	V12	0.9768	17.0179	0.01	0.9237
3	V4	0.9768	15.0359	0.02	0.8928
4	V7	0.9768	13.2048	0.17	0.6791
5	V14	0.9768	11.4479	0.25	0.6189
6	V5	0.9767	10.1014	0.67	0.4141
7	CAP	0.9764	10.3900	2.35	0.1269

As estatísticas para as variáveis independentes na equação resultante estão na Tabela 20.

Tabela 20. Estatísticas para as variáveis independentes na equação resultante

Efeitos	Coeficientes	Erro Padrão	F	Probabilidade
Constante	-0.00142	0.00122	1.36	0.2443
H	0.0009849	0.0001270	60.13	<.0001
NG	0.0005956	0.0000714	69.61	<.0001
CG	0.02899	0.00474	37.37	<.0001
C30	-0.07585	0.01201	39.86	<.0001
V1	9.03391	0.95177	90.09	<.0001
V2	-44.73535	23.08073	3.76	0.0541
V3	-2.94106	1.49853	3.85	0.0512
V6	32.96763	17.43257	3.58	0.0602
V8	2.77371	1.40416	3.90	0.0497
V9	-5.07018	0.84477	36.02	<.0001
V11	30.02396	13.83288	4.71	0.0312
V13	-26.64066	11.69502	5.19	0.0239
V15	4.62677	0.66718	48.09	<.0001

No processo Backward, 13 variáveis foram incluídas na equação resultante, entretanto o aumento em termo de R^2 foi de apenas 0.38% com relação ao processo Stepwise que envolveu apenas 6 variáveis, o que o caracteriza como sendo superior em termos de custos e precisão.

A distribuição dos resíduos para o método Backward se encontra no Gráfico 13.

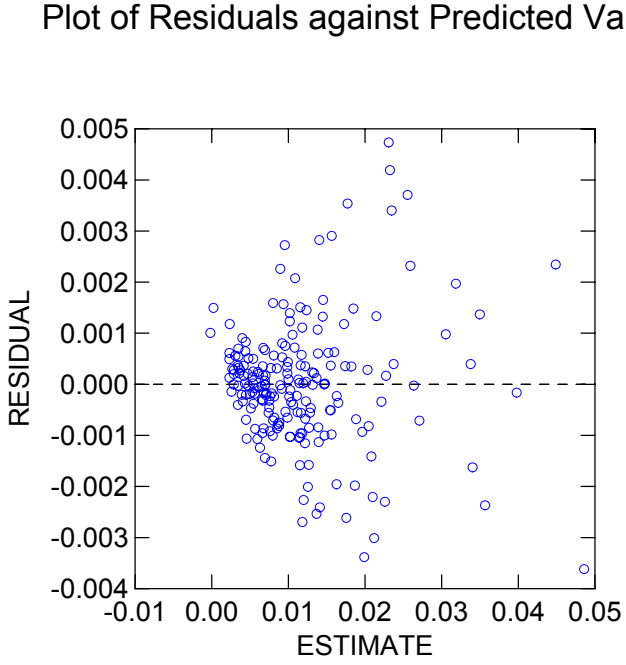

Gráfico 13. Distribuição residual para o Método Backward.

Observa-se que a distribuição dos resíduos é muito semelhante aos métodos Stepwise e Forward, mesmo que neste caso houvesse um maior número de variáveis independentes selecionadas.

4.11. Critério de Akaike

Na realidade o critério de Akaike é uma outra forma de descartar variáveis através dos processos Stepwise, Forward e Backward.

a) Stepwise

As estatísticas para as variáveis independentes na equação resultante estão na Tabela 21.

Tabela 21. Estatísticas para as variáveis independentes na equação resultante

Efeitos	Coeficientes	Erro Padrão	t	Probabilidade
Constante	-0.001632	0.0012040	-1.355	0.17689
H	0.001037	0.0001257	8.250	<0.0001
NG	0.000560	0.0000698	8.599	<0.0001
CG	0.029660	0.0046870	6.328	<0.0001
C30	-0.077050	0.0118400	-6.509	<0.0001
V1	8.874000	0.8737000	10.157	<0.0001
V3	-3.255000	1.2520000	-2.601	0.01004
V9	-4.910000	0.8319000	-5.902	<0.0001
V11	55.790000	18.6000000	2.999	0.00307
V12	2.134000	0.9890000	2.158	0.03220
V13	-46.420000	15.5800000	-2.979	0.00327
V14	-43.990000	15.5300000	-2.832	0.00513
V15	40.700000	13.3300000	3.052	0.00260

AIC= -2667.57

R^2= 0.9770

R^2 ajustado = 0.9753

A distribuição dos resíduos para o Método Stepwise usando a metodologia de Akaike se encontra no Gráfico 14.

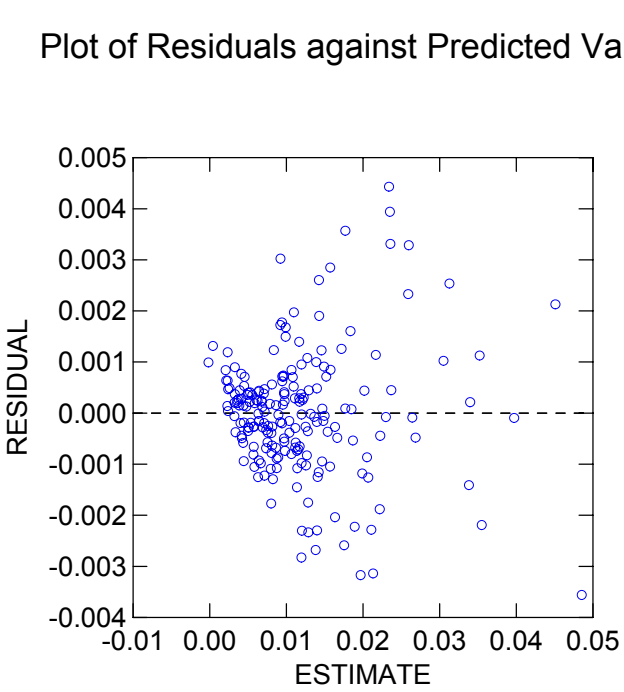

Gráfico 14. Distribuição residual do Método Stepwise (Akaike).

b) Forward

As estatísticas para as variáveis independentes na equação resultante estão na Tabela 22.

Tabela 22. Estatísticas para as variáveis independentes na equação resultante

Efeitos	Coeficientes	Erro Padrão	t	Probabilidade
Constante	-0.001192	0.001235	-0.965	0.3357
V1	11.420000	2.070000	5.516	<0.0001
H	0.000988	0.000128	7.705	<0.0001
NG	0.000578	0.000069	8.344	<0.0001
CG	0.029090	0.004733	6.147	<0.0001
V14	-19.890000	14.120000	-1.409	0.1606
CAP	0.027010	0.019760	1.367	0.1733
C30	-0.100600	0.021600	-4.657	<0.0001
V9	-7.088000	1.751000	-4.049	<0.0001
V7	33.730000	21.060000	1.601	0.1110
V15	22.770000	11.940000	1.907	0.0580
V10	-30.560000	16.730000	-1.827	0.0693

AIC= -2665.24

R^2= 0.9761

R^2 ajustado = 0.9749

A distribuição dos resíduos para o Método Forward usando a metodologia de Akaike se encontra no Gráfico 15.

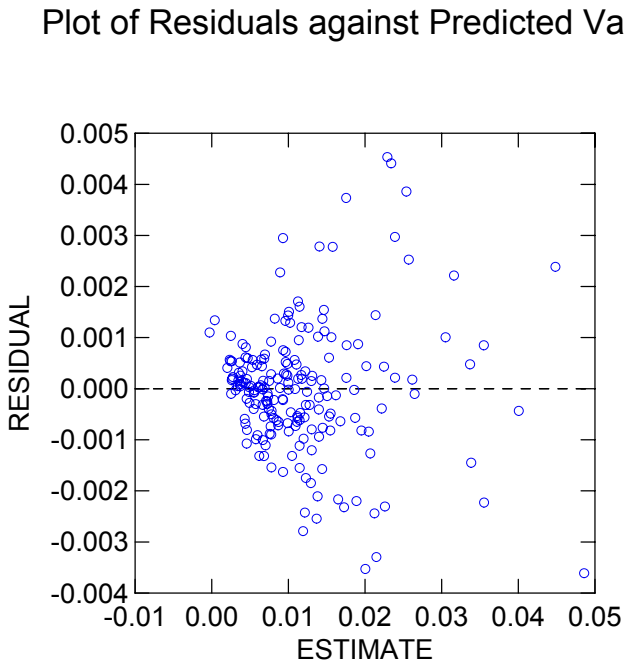

Gráfico 15. Distribuição residual do Método Forward (Akaike)

c) Backward

Resultados semelhantes ao do Stepwise.

Observa-se que os resultados provenientes do Critério de Akaike para os processos de Stepwise, Forward e Backward envolveram mais variáveis nas equações finais, sendo que os aumentos de precisão nos R^2 não chegaram sequer a 1.00%, o que do ponto de vista prático não recomenda o uso de tal critério.

Como foram muitos métodos analisados, faz-se necessário um resumos dos mesmos para que se possa ter idéia do comportamento da seleção de variáveis nos diferentes métodos utilizados. Tal informação se encontra na Tabela 23.

Tabela 23. Resultados das variáveis retidas nos modelos pelos diversos métodos usados

MÉTODOS	VARIÁVEIS RETIDAS NAS EQUAÇÕES FINAIS	R^2
COMPONENTES PRINCIPAIS		
Retenção de k componentes	NG, V2	0.9429
Beale	NG, V2	0.9429
Beale Interativo	NG, H	0.6214
Correlação múltipla	H, V1	0.9536
ANÁLISE DE AGRUPAMENTOS		
LIGAÇÃO MÁXIMA		
Agrupamento interno	NG, V6	0.9443
Agrupamento externo	NG, H	0.6214
Agrupamento aleatório	NG, V1	0.9535
LIGAÇÃO MÉDIA		
Agrupamento interno	H, NG, V6	0.9512
Agrupamento externo	H, NG, CG	0.8130
Agrupamento aleatório	H, NG, V1	0.9608
BUSCA DE t	H, NG, CG, C30, V11	0.9552
STEPWISE	V1, H, NG, CG, V14, CAP	0.9725
FORWARD	V1, H, NG, CG, V14, CAP	0.9725
BACKWARD	H, NG, CG, C30, V1, V2, V3, V6, V8, V9, V11, V13, V15	0.9764
AKAIKE		
Stepwise	H, NG, CG, C30, V1, V3, V6, V11, V12, V13, V14, V15	0,9770
Forward	V1, H, NG, CG, V14, CAP, C30, V9, V7, V15, V10	0.9761
Backward	H, NG, CG, C30, V1, V3, V6, V11, V12, V13, V14, V15	0,9770

Observa-se que os métodos R^2 Múltiplo, Critério Cp de Mallows e Incremento de R^2 Máximo e Mínimo não constam da tabela acima, pelo fato de que os mesmos sempre conduzem à equação final com todas as variáveis independentes. Entretanto, esses métodos podem ser úteis porque eles, na execução de etapas, analisam vários subgrupos de equações que podem auxiliar pesquisador na seleção da equação final, uma vez que nem sempre a melhor equação baseada em algum critério estatístico

é a mais eficiente, já que a equação ideal é aquela que é precisa, com poucas variáveis independentes e de fáceis mensurações.

Na Tabela 23, os métodos de Beale Interativo e o de Ligação Máxima por Agrupamento Externo foram os que apresentaram o menor resultado de R^2 0.6214 que é baixo quando comparado com os demais. As duas equações resultantes desses modelos retiveram as mesmas variáveis NG e H.

Entretanto, analisando a matriz de correlação entre as variáveis independentes (Tabela 2), observa-se que essas duas variáveis retidas nas equações são as que apresentaram o menor coeficiente de correlação entre todas variáveis consideradas r_{ij}=0.253. Portanto, esses dois métodos, não violam o princípio de independência entre as variáveis independentes, mas por outro lado, apresentam respostas de baixa precisão quando comparadas com as demais, e isso não interessa ao pesquisador.

Para os outros métodos, há uma grande variabilidade de seleção de variáveis, sendo que todas as equações apresentaram valores de R^2 próximos, mesmo com estruturas diferentes. Isso se explica pelo fato de que na pesquisa em questão existem 1048576 possíveis modelos (método de todas as possibilidades), e, certamente, existe uma infinidade de equações com variáveis independentes completamente diferentes que podem dar, exatamente, o mesmo resultado. Nessas situações, a análise dos gráficos dos resíduos com relação aos valores estimados é de grande valia para o pesquisador, pois permite decidir entre duas equações com mesmos R^2, a que melhor distribui seus erros (resíduos).

Assim sendo, fica claro que o julgamento subjetivo do pesquisador é um critério que deve ser levado em conta na hora de decidir sobre a seleção de variáveis em modelagem matemática, uma vez que a estatística e os processos matemáticos tratam todas as variáveis como sendo iguais em termos de custos e facilidade de mensuração.

No caso deste trabalho, a variável altura (H), frequentemente usada na construção de tabelas volumétricas, não se mostrou tão eficiente como as variáveis volumétricas tomadas em diferentes alturas acessíveis no tronco da árvore. Também mensurar alturas de árvores em povoamentos florestais é uma tarefa muito difícil e, geralmente, envolve muitos erros.

Depois da análise de todos os métodos considerados, recomenda-se que o modelo $\mathbf{V_{tot} = \beta_0 + \beta_1 V1 + \xi_i}$ seja o indicado, pois basta se medir duas circunferências no tronco da árvore, a 0,30 e 0,90 m no tronco das árvores, medidas de fácil obtenção, que permite aumentar o tamanho da amostra e, consequentemente, diminuir o erro de amostragem.

A equação gerada por esse modelo, apresenta um R^2=0.9477, enquanto que a equação contendo todas as variáveis independentes apresenta um R^2=0,9775, um acréscimo de 2.98%, mas são 20 variáveis, enquanto que o modelo acima só possui uma variável independente.

Vale salientar que os dois modelos, tradicionalmente usados na Engenharia Florestal, $\mathbf{V_{tot} = \beta_0 + \beta_1 CAP^2 + \xi_i}$ (Spurr) e $\mathbf{V_{tot} = \beta_0 . CAP^{\beta 1} . H^{\beta 2} . \xi_i}$ (Schumacher e Hall), apresentaram equações com coeficientes de determinação de 0.864 e 0.876, respectivamente. O modelo indicado neste estudo também foi o selecionado por Ribeiro (2001), para o mesmo conjunto de dados usado neste estudo.

5. CONCLUSÕES

- No geral, os métodos univariados e multivariados empregados na seleção de variáveis independentes para modelos volumétricos, conduzem a respostas semelhantes, mesmo que possuam estruturas diferentes em relação às variáveis independentes, desde que o número dessas variáveis seja elevado.

- Os métodos que mais seguiram os requisitos de análise da variância para modelos de regressão foram os que apresentaram piores resultados, pois as variáveis independentes selecionadas (NG e H) foram aquelas que apresentaram menores coeficientes de correlação com a variável dependente e entre elas.

- Para construção de tabelas volumétricas para *Leucaena leucocephala* a variável altura (H) pode ser substituída com ganho de precisão por medidas de volumes parciais tomadas no tronco das árvores.

- Nos processos de seleção de variáveis, o julgamento do pesquisador sobre a relevância das variáveis descartadas ou adicionadas nas equações selecionadas é de grande importância, principalmente, na redução de custos e do erro de amostragem.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H.(1974). **A new look at the statistical identification model***,* IEEE, Trans. Auto. Catl. AC-19, n^{o} 6, 716-723.

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

ALCÂNTARA, P.B. Recursos genéticos em leguminosas arbóreas arbustivas. **In:** SIMPÓSIO SOBRE USOS MÚLTIPLOS DE LEGUMINOSAS ARBÓREAS E ARBUSTIVAS, 1,1993. Nova Odessa. Anais. Nova Odessa: Instituto de Zootecnia, 1993, p. 1-29.

ASSOCIAÇÃO PARANAENSE DE ENGENHEIROS FLORESTAIS - APEF. **Curso de atualização em manejo florestal**. Curitiba, 1987. p. 30-55.

BARROSO, G. M. **Sistemática das angiospermas do Brasil**. Viçosa: UFV. 1984, v.2, p 70-77.

BEALE, E.M.L. et al*.* (1967). The discarding of the variables in multivariate analasys*.* **Biometrika**, 54, 357-365.

BREWBAKER, J. L.; ITO, G. M. Taxonomic studies of the genus *Leucaena.* **Leucaena Newsletter**, Taipei. v. 1, p. 41-42, 1980.

BREWBAKER, J. L. Can there be such a thing as a perfect tree? **Agroforestry Today**, Nairobi, v. 1, n. 4, p. 4-7, 1989.

BREWBAKER, J. L. Revisions in the systematics of the genus *Leucaena.* **Leucaena Research Reports,** Taipei, v. 6, p. 78-80, 1985.

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

CAMPELLO, F. B. et al. **Diagnóstico florestal da região Nordeste**. Brasília: IBAMA, 1999. 20 p. (Boletim Técnico, n. 2).

CARVALHO, R.F. Sugestões para escolha de espécies de florestais destinadas à experimentação e plantio na Região Nordeste. **Brasil Florestal**, Brasília, n. 33, p. 45-48, 1978.

CLARO, D.P. et. al. 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. Jan/Jun, 1998, p 35-44.

CLUTTER, J. L., at al. **Timber management: A quantitative approach**. New York: John Wiley e Sons, 1983. 333 p.

CORDEIRO, G.M.; PAULA, G.A. Modelos de regressão para análise de dados univariados. **In:** 17^0 Colóquio Brasileiro de Matemática. IMPA, Rio de Janeiro, 1989, 353p.

CORDEIRO, G. M.; NETO, E. A. L. **Modelos paramétricos**, Associação Brasileira de Estatística. 2004. 246.p.

CRUZ, C. D.; REGAZZI, A. J. **Métodos biométricos aplicados ao melhoramento genético***.* Viçosa: UFV, 2001, 390 p.

CUNHA, L. S. *Leucaena*: a árvore milagrosa de grande futuro energético para o Brasil. **Jornal dos Reflorestadores**, v. 1, n. 4, p. 17-19, 1979.

DIAS FILHO, M. B.; SERRÃO, E. A. S. **Introdução e avaliação de leguminosas forrageiras na região de Paragominas**, Pará: EMBRAPA - CPATU, 1982. 18 p. (Circular Técnica, 29).

DRAPER, N.; SMITH, H. **Applied regression analysis**. 2 ed. New York: John Wiley & Sons. 1981. 709 p.

DRUMOND, M. A. Caracterização de hortos caseiros mistos na região de Petrolina-PE. Brasil. **In:** CONGRESSO BRASILEIRO SOBRE SISTEMAS AGROFLORESTAIS, 1., 1994. Porto Velho. **Anais...** Colombo: EMBRAPA-CNPF, 1994. v. 2, p. 321-326. (Documentos, 27).

DRUMOND, M. A; COUTO, L. Uso da agrossilvicultura em áreas degradadas na região Nordeste. **In:** Congresso Brasileiro sobre Sistemas Agroflorestais, 1. 1994. Porto Velho. Anais. Colombo. EMBRAPA-CNPF, 1994, v. 2, p. 279-284. (Documentos, 27).

EDWARDS, A.W.F., CAVALLI-SFORZA, L.L. A method for cluster analysis. **Biometrics**, North Carolina, v. 21, p. 362-375, 1965.

EMBRAPA. **Programa Nacional de Pesquisa Florestal.** Relatório Técnico Anual. Brasília: EMBRAPA, 1981. 64 p.

FERREIRA, A.F. et al. Utilização de técnicas multivariadas na avaliação da divergência genética entre clones de palma forrageira (*Opuntia fícus-indica* Mill.), **Rev. Bras. Zootec.,** v.32, n.6, p.1560-1568, 2003

FIERROS – GONZALEZ, A. M. et al. Site index for *Pinus caribaea* var. *hondurensis* in 'La Sabana' Oaxac, México**. Commonwealth Foresty Review**, Volume 71(1), 1992, p. 47 – 51.

FINGER, C. A. G. **Fundamentos de biometria florestal**. Santa Maria: UFSM. 1992. 269 p.

FRANCO, A. A.; SOUTO, S. M. ***Leucaena leucocephala:*** **uma leguminosa com múltiplas utilidades para os trópicos**. Rio de Janeiro: EMBRAPA, 1986. 7 p. (Comunicado Técnico, 2).

FREESE, F. **Linear regression methods for forest research,** U.S. Forest Service Research Paper 17, 138 p., 1964.

FREITAS, A.R. et. al. ***Leucaena leucocephala*** **(Lam.) de Wit.: cultura e melhoramento**. São Carlos: EMBRAPA - UEPAE, 1991. 93 P. (Documentos, 12)

GALVÃO, A.P.M. **Reflorestamento de propriedades rurais para fins produtivos e ambientais: um guia para ações municipais e regionais**. Brasília: EMBRAPA, 2000. 351 p.

GOLFARI, L.; CASER, R. L. **Zoneamento ecológico da região Nordeste para experimentação florestal**. Belo Horizonte: PNUD, 1977. 116 p. (Série Técnica, 10).

HOFFMANN, R. **Componentes principais e análise fatorial**. 4. ed. Piracicaba: ESALQ/USP, 1999. 40 p. (Série didática, nº 90).

HUSCH et. al. **Forest mensuration,** Second edition, The Ronald Press Company, 1972, 409 p.

JOHNSON, R. A; WICHERN, D. W. **Applied Mu1ltivariate Statistical Analysis**, 1982, 584 p.

JOLLIFE, I. T. Discarding variables in a principal component analysis. i. artificial data**. J**. **Royal Stat. Soc. , Series C, Appl. Stat.,** 21:160 173, 1972.

JOLLIFE, I. T. Discarding variables in a principal component analysis. **ii**. artificial data. **J. Royal Stat. Soc. , Series C, Appl. Stat.,** 22:21 31, 1973.

JUNIOR, R.C.B.S. **Aplicação de modelos matemáticos na estimativa de crescimento de *Leucaena leucocephala* (Lam.) de Wit, no Agreste de Pernambuco**, 2005, 85f, Dissertação (Mestrado em Ciência Florestal) DCFL/UFRPE, Recife – Pernambuco.

KLUTHCOUSKI, J. **Leucena: uma alternativa para a pequena e média agricultura**. 2 ed. Brasília: EMBRAPA, 1982. 12 p. (Circular Técnica, 6).

LA CHENBRUC, P. A. **Discriminant analyssis**. New York: Hafner Press, 1975. 128 p.

LE ROUX, N. J.; STELL. S. J.; LOUW, N. Variable selection and error rate estimation in discriminant analysis. **J. Statist. Comput. Simulation**, v. 59, p. 195-219, 1997.

LEWIS-BECK, M.S. **Applied regression. An introduction.** Sage University Paper 22, Series: Quantitative Applications in the Social Sciences. 1980, 77p.

LIMA, P.C.F. Programa Regional de Pesquisa Florestal do Tópico Semi-árido. **In:** Congresso Florestal Brasileiro, 3,1978. Manaus. **Anais**... Manaus: SBS, 1978, p. 392-393.

LIMA, P.C.F. **Comportamento de *Leucaena leucocephala* (Lam.) de Wit. comparado com *Prosopis juliflora* (SW) DC e *Eucalyptus alba* Reinw ex Blume em Petrolina (PE), região semi-árida do Brasil**. 1982. 96 f. Dissertação (Mestrado em Eng. Florestal - Silvicultura) - Universidade Federal do Paraná, Curitiba, Paraná.

LIMA, P. C. F. Usos múltiplos da leucena: produtividade no semi-árido brasileiro. **In**: CONGRESSO FLORESTAL BRASILEIRO, 5., 1986, Olinda. Silvicultura, SBS, v. 11, n. 41, p. 55-57, 1986. Edição especial.

MACEDO, R. L. G. CAMARGO, I. P. Sistemas agroflorestais no contexto do desenvolvimento sustentável. **In:** CONGRESSO BRASILEIRO SOBRE SISTEMAS AGROFLORESTAIS, 1., 1994. Porto Velho. Anais... Colombo: EMBRAPA - CNPF, v. 2, p. 43-49, 1994.

MACHADO, S. A. Tabela de volume para *Pinus taeda* na região de Telêmaco Borba - PR. **Revista Floresta.** Curitiba, v. 10, n. 1, 1979.

MALLOWS, C. L. Some comments on Cp**. Technometrics**, 15:661 675, 1973.

McCABE, G. P. Principa variables. **Technometrics***,* v. 26, n. 2, p. 137-144, 1984.

McKAY, R. J.; CAMPBELL, N. A. Variable selection techniques in discriminant analysis. II. Allocation. **Bristish J. Math. Statist. Psych.,** v. 35, p. 30-41, 1982*.*

MEDRADO, M. J. S. **Sistemas agroflorestais: aspectos básicos e indicações. In: Galvão, A.P.M., Coord. Reflorestamento de propriedades rurais para fins produtivos e ambientais: guia para ações municipais e regionais**. Brasília: EMBRAPA, 2000. p. 269-312.

MEUNIER, I. M. J. **Crescimento de mudas de *Leucaena leucocephala* (Lam.) de Wit. em função do uso de composto de resíduo urbano, adubação fosfatada e inoculação com *Rhizobium loti***. 1991. 110 f. Dissertação (Mestrado em Agronomia, Ciência do Solo) - Universidade Federal Rural de Pernambuco, Recife – Pernambuco.

MEUNIER, I.J.M. et al. **Inventário florestal: Programa de estudo.** UFRPE, 2002, 189p.

NETER, J.; WASSERMAN, W.; KUTNER, M. H. **Applied linear regression models***.* Homewood (H); Irwin, 1989, 667 p.

NFTA - NITROGEN FIXING TREE ASSOCIATION. **Leucaena: wood production and use**. Hawaii, 1985. 50 p.

NÓBREGA, A. A. Experimentação florestal executada pelo DNOCS na área do polígono das secas. **In**: CONGRESSO FLORESTAL BRASILEIRO, 3., 1978, Manaus. **Anais**... Manaus: SBS, 1978. p. 394-395.

PÉLLICO NETTO, S.; BRENA, D. A. **Inventário florestal**. Curitiba, UFPR/UFSM, 1997. 316 p.

PIRES, I. E.; FERREIRA, C. A. **Potencialidade do Nordeste do Brasil para reflorestamento**. Curitiba: EMBRAPA / URFCS, 1982. 30 p. (Circular Técnica, 6).

RAO, C.R. **Advanced Statistical Methods in Biometric Research,** John Wiley & Sons. 1952, 390p.

RIBEIRO, C. A. S. **Seleção de modelos volumétricos para leucena no Agreste do Estado de Pernambuco**. 2001. 57f. Dissertação (Programa de Pós-graduação em Biometria) - Universidade Federal Rural de Pernambuco, Recife - Pernambuco.

SAS INSTITUTE. **User's guide: Statistics**, Cary (NC): SAS INSTITUTE, 1982, 585p.

SCOLFORO, J. R. **Mensuração florestal 3**: **Relações quantitativas em volume, peso e a relação hipsométrica**. Lavras: ESAL, 1993. 292 p.

SEIFFERT, N. F. Produção biológica de nitrogênio e proteína bruta de acessões de *Leucaena* spp. cultivadas para emprego na suplementação protéica de ruminantes. **Pesquisa Agropecuária Brasileira, Brasília,** v. 19, p. 283-291,1984.

SCHNEIDER, P.R.; TONINI, H. Utilização de variáveis dummy em equações de volume para *Acacia mearnsii* de Wild. **Ciência Florestal,** Santa Maria, v.13, n.2., p.121-129, 2003.

SILVA, J. A. A.; SILVA, I. P. **Estatística experimental aplicada à ciência florestal**, Recife: UFRPE. 1982. 291 p.

SILVA, J. A. A. et al. Equação volumétrica para *Eucalyptus camaldulensis*, na Região de Barbalha, Ceará, usando o volume da 1ª tora como variável independente. **Revista Árvore,** Viçosa, v. 17, n. 1, p. 30-37, 1993.

SILVA, J. A. A., **Dynamics of stand structure in fistilized slash pine plantations.** The University of Georgia,(thesis of Ph. D), 1986, 139 p.

SILVA, A.P.D. Eficient variable screening for multivariate analysis. **Journal of Multivariate Analysis,** v. 76, p.35-62, 2001.

SNEATH, P. H. A; SOKAL, R. R. **Numeric taxonomy: the principles and practice of numerical classification,** San Francisco: W. H. Freemen, 1973, 573p.

SOUZA, F.B. **Leucena: produção e manejo no Nordeste brasileiro**. Sobral: EMBRAPA, 1999. 20 p. (Circular Técnica, 18).

SOUZA, G.S. **Introdução aos modelos de regressão linear e não-linear.** EMBRAPA, 2001, 489p.

SOUZA, C.M.; **Avaliação do crescimento em altura de leucena *leucaena leucocephala* (Lam.) de Wit., no agreste de Pernambuco, por meio da analise multivariada de medidas repetidas.** 2003. 123f. dissertação (Mestrado em Biometria) – UFRPE, Recife - Pernambuco.

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

SUDENE. **Recursos naturais do Nordeste: investigação e potencial.** Recife, 1972, 108p.

VEIGA, R. A. A. Tabelas de volume para *Eucalyptus saligna* Smith. Em ocasião de primeiro corte. **Revista Floresta,** Curitiba, v. 4, n. 3, p. 29-44, 1973.

VENABLES, W.N.; RIPLEY, B.D. **Modern applied statistics with S (Statistical and Computing)**, Springer Verlag, 4th Edition, 495 p. 2002.

WARD, J. H; Hierarchical grouping to optimize in objective function. **Journal of American Statistical Association,** v.58, pág. 263, 244, 1963.