VALDIR ALVES

AVALIAÇÃO DE IMÓVEIS URBANOS BASEADA EM MÉTODOS ESTATÍSTICOS MULTIVARIADOS

CAMPO MOURÃO

2005

VALDIR ALVES

AVALIAÇÃO DE IMÓVEIS URBANOS BASEADA EM MÉTODOS ESTATÍSTICOS MULTIVARIADOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, Setor de Ciências Exatas e Setor de Tecnologia, Universidade Federal do Paraná.

Orientação: Profº. Dr. Anselmo Chaves Neto.

CAMPO MOURÃO

2005

TERMO DE APROVAÇÃO

Valdir Alves

"Avaliação de Imóveis Urbanos Baseada em Métodos Estatísticos Multivariados"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Área de Concentração em Programação Matemática, Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Anselmo Chaves Neto, D.Sc.
PPGMNE / DEST- UFPR

Profª Maria Teresinha Arns Steiner, D.Eng.
PPGMNE / DMAT - UFPR

Prof. Fábio Favaretto, D.Eng.
PPGEPS / PUC-PR

Curitiba, 04 de outubro de 2005.

DEDICATÓRIA

Senhor, que as pessoas se amem se respeitem e que busquem a paz.

AGRADECIMENTOS

À minha família, pelo apoio e compreensão, em especial à minha esposa Lídia e aos meus filhos Ana Carolina e André Felipe.

A todos os alunos do curso pelo apoio e amizade e em especial aos colegas Amauri, Douglas, Flávia e Silvia pela ajuda.

Aos professores do curso e em especial à professora Neida pelas primeiras orientações.

Ao meu orientador, prof. Anselmo, que apoiou o desenvolvimento do trabalho e ofereceu todas as contribuições necessárias para sua realização.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS ..... **vii**
LISTA DE QUADROS ..... **viii**
LISTA DE TABELAS ..... **ix**
RESUMO ..... **x**
ABSTRACT ..... **xi**
1 INTRODUÇÃO **01**
1.1 TEMA DO ESTUDO 01
1.2 OBJETIVOS 03
1.2.1 Objetivo Geral **03**
1.2.2 Objetivos Específicos **04**
1.3 IMPORTÂNCIA DO TRABALHO 04
1.4 ESTRUTURA DO TRABALHO 05
2 REVISÃO DE LITERATURA **06**
2.1 INTRODUÇÃO 06
2.2 ENGENHARIA DE AVALIAÇÃO 06
2.2.1 Introdução **06**
2.2.2 Normas Técnicas **07**
2.2.3 Avaliação de Imóveis Urbanos **07**
2.2.3.1 Avaliação **08**
2.2.3.2 Valor **08**
2.2.4 Classificação dos Imóveis Urbanos **10**
2.2.5 O Mercado **11**
2.2.5.1 Mercado Imobiliário **11**
2.2.6 Características dos Imóveis **14**
2.2.7 Métodos de Avaliação **14**
2.2.8 Nível de Precisão da Avaliação **17**
2.3 ANÁLISE MULTIVARIADA 18
2.3.1 Introdução **18**
2.3.2 Estatísticas Descritivas Multivariadas **21**
2.3.3 Métodos Multivariados **23**
2.3.3.1 Análise de Componentes Principais **24**

2.3.3.2 Análise de Agrupamento (*Clusters Analysis*) 28
2.3.3.3 Discriminação, Classificação e Reconhecimento de Padrões 31
2.3.3.4 Avaliação de Funções de Reconhecimento e Classificação 38
2.4 ANÁLISE DE REGRESSÃO LINEAR 40
2.4.1 Introdução 40
2.4.2 Modelo Linear Geral de Regressão 41
2.4.3 Análise da Variância da Regressão 42
2.4.4 Verificação dos Pressupostos do Modelo 43
2.4.5 Poder de Explicação do Modelo 46
2.4.6 Relação entre Variáveis 46
2.4.7 Seleção de Variáveis Regressoras 48
3 MATERIAL E MÉTODO 51
3.1 MATERIAL 51
3.1.1 Área de Estudo 51
3.1.2 Limitação da Pesquisa 53
3.1.3 Levantamento dos Dados 54
3.1.3.1 As Variáveis Utilizadas 54
3.1.3.2 Amostra 57
3.2 METODOLOGIA DE DESENVOLVIMENTO DA PESQUISA 57
3.2.1 Considerações para a Construção do Modelo 58
3.2.1.1 Identificação das Variáveis Independentes 58
3.2.1.2 Transformações de Variáveis 59
3.2.1.3 Análise Exploratória 60
3.2.1.4 Análise dos Resíduos 60
3.2.1.5 Verificação da Adequação do Modelo 60
3.3 ESTUDO DE CASO 61
4 APLICATIVO AMI DESENVOLVIDO 62
4.1 ALGORITMO DO PROGRAMA AMI 62
4.2 AVALIAÇÃO DO AMI 66
4.3 DESCRIÇÃO DO PROGRAMA AMI 67
4.4 COMPARAÇÃO DOS RESULTADOS ENTRE O PROGRAMA STATGRAPHICS E O AMI 75
4.4.1 Resumo dos Resultados para Apartamentos 77

4.4.2 Resumo dos Resultados para Casas 78
4.4.3 Resumo dos Resultados para Terrenos 80
5 CONSIDERAÇÕES FINAIS 82
5.1 CONCLUSÕES 82
5.2 SUGESTÕES PARA FUTURAS PESQUISAS 84
REFERÊNCIAS BIBLIOGRÁFICAS 85
APÊNDICES 87

LISTA DE FIGURAS

Figura 2.1: Representação da estrutura de componentes principais 24
Figura 2.2: Matriz de confusão 39
Figura 3.1: Mapa com a localização de Campo Mourão 52
Figura 3.2: Mapa dos municípios vizinhos à Campo Mourão 53
Figura 4.1: Diagrama de fluxo do programa AMI 65
Figura 4.2: Entrada de dados usando um arquivo do *Excel* 67
Figura 4.3: Mostrando os tipos de opções de cálculo para a regressão 67
Figura 4.4: Com a opção 1 mostrando os tipos de distância e ligações 68
Figura 4.5: Dendrograma apresentado para a escolha de número de grupos 68
Figura 4.6: Grupos formados 69
Figura 4.7: Processo 2: análise de Componentes Principais para o grupo 1 69
Figura 4.8: Perguntando o número de componentes e o processo de regressão 70
Figura 4.9: Calculando o valor da observação do 45° apartamento 71
Figura 4.10: Estimando o valor da última observação 71
Figura 4.11: Analisando os apartamentos por Análise Fatorial 72
Figura 4.12: AMI perguntando o número de fatores 72
Figura 4.13: Regressão com Análise Fatorial 73
Figura 4.14: Regressão considerando todas as variáveis 74
Figura 4.15: Resultados da regressão 74

LISTA DE QUADROS

Quadro 2.1: Análise de Variância (ANOVA) 43
Quadro 3.1: Variáveis independentes para apartamento 55
Quadro 3.2: Variáveis independentes para casas residenciais 56
Quadro 3.3: Variáveis independentes para terrenos 57

LISTA DE TABELAS

Tabela 4.1: Comparação dos valores estimados pelas 3 opções 75
Tabela 4.2: Comparação dos valores estimados pelas 3 opções 80
Tabela 4.3: Comparação dos valores estimados pelas 3 opções 81

RESUMO

Neste trabalho são mostrados conceitos de Estatística Multivariada e também apresenta um programa computacional denominado AMI (Análise Multivariada de Imóveis), em MATLAB. Tal programa tem por finalidade avaliar imóveis urbanos considerando cada imóvel como um vetor de características aleatórias e construção de um modelo estatístico que estima o valor de um novo imóvel de maneira automática, evitando subjetividades de avaliadores. O programa oferece ao usuário três opções de regressão: **1)** Análise de Agrupamento (*Clusters Analysis*) e Componentes Principais aplicadas às características dos imóveis para obtenção de classes homogêneas, também oferece 5 (cinco) tipos de distâncias e 4 (quatro) tipos de ligações (oferecendo a melhor opção de ligação conforme o índice cofenético) e essas variáveis foram reduzidas e transformadas de modo a evitar problemas de multicolinearidade, comuns a esse tipo de dados. Assim, as novas variáveis foram utilizadas para a determinação dos modelos de Regressão Linear Múltipla de cada classe ou grupo homogêneo de itens e para cada tipo de imóveis (apartamento, casa e terreno). **2)** Análise Fatorial que explica as correlações entre as variáveis originais em termos de um conjunto de poucas variáveis não observáveis, faz a rotação dos fatores (Varimax) para evitar a multicolinearidade. **3)** Regressão múltipla simples trabalha com todas as variáveis, sem nenhum tratamento. Essa opção não é confiável, pois os dados não sofreram nenhum tratamento para evitar problemas de multicolinearidade. As três opções apresentam a função de regressão, os parâmetros estatísticos (R^2, F e p) e o valor estimado para um novo imóvel. O programa computacional desenvolvido, bem como sua proposta de avaliação, foi aplicado a um conjunto de dados oriundos da pesquisa realizada por Silvia Neide Bráulio (2004), referentes a 44 apartamentos, 51 casas e 24 terrenos da cidade de Campo Mourão – PR, podendo ser aplicado em outras localidades e também a outros produtos que necessitem de avaliação. O modelo de cada classe dos três tipos de imóveis avaliados apresentou um bom ajuste aos dados e uma boa capacidade preditiva, atendendo todas as suposições teóricas de regressão linear múltipla.

Palavras-chave: avaliação de imóveis, regressão linear múltipla, análise de agrupamento, análise fatorial.

ABSTRACT

In this work concepts of Multivariate Statistics are shown and it also presents a computational program denominated MAI (Multivariate Analyses of Immobiles), in MATLAB ®.That program has the purpose to evaluate immobile urban considering each immobile one as a vector of aleatory characteristics and construction of a statistical model that esteems the value of a new property in an automatic way, avoiding appraisers' subjectivities. The program offers to the user three regression options: 1) Clusters Analysis and Main Components applied to the characteristics of the properties for obtaining of homogeneous classes, it also offers 5 (five) types of distances and 4 (four) types of connections (offering the best connection option according to the cophenet index) and those variables were reduced and transformed in way to avoid multicolinearity problems, common to that type of data. Like this, the new variables were used for the determination of the models of Multiple Linear Regression of each class or homogeneous group of items and for each type of properties (apartment marries and land). 2) Factorial Analysis that explains the correlations among the original variables in terms of a group of little you varied you didn't observe, it makes the rotation of the factors (Varimax) to avoid the multicolinearity. 3) Simple Multiple Regression uses all the variables, without any treatment. That option is not reliable, because the data didn't suffer any treatment to avoid multicolinearity problems. The three options present the regression function, the statistical parameters (R^2, F and p) and the esteem value for a new one immobile. The program developed for computer, as well as your evaluation proposal, the was applied a group of data originating from of the research accomplished by Silvia Neide Bráulio (2004), referring to 44 apartments, 51 houses and 24 lands of Campo Mourão's city - PR, it could be applied at other places and also the other products that need evaluation. The model of each class of the three types of appraised properties presented a good adjustment to the data and a good capacity predictable, assisting all the theoretical suppositions of multiple linear regression.

Word-key: evaluation of properties, multiple linear regression, clusters analysis, factorial analysis.

1 INTRODUÇÃO

1.1 TEMA DO ESTUDO

A avaliação de imóveis, urbanos e rurais, é utilizada na grande maioria dos negócios, discussões e pendências interpessoais e sociais em toda e qualquer comunidade. Ocorre em geral na compra ou na venda de casas, lojas comerciais, instalações industriais, aluguéis, na reavaliação de ativos de empresas, em atendimento à legislação vigente, na partilha oriunda de heranças, meações ou divórcios, no lançamento de impostos, nas hipotecas imobiliárias, nas divergências que originam ações demarcatórias, possessórias, nas indenizações, nas desapropriações e servidões, enfim, em um número expressivo de ações oriundas de problemas inerentes aos relacionamentos humanos, onde o valor de um bem assume importância fundamental.

Toda vez que uma empresa necessita de um empréstimo para investimento na construção de instalações ou capital de giro recorre a bancos de desenvolvimento. E esses bancos atendem a essas solicitações desde que o solicitante apresente garantias reais. Entende-se por garantias reais imóveis ou máquinas e equipamentos. Sendo assim, toda vez que se faz um investimento com apoio em bancos de desenvolvimento há necessidade de emissão de um laudo de avaliação do imóvel, máquina ou equipamento. Neste trabalho será estudado o caso de avaliação de imóveis urbanos.

Apesar do conceito de valor ser de difícil definição, sujeito e suscetível às mudanças filosóficas, tornam-se importante no relacionamento humano e social adotar critérios para que se exerça um caráter de justiça em sua aplicação prática. Assim, um trabalho de avaliação imobiliária constitui-se de uma seqüência de operações que resultam no que poderia ser chamado de uma "formação de juízo" sobre o valor de um imóvel ou um direito sobre ele.

A norma brasileira NBR5676 (ABNT, 1989) trabalha com o conceito de que o valor é aquele fornecido para um dado instante, único, não importando qual a finalidade da avaliação. Esse valor corresponde ao preço que se definiria, para um determinado imóvel, em um mercado de concorrência perfeito, sujeito às seguintes premissas:

a) Homogeneidade dos bens levados a mercado;

b) Números elevados de compradores e vendedores (o mercado não pode por eles ser alterado);

c) Sem influência externa;

d) Conhecimento pleno e absoluto sobre o mercado, sobre os bens e das tendências de avaliação por parte dos compradores e vendedores;

e) Vendedores e compradores oferecendo liquidez com liberdade plena de entrada e saída do mercado.

A partir destas considerações, pode-se afirmar que a avaliação passa a ser uma determinação técnica do valor ou de um direito sobre o imóvel. Dentre outros fatores deve-se levar em conta que o valor de um bem está diretamente ligado à sua capacidade de produzir renda, sua utilização potencial, o atendimento de uma necessidade ou a sua raridade. Os imóveis urbanos podem ser definidos como bens que não são móveis, localizados nas cidades, geralmente classificados como glebas urbanizáveis, áreas ou lotes e terrenos com benfeitorias (casas, prédios residenciais, prédios comerciais, galpões e outros). A NBR 5676-ABNT (1989) define como sendo uma gleba urbanizável, "uma grande extensão de terreno passível de receber obras e infraestruturas urbanas, por sua localização, seus aspectos físicos, sua destinação legal e pela existência de um mercado comprador".

Entretanto, em grande parte dos mercados imobiliários, a base de cálculos que estima os valores dos imóveis, seja para a venda ou para cálculo de impostos, é feita de forma subjetiva, ou seja, sem nenhum procedimento científico.

Assim, neste trabalho desenvolveu-se um programa computacional que, a partir das características particulares dos imóveis, estima o seu valor de forma objetiva e dentro da realidade de mercado. Nesse sentido, as características e os atributos dos imóveis constituem os dados pesquisados e que exercem influência na formação do valor. Estes dados devem ser tratados por inferência estatística, respeitados os níveis de rigor definidos por norma técnica, e transformados em informação de mercado.

O nível de rigor corresponde à precisão do trabalho e será tanto maior quanto menor for a subjetividade contida na avaliação. O rigor de uma avaliação está condicionado à

pesquisa efetivada, à confiabilidade dos dados coletados e à qualidade do modelo aplicado no processo de avaliação. A norma NBR 5676 classifica o trabalho nos seguintes níveis: expedito, normal, rigoroso ou rigoroso especial.

As variáveis utilizadas na inferência estatística merecem destaque e, para cada tipo de problema devem ser classificadas, estudadas e aceitas através de trabalho estatístico. Assim, por exemplo, para se avaliar um lote urbano devem-se levar em conta algumas variáveis tais como: a dimensão de testada, a profundidade, a área total, a localização, o uso do solo, as posturas municipais, o zoneamento urbano, as distâncias a pólos que os valorizem ou os desvalorizem, a taxa de ocupação, a topografia, a suscetibilidade a enchentes ou a danos ambientais, o padrão de construções na vizinhança, a infra-estrutura urbana, a paisagem visual a partir do imóvel. Estas e outras variáveis permitem ao final a determinação do valor unitário do terreno pesquisado com relação à sua área total.

Estas variáveis devem ser ponderadas para que conceitos estatísticos possam ser aplicados com vistas à determinação do melhor modelo de ajuste. Embora muitas publicações tratem deste assunto, poucas dão um tratamento matemático rigoroso ao problema, e dificilmente empregam métodos estatísticos multivariados. É importante salientar que no tópico referente à detecção dos erros grosseiros, trabalham somente com critérios práticos, não aplicando testes estatísticos. Assim, visando dar um tratamento mais adequado a este tipo de erro está sendo proposta, no presente trabalho, a aplicação dos métodos estatísticos multivariados juntamente com o programa computacional desenvolvido no *Software Matlab 7.0*.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um programa computacional que avalie imóveis urbanos utilizando estatísticas multivariadas e considerando cada imóvel como um vetor de características aleatórias.

1.2.2 Objetivos Específicos

O objetivo geral pode ser atingido desde que se alcancem os seguintes objetivos específicos:

1. Elaborar uma rotina de programa que construa classes homogêneas das particularidades de cada tipo de imóvel.
2. Elaborar uma rotina para reduzir o número de variáveis a serem analisadas e que facilite a interpretação, sem grande prejuízo de informações.
3. Construir modelos de regressão múltipla.
4. Interpretar os resultados obtidos.

1.3 IMPORTÂNCIA DO TRABALHO

Uma considerável parcela de bens públicos, de pessoas físicas ou jurídicas, consiste de bens imóveis. A amplitude desse recurso primordial da sociedade cria a necessidade de informes avaliatórios como suporte para tomadas de decisões referentes ao uso e disposições desses bens (Abunahman, 1998 citado por Bráulio, 2005). Imóveis são dados como garantias reais em empréstimos junto às instituições financeiras.

A necessidade de estabelecer valor é uma das atividades mais freqüentes do homem moderno, sendo os imóveis um de seus bens mais valorizados. Então, a avaliação de imóveis é importante nas operações de compra e venda locações, garantias de financiamentos, apólices de seguro, instalações comerciais e industriais, cobrança de impostos, ou seja, atividades cotidianas inerentes ao relacionamento humano.

O valor de um imóvel pode ser subjetivo dependendo das circunstâncias e modo de avaliação. Por outro lado, o valor de um imóvel é dado pela soma dos valores da edificação e pelo valor do terreno, que depende da sua localização. No entanto, o valor de um bem não pode ser confundido com o preço do bem, que representa a quantidade de dinheiro paga pelo mesmo, o que depende da lei da oferta e da procura.

Nesse sentido, houve, portanto, a necessidade de se desenvolver um programa computacional que torne mais precisas as formas de avaliação de um imóvel, aproximando ao máximo de seu valor de mercado, haja vista que em Campo Mourão (local do presente estudo), assim como em outras cidades, essas avaliações são feitas de forma subjetiva e esse trabalho visou facilitar e melhorar a qualidade das avaliações. Isso pode se dar com o uso adequado do programa no tratamento das características desses imóveis através da inferência estatística, da regressão múltipla e das técnicas de Análise Multivariada.

1.4 ESTRUTURA DO TRABALHO

Com o objetivo de facilitar o entendimento do leitor, de forma clara e objetiva, esta pesquisa está estruturada de tal maneira que no primeiro capítulo encontram-se os objetivos e o verdadeiro significado da escolha do tema. No segundo capítulo tem-se uma revisão de literatura envolvendo a teoria sobre a avaliação de imóveis, sua normatização e os conceitos matemáticos acerca da Análise Multivariada, que sustenta todo o desenvolvimento do programa computacional. O programa computacional desenvolvido, bem como sua implementação, se encontra no terceiro capítulo. O quarto capítulo apresenta, de forma detalhada, os resultados da utilização do programa em relação a dados reais e a análise de todo o procedimento estatístico utilizado. E finalmente, as conclusões com base no estudo realizado e sugestões para trabalhos futuros são encontradas no quinto capítulo.

2 REVISÃO DE LITERATURA

2.1 INTRODUÇÃO

O principal objetivo desse capítulo é apresentar, numa linguagem simples, um texto sobre avaliação de imóveis, dando informações sobre programação e estatística multivariada, suficientes para o entendimento do modelo estatístico proposto.

2.2 ENGENHARIA DE AVALIAÇÃO

2.2.1 Introdução

Os primeiros estudos sobre avaliação de imóveis no Brasil datam de 1918, e já em 1923 foram introduzidos novos métodos de avaliação de terrenos, que a partir de 1929 começaram a ser sistematicamente aplicados. A partir daí a engenharia de avaliação no Brasil tomou corpo e continua crescendo e evoluindo nas técnicas de avaliação. Atualmente um grande número de profissionais desenvolve estudos nesse campo, visando dar ao tema o suporte científico necessário aos métodos técnicos até então utilizados (Fiker, 1997).

Para Dantas (1998), a Engenharia de Avaliações é uma especialidade da engenharia que reúne um conjunto amplo de conhecimentos na área de engenharia e arquitetura, bem como em outras áreas das ciências sociais, exatas e da natureza. Ela tem com o objetivo determinar tecnicamente o valor de um bem, de seus direitos, frutos e custos de reprodução. As Avaliações são de grande interesse para o mercado imobiliário, tais como: bancos de crédito imobiliário, compradores e vendedores de imóveis, prefeituras, empresas seguradoras, etc.

A Engenharia de Avaliações pode ser praticada por engenheiros, arquitetos e agrônomos, cada um atuando em sua habilitação profissional, conforme normas e regulamentos do CREA, CONFEA, ABNT, leis municipais, estaduais e federais.

2.2.2 Normas Técnicas

A ABNT - Associação Brasileira de Normas Técnicas é o Fórum Nacional de Normalização. As Normas Brasileiras, cujos conteúdos são de responsabilidade dos Comitês Brasileiros (ABNT/CB) e dos Organismos de Normalização Setorial (ABNT/ONS) e das Comissões de Estudos Especiais Temporários (ABNT/CEET), são elaboradas por Comissões de Estudo (CE), formadas por representantes dos setores envolvidos, delas fazendo parte: produtores, consumidores e neutros (universidades, laboratórios e outros) (ABNT NBR 14653-2, 2004).

Em meados de 1950 surgiram as primeiras normas de avaliação de imóveis organizadas por entidades públicas e institutos. Devido à ocorrência de grande quantidade de desapropriações em São Paulo, ocasionado pela expansão da cidade e construção do metrô na década de 1960, as normas ganharam maior relevância. Porém, o primeiro anteprojeto de normas da ABNT na Engenharia de Avaliação data de 1957. Em 1977, estudos feitos por comissões de profissionais dedicados às perícias e avaliações judiciais, em essência, deram origem à primeira Norma Brasileira para Avaliação de Imóveis Urbanos, a NB-502/77 da ABNT (Dantas, 1998).

Depois de passar por uma revisão em 1989, a Norma Brasileira para Avaliação de Imóveis Urbanos foi registrada no INMETRO como NBR 5676. Os níveis de precisão foram transformados em níveis de rigor. Segue-se a ela a Norma para Avaliação de Servidões. Alguns institutos, com base na NBR 5676, produziram, paralelamente, normas específicas de forma mais detalhada observando as características de cada região. Atualmente existe uma proposta levada à ABNT para sintetizar o tema de Engenharia de Avaliações em uma norma única. Trata-se da Norma para Avaliação de Bens, formada por uma parte principal, contendo os conceitos, métodos e definições comuns a todos os bens e, nos apêndices, a parte específica para cada tipologia de bem a avaliar.

2.2.3 Avaliação de Imóveis Urbanos

Alguns conceitos pertinentes à técnica de avaliar são apresentados a seguir.

2.2.3.1 Avaliação

A Norma NB-502/89 (NBR-5676) da ABNT, de avaliação de imóveis urbanos, define avaliação como a determinação técnica do valor de um imóvel ou de um direito sobre o imóvel.

A avaliação de imóveis é a definição técnica do valor de mercado dos bens ou de direitos sobre eles. Esta definição é feita dentro de procedimentos técnicos para a realização das análises de valor. Segundo Moreira (1994), avaliar é a arte de estimar valores apropriados e específicos, em que o conhecimento técnico e o bom-senso são condições fundamentais.

Abunahman (1998) define avaliação como sendo uma aferição de vários fatores econômicos definidos em relação a propriedades descritas com data prevista, tendo como base a análise de dados relevantes.

2.2.3.2 Valor

Desde sua origem, o homem busca estabelecer preços para os bens, tais que satisfaçam sua noção de valor numa transação e de maneira que se efetivem trocas, sejam elas diretas como o escambo ou indiretas como a moeda.

O conceito de valor de um bem, de modo geral, é intuitivo e subjetivo, quer seja vendedor ou comprador deste bem, podendo variar entre os participantes de um mercado. No entanto, o preço é uma característica que representa a quantidade de dinheiro paga pelo bem.

Segundo Bráulio (2005), muitas medidas de valor podem estar relacionadas a um bem, dentre elas o custo de produção, ao qual são agregados outros custos como matéria-prima, estocagem e comercialização desde o produtor até o produto final onde será formado o preço e o valor de mercado, não havendo necessariamente uma relação matemática entre eles. No entanto, em mercados que se aproximam daquele de concorrência perfeita, os preços são estabelecidos pela lei da oferta e procura independentemente dos custos de produção. Em vista disso, no mercado considerado, o valor do bem poderá não apresentar nenhuma relação com os custos citados (podendo mesmo ser inferior).

Quando o mercado permanece estável por um longo período, o preço e a quantidade acabam sendo negociados através da oferta e procura. Toda a subjetividade e intuição que leva os participantes do mercado a tentar suprir sua satisfação, tornam-se, portanto em

quantidades vendidas e ofertadas e seus respectivos preços. Logo, o preço estabelecido pelo mercado é considerado uma representação justa do valor do bem analisado.

Nos mercados onde são efetuadas trocas indiretas, os preços (valores) são expressos em moeda corrente, podendo ser transformados em outras moedas. As avaliações pelo valor de mercado podem ser consideradas instantâneas, ou seja, são válidas, apenas, por um intervalo curto de tempo.

Encontram-se várias definições e interpretações e que são suscetíveis a mudanças sobre os conceitos de valor, valor de mercado e preço. No entanto, torna-se importante determinar alguns critérios para sua aplicação prática. Assim, um trabalho de avaliação imobiliária constitui-se de uma série de operações e etapas até que se chegue a uma definição de valor. Dentre os diversos conceitos de valor, a Norma NB-502/89 (NBR-5676) da ABNT, de Avaliação de Imóveis Urbanos, define valor como sendo aquele fornecido para um bem em um dado instante, único, não importando qual a finalidade da avaliação. Esse valor corresponde ao valor real que se definiria em um mercado de concorrência perfeita caracterizado pelos seguintes silogismos:

a) Igualdade dos bens levados ao mercado;

b) Número elevado de compradores e vendedores, não sendo o mercado alterado por eles;

c) Sem influências externas;

d) Conhecimento pleno e absoluto entre os participantes sobre o bem, o mercado e as tendências deste;

e) Os participantes oferecendo liquidez com plena liberdade de entrada e saída do mercado.

Existem casos onde a necessidade de estimar um valor acontece a nível particular. Assim, as partes envolvidas vendedor e comprador do bem, chegam ao comum acordo da quantidade necessária (moedas) em um determinado instante (Molina, 1999). Porém, quando ocorre a necessidade da determinação do valor, de uma maneira mais ampla, isto é, sem ser a nível particular, portanto ampliado a outras pessoas além dos diretamente envolvidos na transação de ordem privada ou pública, procura-se uma perspectiva técnica. Surge, então, a

"Ciência da Avaliação", ou seja, a Engenharia de Avaliações, que infere sobre o valor de um bem de forma fundamentada.

Segundo Barbosa Filho (1998) o valor de um bem antes de tudo é um fenômeno social e pode estar associado a um vetor composto por um conjunto de variáveis que abrange todas as suas características físicas, do seu entorno, da sua utilidade e dos fatores subjetivos que a própria coletividade cria no contexto em que está situado a cada instante.

Moreira (1994) afirma que valor é empregado costumeiramente em diversos sentidos. No entanto, quando é aplicado relativamente à propriedade, a palavra valor demonstra um sentido de posse, domínio ou troca de propriedade, medida em termos de uma unidade monetária. Fiker (1997) afirma que valor é a relação entre a intensidade das necessidades econômicas do homem e a quantidade de bens disponíveis para satisfazê-las, sendo determinado dependendo da oferta e da demanda do bem.

Atualmente o valor de mercado de um imóvel é atribuído pelo preço fixado pelo vendedor e comprador. Assim, não são forçados e não estão sujeitos a pressões anormais tendo pleno conhecimento das condições de compra e venda e de como este imóvel deve ser e será utilizado. Contudo, o mercado imobiliário não é, por natureza, de concorrência perfeita. Dessa forma, estima-se o preço médio de mercado, através de uma amostragem de preços que trazem todas as imperfeições deste mercado.

2.2.4 Classificação dos Imóveis Urbanos

De acordo com a Norma NB-502/89 (NBR-5676) da ABNT, os imóveis são classificados em:

a) **Quanto ao uso**

O imóvel urbano pode ser: residencial, comercial, industrial, institucional e misto.

b) **Quanto ao tipo do imóvel**

O imóvel urbano pode ser: terreno (lote ou gleba), apartamento, casa, escritório (sala ou andar corrido), loja, galpão, vaga de garagem, misto, hotéis, hospitais, cinemas e teatros, clubes e recreativos.

c) **Quanto ao agrupamento**

Os imóveis urbanos se agrupam da seguinte forma: loteamento, condomínio de casas, prédio de apartamentos, conjunto habitacional (casas, prédios ou mistos), conjunto de salas comerciais, prédio comercial, conjunto de prédios comerciais, conjunto de unidades comerciais, *shopping-centers* e complexo industrial. Este trabalho utilizou dados correspondentes apenas aos imóveis dos tipos: apartamentos, casas residenciais e terrenos.

2.2.5 O Mercado

Entende-se por mercado o local onde são efetuadas as transações comerciais envolvendo troca de bens, tangíveis ou intangíveis, ou direitos sobre os mesmos. Neste sentido, o termo mercado refere-se àquele de concorrência perfeita e contendo, em geral, as características dos bens. Os participantes do mercado o fazem voluntariamente e têm conhecimento pleno das condições em vigor; nenhum participante, sozinho, é capaz de alterar as condições estabelecidas; cada transação é feita de maneira independente das demais; o número de ofertas e/ou transações é suficientemente grande, de maneira que a retirada de uma amostra não afeta o mercado.

2.2.5.1 Mercado Imobiliário

Para Moscovitch (1997), o mercado imobiliário é a jurisdição de determinação dos preços de imóveis urbanos que, como quaisquer outras mercadorias, passam pela medida da oferta e da demanda.

De acordo com Dantas (1998), ele é formado por três secções:

a) a dos imóveis a serem vendidos;

b) a das partes que desejam vendê-los (vendedores);

c) a das partes interessadas em adquiri-los (compradores).

O mercado imobiliário pode ser subdivido ainda, em várias especialidades, tais como a de apartamentos, de casas e de terrenos que foram especificamente analisadas neste trabalho.

Cada mercado tem seu próprio comportamento e suas características específicas. No entanto, existem inúmeras divergências e desigualdades entre os imóveis, que faz o mercado imobiliário comportar-se de forma acentuadamente diferente de outros mercados de bens, devido às características especiais dos imóveis. Por sua localização fixa, qualquer alteração no ambiente provoca modificações no valor do imóvel. Como as influências não são análogas, as variações provocadas são claramente notáveis, causando progressivamente as diferenças.

Por outro lado, como todo bem econômico, a escassez relativa à lei da oferta e procura define o preço dos imóveis. Os governos e as economias globais são grandes influenciadores sobre o preço dos imóveis. E, por sua importância e significado social, as leis propiciam tratamentos especiais.

Dentro do mercado onde ocorrem as transações imobiliárias, identificam-se alguns fenômenos como o dinamismo da atividade imobiliária e o processo de estruturação interna das áreas urbanas. Existem, também, influências externas, que alteram continuamente os valores e usos do solo.

O estudo de todos estes fatores constitui o processo de formação de valores, ou seja, como os valores dos imóveis são compostos. Esses valores muitas vezes sofrem transformações como condições de mercado e também valores que são praticados, devido à falta de ordenação entre empreendedores, intermediários, poder público e também a própria população. Por se tratar de bens econômicos, todas as mudanças que causam maior ou menor disponibilidade refletem em modificações, alterações de valor. Sabe-se que muitas alterações como a oferta de crédito, a inflação, a condução da economia, as políticas fiscais, o crescimento demográfico e a confiança no governo são importantes na flutuação de preços.

No entanto, ocorrerá em certo momento um fenômeno chamado de "equilíbrio instantâneo" resultando num valor de mercado. Todas as variações ocorridas nas condições do mercado são absorvidas e armazenadas pelos imóveis, influenciando seus valores, que oscilam no tempo e no espaço e que, em última análise, são resultados da oferta e demanda por este bem. E assim, outras mudanças na oferta ou na demanda causarão novo equilíbrio, em outros níveis de preço.

Faz-se interessante observar que por existirem inúmeros fatores e influências, uma parte das variações dos valores imobiliários é considerada aleatória, podendo-se pensar no

preço final, como baseado em um valor mais provável que aumenta ou diminui por uma parcela imprevisível, de acordo com as influências casuais.

Assim, o valor de um imóvel segue o modelo estatístico,

$$Y = \mu + \varepsilon \qquad (2.1)$$

Onde: Y é o valor negociado;

μ é o valor mais provável, ou seja, $E(Y) = \mu$;

ε é a perturbação estocástica.

Segundo Bráulio (2005), num mercado de concorrência perfeita a situação ideal é aquela onde a oferta e a procura encontra-se equilibrada sendo a relação entre imóvel, vendedor e comprador que são os formadores do mercado, especialmente importantes para a construção do preço. No entanto, pode ocorrer o mercado de concorrência imperfeita, ocasionando casos de monopólio e oligopólio. Nos casos de monopólio o mercado torna-se comandado por um único vendedor, mas é um caso mais raro de ser encontrado. Entretanto é mais comum ser encontrado casos de oligopólio, onde a oferta é concentrada em um pequeno número de vendedores que também se preocupam com propagandas e qualidade dos imóveis, fazendo com que haja sempre uma alta nos preços dos imóveis. Ocorrem ainda os casos de monopsônio, existindo apenas um comprador e também oligopsônio com muitos compradores (Dantas, 1998).

O mercado imobiliário não pode ser considerado um mercado de concorrência perfeita diante de muitas análises teóricas, pois este mercado pode ser visto como uma passagem livre, que os bens são idênticos e que todos os participantes têm as informações perfeitas e que não sofrem nenhuma pressão agindo livremente. No mercado imobiliário, ocorrem fatores que o dificultam: a falta de uniformidade dos imóveis e a falta de informações são exemplos disto. Não se pode esquecer que os altos custos impossibilitam grande parte da população a participar deste mercado de compra e venda, deixando-as ligadas a locação e a financiamentos indisponíveis. Finalmente, pode-se concluir que o mercado imobiliário é um mercado de concorrência imperfeita.

2.2.6 Características dos Imóveis

Muitos fatores e fontes de divergências diferenciam os imóveis em si: vida útil, localização, singularidade, custos das unidades e tudo isto faz com que o mercado imobiliário tenha um comportamento diferente de outros mercados de bens. Esta combinação de fatores e divergências, em um dado momento, permite explicar grande parte das diferenças de valores entre os imóveis (González e Formoso, 2000).

O preço de um imóvel é conhecido integralmente. Nele estão orçados os atributos como localização, terreno, área e vida útil. Assim, os preços dos imóveis podem ser compreendidos como a soma dos produtos das quantidades de cada um desses serviços pelos seus preços implícitos (localização, área, terreno e vida útil).

Ao final, é importante verificar que, por existirem inúmeras influências, uma parte das variações dos valores imobiliários pode ser considerada aleatória, ou seja, pode-se pensar no preço final como baseado em um "valor mais provável" que é sujeito a variações, de acordo com as influências pontuais do caso e conforme o modelo (2.1).

2.2.7 Métodos de Avaliação

A escolha da metodologia mais apropriada para uma dada avaliação depende das condições atuais do mercado, do tipo de serviço a que se presta e da precisão que se deseja. No entanto, independentemente da metodologia utilizada, essa deverá apoiar-se em pesquisa de mercado e considerar os preços comercializados e/ou ofertados, bem como outros elementos e atributos que influenciam o valor (NBR-5676/90).

De acordo com Montenegro Duarte (1999), a avaliação como ciência usada para encontrar um valor, não é exata, mas pode ser altamente precisa. Deve ser objetiva e esclarecedora, identificando o bem a ser avaliado e o método a ser utilizado e quando realizada de forma científica, ou seja, baseada em teorias e métodos adequados, utiliza instrumental tecnológico próprio buscando a objetividade. Assim os resultados das estimativas realizadas por diferentes grupos de avaliadores, deverão ser próximos uns dos outros.

Desta forma, podem-se definir as metodologias de avaliação como formas de se atribuir valor a um imóvel. De acordo com a ABNT (NBR-5676/90), os métodos de avaliação

são divididos em dois grandes grupos: métodos diretos e métodos indiretos. A seguir são apresentados detalhamentos desses métodos.

a) **Métodos Diretos**

Quando o valor do resultado da avaliação independe de outros, o método é chamado de direto (Dantas, 1998). Os métodos diretos se subdividem:

- *Método comparativo de dados*

É um método que define o valor de comparação com dados de mercado que são semelhantes quanto a características peculiares ou não. É o método mais indicado para trabalhos de avaliação, porém para a sua aplicação, existe uma condição fundamental que é a existência de um conjunto de dados que possa ser utilizada, estatisticamente, como amostra de mercado imobiliário. As características e os atributos dos dados pesquisados que exercem influência na formação dos preços, devem ser ponderados por homogeneização ou por dedução estatística, respeitados os níveis de rigor definidos nessa norma.

- *Método comparativo de custo de reprodução de benfeitorias*

É o método que apropria o valor das benfeitorias, através da reprodução dos custos de seus componentes. Estes custos são compostos com base em orçamento detalhado em função da avaliação realizada. Devem ser justificados e quantificados os efeitos do desgaste físico e do obsoletismo funcional das benfeitorias.

A utilização dos métodos diretos tem preferência e sempre que existirem dados suficientes para utilização do método comparativo ele deve ser escolhido (Dantas, 1998).

b) **Métodos Indiretos**

O método é considerado indireto quando necessita de resultados provenientes de outro procedimento. Os métodos indiretos são:

- *O método da renda*

Avalia o valor do imóvel ou de suas partes componentes em função de um rendimento já existente ou previsto pelo bem no mercado, ou seja, o valor econômico do bem.

- *Método involutivo*

O valor do terreno é estimado por estudos da viabilidade técnica-econômica do seu aproveitamento, considerando como aproveitamento eficiente à realização de um empreendimento imobiliário hipotético compatível com as características do imóvel e com as condições do mercado (Moreira Filho, Frainer e Moreira 1993).

- *Método residual*

Obtém-se o valor do terreno a partir da diferença entre o valor total do imóvel e o valor das benfeitorias, levando-se em conta o fator de comercialização (Fiker, 1997).

Analisando os vários métodos de avaliação, mostrados anteriormente, pode-se observar que de uma forma, ou de outra, todos são comparativos. Sendo assim, Dantas, 1998, explica que no método comparativo comparam-se bens semelhantes; no método de custo, comparam-se os próprios custos no mercado; nos métodos da renda e involutivo compara-se à possibilidade de renda do bem; e no método residual, compara-se o nível de comercialização do mercado.

O método comparativo, quando utilizado, permite que se tenha uma estimativa mesmo com as diferentes tendências de mercado. Ele estima o valor baseado na comparação com outros semelhantes, iniciando de um grupo de dados e somando-se com outras informações resultando numa amostragem estatística de dados do mercado imobiliário. Portanto, o método comparativo de dados de mercado é o método mais utilizado e indicado para avaliações de mercado.

Na prática, geralmente ocorre a falta de algum componente que influencie no valor final do imóvel e isto faz com que a semelhança entre o imóvel avaliado e os componentes da amostra seja imperfeita e incompleta. Assim, os atributos dos dados pesquisados que influenciam o valor devem ser ponderados por homogeneização ou inferência estatística, respeitando os níveis de rigor definidos na NBR-5676/89. Com o uso de técnicas estatísticas obtém-se uma avaliação isenta de subjetividade e de grande confiabilidade (Moreira Filho, Frainer e Moreira 1993; González e Formoso, 2000).

Tabelas e também modelos de preços hedônicos, tais como a regressão linear múltipla, são utilizados para estimar valores. No entanto, ambos sofrem contestações sobre sua eficácia. As tabelas que eram tradicionalmente utilizadas são criticadas por serem imprecisas e de pouca confiabilidade. A regressão linear múltipla tem demonstrado sérios

problemas de multicolinearidade nas variáveis explicativas e também de inclusão de *outlier* na amostra. Além disso, a colinearidade dentro dos dados pode tornar a regressão linear múltipla um modelo inadequado para um mercado que requer respostas rápidas e de alta precisão, sendo aceitável apenas quando realizada por um profissional capacitado (Worzala, Lenk e Silva, 1995).

Neste trabalho procura-se evitar o problema da multicolinearidade trabalhando-se com componentes principais.

2.2.8 Nível de Precisão da Avaliação

Os níveis de precisão que caracterizam uma determinada avaliação são normatizados pela NBR-5676/90. De acordo com NBR-5676, em seu item 7, o nível de rigor almejado numa dada avaliação está diretamente relacionado com as informações extraídas do mercado. Sendo assim, a precisão do mercado será determinada por esse nível que será maior ou menor, dependendo da avaliação. Este rigor relaciona-se diretamente com a abrangência da pesquisa, à confiabilidade e adequação dos dados coletados, à qualidade do processo avaliatório e ao grau de subjetividade do avaliador. Desta forma, os trabalhos avaliatórios podem, de acordo com a norma, ser classificados como de nível de rigor expedito, normal, rigoroso e rigoroso especial.

Rigor expedito: o valor é obtido sem a utilização de qualquer instrumento matemático, apenas com o nível de conhecimento de mercado do avaliador. É muito freqüente, porém não é este o objetivo deste trabalho. A ausência de método científico determina que o valor seja atribuído através de escolha arbitrária, não caracterizando o aspecto técnico da avaliação.

Rigor normal: para esta avaliação utilizam-se métodos estatísticos e existem exigências com relação à coleta e tratamento dos dados.

Rigorosa: nestas avaliações os dados devem se basear em processos estatísticos que permitam calcular estimativas não tendenciosas do valor, tendo a total isenção de subjetividade. O resultado final da avaliação, através do tratamento estatístico adotado, deve estar contido em um intervalo de confiança fechado e com um nível de confiança máximo de 80%, desde que as hipóteses nulas sejam testadas ao nível de significância máximo de 5%.

Rigorosa especial: caracteriza-se pelo encontro de um modelo estatístico o mais abrangente possível, ou seja, que incorpore o maior número de características que contribuem para a formação do valor.

A função estimada da formação de valor deve ser eficiente e não tendenciosa. Portanto, a hipótese nula da equação de regressão deve ser rejeitada ao nível de significância máximo de 1% (ANOVA). Já as hipóteses nulas sobre os parâmetros do modelo de regressão ao nível de significância máximo de 10% para o teste unicaudal (teste "*t*") ou 5% em cada ramo do teste bicaudal. Devem ser analisadas as seguintes condições básicas referentes aos resíduos do modelo ajustado aos dados: Gaussianidade, homogeneidade da variância e independência. Desta forma os resíduos devem ser Gaussianos, independentes e identicamente distribuídos, ou seja, $\varepsilon_i \sim N(0, \sigma^2)$.

Logo, deve-se testar a Gaussianidade por um método adequado. Os mais usados são: o teste de Shapiro-Wilks e o de Kolmogorov-Smirnov. A homogeneidade da variância das variáveis pode ser verificada graficamente por meio do gráfico dos resíduos contra os valores ajustados, ou seja, $\hat{\varepsilon}_i \; x \; \hat{y}_i \; i = 1, 2, ..., n$. E, a premissa da independência dos resíduos pode ser identificada por meio do gráfico dos resíduos contra a ordem, ou seja, $\hat{\varepsilon}_i \; x_i$.

2.3 ANÁLISE MULTIVARIADA

2.3.1 Introdução

Ao se tomar uma decisão, muitos fatores costumam estar envolvidos nela; certamente nem todos têm a mesma importância. Quando a intuição é utilizada nessa tomada de decisão, nem todos os fatores costumam ser identificados, ou seja, não serão definidas as variáveis que afetam a decisão. Assim também, nota-se que um grande número de variáveis envolve os acontecimentos sejam eles culturais ou naturais.

Muitas ciências buscam conhecer a realidade e, assim, interpretar os acontecimentos e fenômenos baseados no conhecimento das variáveis intervenientes consideradas importantes nestes eventos. A finalidade da ciência é estabelecer relações e encontrar ou propor leis

explicativas. Para isso é necessário absorver todas as informações que são consideradas mais relevantes para obter o entendimento do fenômeno analisado.

São diversas as dificuldades encontradas na transformação das informações (dados) obtidas em conhecimento. Porém, a maior delas é de natureza epistemológica: a ciência que tenta representar a realidade através de modelos e teorias dos diversos ramos do conhecimento. Outra dificuldade é a aspiração de universalidade das explicações científicas, resultando desta forma, ao condicionamento da pesquisa a uma "padronização" metodológica, tendo um aspecto essencial que é a avaliação estatística das informações.

A cada dia amplia-se cientificamente a facilidade de obtenção de informações sobre acontecimentos e fenômenos que estão sendo analisados, sendo que estas informações são parcelas de dados que devem ser trabalhados e processados e então transformados em conhecimentos.

Assim, cada vez mais surge a necessidade da utilização de ferramentas estatísticas que apresentem uma visão por inteiro sobre o fenômeno analisado e que melhore as informações obtidas com uma abordagem univariada.

A designação "Análise Multivariada" corresponde a um conjunto de técnicas que utiliza simultaneamente todas as variáveis que caracterizam um item na análise estatística. Assim, ao invés de trabalhar com uma variável explicativa X ela considera o vetor $\underline{X}$ cujas componentes são variáveis aleatórias explicativas. A metodologia da Análise Multivariada detém um grande potencial de aplicação, pois facilita o entendimento do relacionamento entre as diversas variáveis aleatórias. As técnicas multivariadas, são técnicas que não tratam apenas uma dimensão de análise de dados, mas também uma escala de cruzamento entre várias dependentes ou não e também um cruzamento de dados que envolvem informações dependentes, oferecendo assim ao pesquisador, uma nova dimensão, mais abundante que normalmente em abordagem univariada.

A estatística univariada clássica fixou-se no estudo de uma única característica (ou variável) medida para um conjunto pequeno de indivíduos, desenvolvendo as noções de estimativa e de testes fundamentados em hipóteses muito restritivas. Entretanto, na prática, os indivíduos observados são freqüentemente caracterizados por um grande número de características (ou variáveis).

Os métodos de análise de dados permitem um estudo global dessas variáveis, pondo em evidência ligações, semelhanças ou diferenças. Por isso, "mergulham-se" indivíduos e variáveis em espaços geométricos, fazendo-se a máxima economia de hipóteses, e transformam-se os dados para visualizá-los num plano ou classificá-los em grupos homogêneos, perdendo o mínimo de informações.

A Análise Multivariada é a parte da estatística que estuda, interpreta e elaboram o material estatístico sobre um conjunto de $n > 1$ variáveis (quantitativa e/ou qualitativa). Os dados onde cabe uma Análise Multivariada são, portanto, de caráter multidimensional (Cuadras, 1981). Pode ser usada para a redução ou simplificação de dados, distribuição e agrupamentos, investigação da dependência entre variáveis, predição, teste de hipótese, e muitas outras.

Pla (1986) aponta alguns dos objetivos mais importantes dos métodos multivariados:

a) Encontrar a adequação de representar o universo em estudo, simplificando a estrutura dos dados. Pode-se obter através da transformação (combinação linear ou não linear) de um conjunto de variáveis interdependentes em outro conjunto independente e/ou num conjunto de menor dimensão.

b) Classificação; esta análise permite estabelecer as observações dentro de grupos ou, então, concluir que os indivíduos estão aleatórios no multiespaço, sendo também possível, alocar novos itens em grupos identificados.

c) Análise da interdependência; tem como objetivo examinar a interdependência entre as variáveis, a qual abrange desde a independência total até a colinearidade quando uma delas é combinação linear de outras ou, em termos mais gerais, é uma função $f(x)$ qualquer das outras.

De acordo com Johnson e Wichern (1998), em problemas que envolvem p variáveis ($p > 1$), tomando-se n observações de cada vetor aleatório $\underline{X}$ de dimensão p tem-se que as medidas observadas x_{ij}, com $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$, podem ser arranjadas em uma matriz de dados genérica de ordem $n \times p$, ${}_nX_p$, conforme 2.2.

$$ {}_{n}X_{p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \qquad (2.2) $$

A representação da matriz de dados correspondente a n observações do vetor $\underline{X}' = [X_1, X_2,..., X_p]$ de dimensão p, composto por p varáveis aleatórias, pode ser ${}_{n}X_{p} = (X_{ij})$. No entanto, essa matriz corresponde a uma amostra aleatória de tamanho n do vetor p-dimensional $\underline{X}$, ou seja, $[\underline{x}_1, \underline{x}_2, \underline{x}_3,..., \underline{x}_n]$.

Na grande maioria das vezes, quando se estuda áreas de pesquisas e aplicação de técnicas estatísticas, várias características (variáveis) são observadas. Essas variáveis devem ser analisadas em conjunto por não serem independentes e a Análise Multivariada é a área que trata desse tipo de análise e visa trabalhar, conjuntamente, com mais do que uma variável.

As técnicas multivariadas mais utilizadas são aquelas relacionadas ao estudo da estrutura de covariância do vetor observado, ao estudo do agrupamento de itens e ao estudo do reconhecimento de padrões e classificação. Especificamente pode-se citar: Análise de Componentes Principais, Análise Fatorial, Análise de Correlação Canônica, Análise de Agrupamento (*Cluster Analysis*) e Reconhecimento e Classificação de Padrões.

Várias são as técnicas que podem ser aplicadas aos dados. Sua utilização depende do tipo de dados que se deseja analisar e dos objetivos do estudo. Nesta pesquisa trabalhou-se com as seguintes técnicas multivariadas:

- Análise de Agrupamentos;

- Análise de Componentes Principais;

- Reconhecimento de Padrões e Classificação.

2.3.2 Estatísticas Descritivas Multivariadas

A Ciência Estatística trabalha com amostras. As informações amostrais podem ser resumidas em números sumários conhecidos como estatísticas e que podem constar nas observações multivariadas $[\underline{x}_1, \underline{x}_2, \underline{x}_3, ..., \underline{x}_n]$. As estatísticas são usadas na inferência sobre os parâmetros, ou seja, na estimação do vetor médio $\underline{\mu}$, da matriz de covariância Σ ou da matriz

de correlação ρ, entre outros. De maneira que o vetor médio populacional $\underline{\mu}$ deve ser estimado pelo vetor médio amostral, $\underline{\overline{X}}$, definido pela expressão,

$$\underline{\overline{X}} = \frac{\sum_{i=1}^{n} \underline{x}_i}{n}, \tag{2.3}$$

onde $\underline{x}_j$ com $i = 1, 2, ..., n$ corresponde às observações amostrais do vetor $\underline{X}$ e n é o tamanho da amostra observada. Outros parâmetros de uma população multivariada $f(\underline{x})$ podem ser avaliados, tais como a matriz de covariância Σ, definida por:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} \tag{2.4}$$

onde se tem na diagonal principal as variâncias das variáveis aleatórias e, fora da diagonal principal, as covariâncias entre elas. E, a matriz de correlação ρ, definida por:

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{2.5}$$

com as correlações entre as variáveis fora da diagonal principal. Então estes parâmetros, Σ e ρ, são estimados, respectivamente, pela matriz de covariância amostral S e pela matriz de correlação amostral R ou $\hat{\rho}$, cujas expressões são:

$$S = \frac{\sum_{i=1}^{n} (\underline{x}_i - \underline{\overline{x}})(\underline{x}_i - \underline{\overline{x}})'}{n-1} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix} \tag{2.6}$$

sendo s_j^2 a variância amostral da variável aleatória X_j,

$$s_j^2 = \frac{\sum_{i=1}^{n}\left(X_{ij} - \overline{X}_j\right)^2}{n-1} \tag{2.7}$$

e s_{jk} a covariância amostral entre as variáveis aleatórias X_j e X_k, ou seja,

$$s_{jk} = \frac{\sum_{i=1}^{n}\left(X_{ij} - \overline{X}_j\right)\left(X_{i1} - \overline{X}_{1j}\right)}{n-1} \tag{2.8}$$

e

$$R = \hat{\rho} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{2.9}$$

com as correlações amostrais fora da diagonal principal e dadas pelo quociente entre a covariância amostral e o produto dos desvios padrões amostrais, ou seja:

$$r_{jk} = \frac{s_{jk}}{s_j s_k} \text{ para } j \neq k. \tag{2.10}$$

Esses estimadores são os melhores para se determinar os parâmetros. Os primeiros são EUMV (Estimador Uniformemente de Mínima Variância) e o último é EMV (Estimador de Máxima Verossimilhança).

2.3.3 Métodos Multivariados

Métodos Multivariados são técnicas estatísticas importantes cujo uso está particularmente disseminado nas ciências físicas, sociais e médicas. É também complexa porque é difícil de identificar técnicas que são projetadas para estudar relações dependentes e interdependentes. Existem vários métodos de Análise Multivariada, com diversas finalidades. Portanto, é necessário saber que tipo de conhecimento se pretende gerar. Os métodos estatísticos são escolhidos de acordo com os objetivos da pesquisa. Aplicou-se neste trabalho alguns destes métodos sendo detalhados nas próximas seções.

2.3.3.1 Análise de Componentes Principais

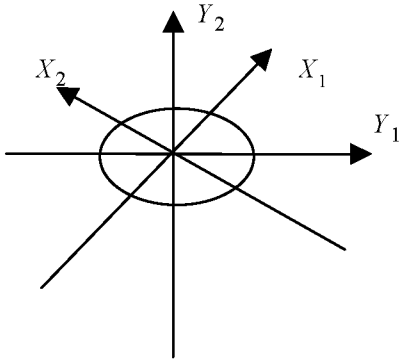
O método das Componentes Principais procura explicar a estrutura da variância e covariância de um vetor aleatório através de poucas combinações lineares das variáveis originais. Seu objetivo geral baseia-se tanto em reduzir os dados como em facilitar a interpretação, pois consiste numa transformação, de eixos, tornando as novas variáveis (combinações lineares) não correlacionadas. De forma que uma matriz de dados X de ordem n x p pode ser substituída por outra de ordem n x m sendo $m << p$.

De acordo com Johnson e Wichern (1998), embora as p componentes sejam necessárias para reproduzir toda a variabilidade presente na estrutura de covariância do vetor $\underline{X}$ de dimensão p, freqüentemente, uma grande parte desta variabilidade poderá ser explicada por um número $m < p$ de Componentes Principais. Neste caso existe praticamente a mesma quantidade de informações nas m Componentes Principais do que nas p variáveis originais. A Análise das Componentes Principais freqüentemente revela relações que não eram previamente consideradas e assim permitem interpretações que não iriam, de outro modo, aparecer.

a) Componentes Principais Populacionais

Algebricamente, as componentes principais são combinações lineares das p variáveis originais X_1, X_2,..., X_p que compõem o vetor aleatório $\underline{X}$. Geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, obtido por rotação do sistema original, sendo que os novos eixos representam as direções com variabilidade máxima. Como exemplo, tem-se, na figura 2.1, a representação da estrutura de componentes principais para $p = 2$:

Figura 2.1: Representação da estrutura de componentes principais



onde:

X_1 e X_2 são eixos originais.

Y_1 e Y_2 são novos eixos (eixos originais rotacionados: centrado na média amostral).

Obtêm-se as Componentes Principais a partir da matriz de covariância Σ ou da matriz de correlação ρ, que resumem a estrutura de relacionamento das p variáveis originais que compõem o vetor $\underline{X}$. Então, da matriz de covariância Σ ou da matriz de correlação ρ, obtém-se os autovalores $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p$ e os respectivos autovetores $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p$. E, com estes entes algébricos se constrói as combinações lineares que definem as componentes principais, ou seja,

$$Y_i = \underline{e}_i' \underline{X}, \; i = 1, 2, \ldots, p. \tag{2.11}$$

As Componentes Principais são combinações lineares, Y_i i = 1, 2,..., p, não correlacionadas, uma vez que a matriz dos autovetores P, em (2.12), é ortogonal (PP' = P'P = I),

$$P = \begin{bmatrix} e_{11} & e_{12} & \ldots & e_{1p} \\ e_{21} & e_{22} & \ldots & e_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ e_{p1} & e_{p2} & \ldots & e_{pp} \end{bmatrix}. \tag{2.12}$$

A variância da Componente Principal $Y_i = \underline{e}_i' \underline{X}$, $i = 1, 2, \ldots, p$ é dada por,

$$V(Y_i) = V(\underline{e}_i' \underline{X}) = \underline{e}_i' V(\underline{X}) \underline{e}_i' = \underline{e}_i' \Sigma \underline{e}_i' \tag{2.13}$$

e a covariância entre as componentes Y_j e Y_k é nula, ou seja, cov $(Y_j, Y_k) = 0$.

Portanto define-se:

- A primeira componente principal como a combinação linear $Y_1 = \underline{e}_1' \underline{X}$ que maximiza a variância de Y_1, sob a restrição $\underline{e}_1' \underline{e}_1 = 1$.

- A segunda componente principal como a combinação linear $Y_2 = \underline{e}_2' \underline{X}$ que maximiza $V(\underline{e}_2' \underline{X})$ sujeita a restrição $\underline{e}_2' \underline{e}_2 = 1$ e $\text{cov}(Y_1, Y_2) = 0$.

- A i-ésima componente principal como a combinação linear $Y_i = \underline{e}_i' \underline{X}$ que maximiza $V(\underline{e}_i' \underline{X})$ sujeita a restrição $\underline{e}_i' \underline{e}_i = 1$ e $\text{cov}(Y_i, Y_k) = 0 \; \forall \; k \neq i$.

b) Componentes Principais Amostrais

Comumente os parâmetros da estrutura de covariância, Σ ou ρ, são desconhecidos. Então, a obtenção das componentes principais é feita a partir de seus estimadores, que são a matriz de covariância amostral S ou a matriz de correlação amostral R. Estas estatísticas são definidas por:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' \tag{2.15}$$

$$R = D^{-1} S D^{-1} \tag{2.16}$$

onde D é a matriz desvio padrão amostral e $\overline{\underline{x}}$ é o vetor médio amostral, dados respectivamente por:

$$D = \begin{pmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & s_p \end{pmatrix} \tag{2.17}$$

$$\overline{\underline{x}} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_p \end{bmatrix} \tag{2.18}$$

Então, obtêm-se as estimativas dos elementos da estrutura de covariância do vetor aleatório $\underline{X}$, ou seja, os autovalores $\hat{\lambda}_i$ i = 1, 2,..., p e os correspondentes autovetores $\hat{\underline{e}}_i$ e constroem-se as componentes principais amostrais:

$$\hat{Y}_i = \hat{\underline{e}}_i' \underline{X},\ i = 1, 2, \ldots, p. \qquad (2.19)$$

As propriedades das componentes principais se mantêm e são obtidas com base em estimadores.

A obtenção das componentes principais com base nas informações da matriz de correlação é preferida, devido ao fato de se conseguir eliminar o efeito de escala nos valores das componentes do vetor de variáveis originais $\underline{X}$. A matriz de correlação é uma matriz de covariância, mas de variáveis padronizadas. Assim, consegue-se eliminar a influência da escala das variáveis na magnitude das variâncias.

Os autovetores definem as direções da máxima variabilidade e os autovalores especificam as variâncias. Eles são a essência do método das componentes principais. Quando os primeiros autovalores são muito maiores que os demais, a maior parte da variância total pode ser explicada por um número menor do que as p dimensões do vetor $\underline{X}$. O desenvolvimento da Ciência Estatística, com o passar do tempo, favoreceu o aparecimento de outro método de extração dos fatores de uma Análise Fatorial, que é o da máxima verossimilhança. Os dois métodos estão disponíveis nos modernos programas computacionais.

c) Critérios para Definição do Número de Componentes Principais Extraídas

Um critério para a determinação do número de componentes a serem extraídas foi sugerido por Kaiser em 1960. Segundo Johnson e Wichern (1988), Kaiser propôs escolherem-se somente as componentes correspondentes aos autovalores (raízes latentes) de magnitudes maiores do que um. Outra maneira de se definir o número de componentes é através da percentagem de variação explicada. O pesquisador, neste caso, deve julgar se m componentes explicam suficientemente o relacionamento entre as p variáveis originais. Geralmente, um bom grau de explicação é superior a 75% para um m pequeno. Um procedimento que visualiza muito bem o Critério de Kaiser é grafar os autovalores contra o número de componentes na ordem de extração (*Scree Plot*). Fixando-se um nível de corte fica fácil decidir o número de m.

Uma propriedade muito importante das Componentes Principais é a independência entre elas. Desta forma, podem substituir as variáveis originais e eliminar o problema de multicolinearidade.

2.3.3.2 Análise de Agrupamento (*Clusters Analysis*)

É uma técnica multivariada que busca a formação de grupos homogêneos de objetos ou variáveis. Estes grupos são formados calculando-se as distâncias entre os itens, representados por vetores compostos pelas suas características, construindo-se uma matriz de distâncias e juntando os itens em grupos de acordo com suas proximidades.

Segundo Crivisqui (1993) os chamados Métodos de Agrupamento, ou *Cluster Analysis*, ou ainda Métodos de Classificação Automática, são métodos estatísticos destinados a dividir em subconjuntos um conjunto de dados observados. Aplicar estes métodos significa definir nesse conjunto as classes em que se distribuem os elementos do conjunto.

Se n indivíduos sobre os quais se observaram p características estão representados num espaço de p dimensões, chamam-se classes aos subconjuntos de indivíduos desse espaço de representação que são identificáveis.

Não se pode requerer a existência de classes num conjunto de observações. Só é possível verificar a existência de níveis de síntese significativos correspondentes à organização em classes e subclasses dos elementos, de modo que os elementos de uma matriz de dados qualquer não são necessariamente classificáveis. Por isso, é necessário explorar previamente a estrutura da informação disponível, antes de orientar-se em direção a um algoritmo de classificação.

a) Medidas de similaridade e dissimilaridade

Quando são agrupados itens, a proximidade é usualmente indicada por uma espécie de distância. Por outro lado, as variáveis são usualmente agrupadas com base nos coeficientes de correlação ou outras medidas de associação. Assim, quanto maior ou menor o valor do índice de similaridade ou dissimilaridade mais ou menos parecidos são os objetos.

Existem vários índices de similaridades, sendo que a sua principal medida é o coeficiente de correlação. No entanto, os itens podem sem comparados por uma distância. Existem várias métricas que podem ser usadas:

Distância Euclidiana: *é,* simplesmente, a distância geométrica no espaço Multidimensional. A distância entre os itens $\underline{x}$ e $\underline{y}$ é definida por:

$$d(\underline{x}, \underline{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \tag{2.20}$$

Outras medidas para a distância entre $\underline{x}$ e $\underline{y}$ são definidas por:

Quadrado da distância Euclidiana:

$$d^2(\underline{x}, \underline{y}) = \sum_{i=1}^{p} (x_i - y_i)^2 \tag{2.21}$$

Distância city-block (Manhattan):

$$d(\underline{x}, \underline{y}) = \sum_{i=1}^{p} |x_i - y_i| \tag{2.22}$$

Distância de Mahalanobis (distância estatística):

$$d(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})' S^{-1} (\underline{x} - \underline{y})} = \sqrt{\frac{(x_1 - y_1)^2}{S_1^2} + \cdots + \frac{(x_p - y_p)^2}{S_p^2}} \tag{2.23}$$

Distância de Minkowski:

$$d(\underline{x}, \underline{y}) = \sqrt[n]{|x_1 - y_1|^n + |x_2 - y_2|^n + \cdots + |x_p - y_p|^n} = \sqrt[n]{\sum_{i=1}^{p} |x_i - y_i|^n} \tag{2.24}$$

b) Método de agrupamento *hierárquico*

A utilização deste método segue o seguinte procedimento: iniciam-se com g grupos, sendo que cada um é formado por um único objeto; calcula-se a matriz simétrica de distâncias $n \times n$, $D = (d_{ij})$, onde d_{ij} é a distância ou similaridade entre o objeto i e o objeto j, onde: $d_{11} = d_{22} = \ldots = d_{nn} = 0$.

Na matriz de distâncias D, acha-se o par de grupos mais próximo (menor distância) e juntam-se os grupos. O novo grupo formado é, por exemplo, (AB). Nova matriz de distâncias é construída, simplesmente apagando-se as linhas e colunas correspondentes aos grupos A e B e adicionando-se a linha e a coluna dadas pelas distâncias entre (AB) e os grupos remanescentes. Repetem-se estes passos anteriores, num total de $(g - 1)$, vezes observando-se as identidades dos grupos que são agrupados.

A função de um item ou grupo a outro grupo é feita usando-se um tipo de procedimento chamado ligação. Os tipos de ligações mais comuns são: Ligação Simples (vizinho mais próximo), Ligação Completa (vizinho mais distante), Método de Ward, Método das Médias das Distâncias e Método do Centróide. A seguir serão detalhados os três primeiros tipos de ligações.

Ligações Simples (vizinho mais próximo)

Nesta ligação o agrupamento é feito juntando-se dois grupos com menor distância ou maior similaridade. Quando formado o novo grupo, por exemplo, (AB), na ligação simples, a distância entre (AB) e algum outro grupo C é calculado e os resultados são apresentados graficamente em um diagrama de árvore ou dendrograma.

$$d_{(AB)C} = \min \{d_{AC}, d_{BC}\} \tag{2.25}$$

Ligação Completa (vizinho mais longe)

O procedimento é muito semelhante ao da ligação simples, com uma única exceção: o algoritmo aglomerativo começa determinando a menor distância d_{ik}, constrói-se a matriz de distâncias $D = (d_{ik})$ e os grupos vão se juntando. Se A e B são dois grupos de um único elemento, tem-se (AB) como novo grupo. A distância entre (AB) e outro grupo C é dada por:

$$d_{(AB)C} = \max \{d_{AC}, d_{BC}\} \tag{2.26}$$

Método de Ward

Este método é usado, avaliando a perda de informações utilizando o critério da soma ao quadrado dos erros, *(SQE),* entre dois agrupamentos, para todas as amostras. Quando se tem a distância mínima unem-se os grupos próximos, e volta-se a iterar nos grupos. Então tem-se:

$$SQE = \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x}) \tag{2.27}$$

Sendo:

$\bar{x}$ a média das amostra;

x_i uma medida multivariada associada com o i-ésimo item.

Os resultados são fornecidos na forma de dendrograma e o eixo vertical dá os valores de *SQE*. A decisão do número de classes ou tipos para análise é tomada a partir do exame do dendrograma ou árvore hierárquica, onde podem ser lidos os índices de nível ou índices de similaridade que são as distâncias euclidianas em que ocorrem as junções dos pontos observados para formar grupos. Uma grande distância indica que a agregação reuniu dois grupos muito dissimilares e, por isso, deve-se definir o número de grupos anterior a esta distância.

2.3.3.3 Discriminação, Classificação e Reconhecimento de Padrões

Estatisticamente, a construção e a avaliação de regras de reconhecimento e classificação de padrões podem ser baseadas em três métodos principais: Função Discriminante Linear de Fisher, Regressão Logística e Método das *k*-médias. Posteriormente, surgiu a tecnologia de Redes Neurais (tecnologia emergente), métodos de Programação Matemática e outros métodos para formação do conjunto de procedimentos usados no reconhecimento e classificação de objetos e indivíduos. Abaixo se detalha o método da Análise Discriminante.

- Análise Discriminante

Segundo Johnson e Wichern (1988), é uma técnica multivariada que tem por objetivo tratar dos problemas relacionados com separar conjuntos distintos de objetos (itens ou observações) e alocar novos objetos em conjuntos previamente definidos. Quando empregada como procedimento de classificação não é uma técnica exploratória, uma vez que ela conduz a regras bem definidas, as quais podem ser utilizadas para classificação de outros objetos.

Tem como objetivos imediatos, quando usada para discriminação e classificação, os seguintes:

1. Descrever algébrica ou graficamente as características diferenciais dos objetos (observações) de várias populações conhecidas a fim de achar "discriminantes" cujos valores numéricos sejam tais que as populações possam ser separadas tanto quanto possível.

2. Agrupar os objetos (observações) dentro de duas ou mais classes determinadas. Tenta-se encontrar uma regra que possa ser usada na alocação ótima de um novo objeto (observação) nas classes consideradas.

Uma função que separa pode servir para alocar um objeto e, da mesma forma, uma regra alocadora pode sugerir um procedimento discriminatório. Na prática, os objetivos 1 e 2, freqüentemente, sobrepõem-se e a distinção entre separação e alocação torna-se confusa.

- Discriminação e Classificação: Método de Fisher

1. Método de Fisher para Duas Populações

A idéia de Fisher foi transformar as observações multivariadas $\underline{X}$'s em observações univariadas Y's tais que os Y's das populações π_1 e π_2 sejam separados tanto quanto possível. Fisher tomou combinações lineares de $\underline{X}$ para criar os Y's, dado que as combinações lineares são funções de $\underline{X}$ e por outro lado são de fácil cálculo. Assim, sendo μ_{1y} a média dos Y's obtidos dos $\underline{X}$'s pertencentes a π_1 (população 1) e μ_{2y} a média dos Y's obtidos dos $\underline{X}$'s pertencentes a π_2 (população 2), Fisher selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos Y's. Assim, seja:

$\underline{\mu}_1 = E(\underline{X}|\pi_1)$ = valor esperado de uma observação multivariada de π_1. (2.28).

$\underline{\mu}_2 = E(\underline{X}|\pi_2)$ = valor esperado de uma observação multivariada de π_2. (2.29).

e supondo a matriz de covariância

$$\Sigma = E[(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)'] \quad i = 1, 2 \tag{2.30}$$

como sendo a mesma para as duas populações, e ainda considerando a combinação linear,

$$\underset{1x1}{Y} = \underset{1xp}{\underline{c}'} \underset{px1}{\underline{X}} \tag{2.31}$$

tem-se:

$$\mu_{1y} = E(Y|\pi_1) = E(\underline{c}'\underline{X}|\pi_1) = \underline{c}'E(\underline{X}|\pi_1) = \underline{c}'\underline{\mu}_1, \tag{2.32}$$

$$\mu_{2y} = E(Y|\pi_2) = E(\underline{c}'\underline{X}|\pi_2) = \underline{c}'E(\underline{X}|\pi_2) = \underline{c}'\underline{\mu}_2 \tag{2.33}$$

e

$$V(Y) = \sigma_y^2 = V(\underline{c}'\underline{X}) = \underline{c}'\, V(\underline{X})\, \underline{c} = \underline{c}'\Sigma\underline{c}, \tag{2.34}$$

que é a mesma para ambas as populações. Então, segundo Fisher, a melhor combinação linear é a derivada da razão entre o "quadrado da distância entre as médias" e a "variância de Y".

$$\frac{(\mu_{1y}-\mu_{2y})^2}{\sigma_y^2}=\frac{(\underline{c}'\underline{\mu}_1-\underline{c}'\underline{\mu}_2)^2}{\underline{c}'\sum\underline{c}}=\frac{\underline{c}'(\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)'\underline{c}}{\underline{c}'\sum\underline{c}}=\frac{(\underline{c}'\underline{\delta})^2}{\underline{c}'\sum\underline{c}} \qquad (2.35)$$

onde

$$\underline{\delta}=\underline{\mu}_1-\underline{\mu}_2. \qquad (2.36)$$

Então, com $\underline{\delta}=\underline{\mu}_1-\underline{\mu}_2$ e $Y=\underline{c}'\underline{X}$, $\dfrac{(\underline{c}'\underline{\delta})^2}{\underline{c}'\sum\underline{c}}$ é maximizada por:

$$\underline{c}=k\,\Sigma^{-1}\,\underline{\delta}=k\,\Sigma^{-1}\,(\underline{\mu}_1-\underline{\mu}_2) \text{ para qualquer } k\neq 0. \qquad (2.37)$$

Para $k = 1$ tem-se:

$$\underline{c}=\Sigma^{-1}\,(\underline{\mu}_1-\underline{\mu}_2) \text{ e } Y=\underline{c}'\underline{X}=(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{X}, \qquad (2.38)$$

que é conhecida como Função Discriminante Linear de Fisher. Ela transforma as populações multivariadas π_1 e π_2 em populações univariadas, tais que as médias destas populações são separadas tanto quanto possível relativamente à variância populacional, considerada comum. Logo, para classificar a observação multivariada $\underline{X}_0$ usa-se o modelo:

$$Y_0=(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{X}_0 \qquad (2.39)$$

Y_0 é o valor da Função Discriminante de Fisher para a nova observação $\underline{X}_0$, e considerando o ponto médio entre as médias das duas populações univariadas,

$$m=\frac{1}{2}(\mu_{1y}+\mu_{2y}), \qquad (2.40)$$

como

$$m=\frac{1}{2}(\underline{c}'_1\underline{\mu}_1+\underline{c}'_2\underline{\mu}_2)$$

$$m=\frac{1}{2}\left[(\underline{\mu}_1-\underline{\mu}_2)'\sum{}^{-1}\underline{\mu}_1+(\underline{\mu}_1-\underline{\mu}_2)'\sum{}^{-1}\underline{\mu}_2\right]$$

$$m = \frac{1}{2}\left[(\underline{\mu}_1 - \underline{\mu}_2)'\sum{}^{-1}(\underline{\mu}_1 + \underline{\mu}_2)\right] \tag{2.41}$$

e tem-se que:

$$E(Y_0 \mid \pi_1) - m \geq 0 \tag{2.42}$$

e

$$E(Y_0 \mid \pi_2) - m < 0, \tag{2.43}$$

ou seja, se $\underline{X}_0$ pertence a π_1, se espera que Y_0 seja igual ou maior do que o ponto médio. Por outro lado se $\underline{X}_0$ pertence a π_2, o valor esperado de Y_0 será menor que o ponto médio. Portanto, a regra de classificação é:

alocar $\underline{x}_0$ em π_1 se $y_0 - m \geq 0$

alocar $\underline{x}_0$ em π_2 se $y_0 - m < 0$

Os parâmetros $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ geralmente são desconhecidos. Então, supondo que se tem n_1 observações da v.a. multivariada $\underline{X}_1$ de dimensão p, ou seja, tem-se uma amostra aleatória da população π_1 e n_2 observações da v.a. multivariada $\underline{X}_2$ de dimensão p, que corresponde a uma mostra aleatória da população π_2, os resultados amostrais correspondentes são:

$$\bar{\underline{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\underline{x}_{i1};\ S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\underline{x}_{i1} - \bar{\underline{x}}_1)(\underline{x}_{i1} - \bar{\underline{x}}_1)' \tag{2.44}$$

$$\bar{\underline{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\underline{x}_{i2};\ S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(\underline{x}_{i2} - \bar{\underline{x}}_2)(\underline{x}_{i2} - \bar{\underline{x}}_2)' \tag{2.45}$$

Assumindo que as populações sejam assemelhadas, é natural considerar a variância como a mesma daí estima-se a matriz de covariância comum Σ pela matriz de covariância amostral calculada com a amostra conjunta,

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)} \tag{2.46}$$

que é um estimador não-viciado daquele parâmetro Σ.

Conseqüentemente, a Função Discriminante Linear de Fisher Amostral é dada por:

$$Y = \underline{\hat{c}}' \underline{X} = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S_p^{-1} \underline{x} \tag{2.47}$$

e a estimativa do ponto médio entre as duas médias amostrais univariadas,

$$\bar{y}_1 = \underline{\hat{c}}' \underline{\bar{x}}_1 \tag{2.48}$$

e

$$\bar{y}_2 = \underline{\hat{c}}' \underline{\bar{x}}_2 \tag{2.49}$$

é dada por:

$$\begin{aligned} \hat{m} &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}\left[(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S_p^{-1} \underline{\bar{x}}_1 + (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S_p^{-1} \underline{\bar{x}}_2\right] \\ \hat{m} &= \frac{1}{2}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S_p^{-1} (\underline{\bar{x}}_1 + \underline{\bar{x}}_2) \end{aligned} \tag{2.50}$$

Finalizando a regra de classificação é a seguinte:

$y_0 - \hat{m} \geq 0$ $\quad$ x_0 é alocado em π_1

$y_0 - \hat{m} < 0$ $\quad$ x_0 é alocado em π_2

A combinação linear particular $Y = \underline{\hat{c}}' \underline{x} = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S_p^{-1} \underline{x}$ maximiza a razão:

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\underline{\hat{c}}_1 \underline{\bar{x}}_1 - \underline{\hat{c}}_2 \underline{\bar{x}}_2)^2}{\underline{\hat{c}}' S_p \underline{\hat{c}}} = \frac{(\underline{\hat{c}}' \underline{d})^2}{\underline{\hat{c}}' S_p \underline{\hat{c}}} \tag{2.51}$$

onde:

$$\underline{d} = \underline{\bar{x}}_1 - \underline{\bar{x}}_2 \tag{2.52}$$

e

$$S_y^2 = \frac{\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2}{n_1 - n_2 - 2} \tag{2.53}$$

2. Discriminação entre Diversas Populações

O método anterior válido para duas populações pode ser estendido para diversas populações. O primeiro objetivo de Fisher com o método foi o de separar populações, podendo ser usado também para classificar novos itens em uma das populações. Esse método não necessita da suposição de que as diversas populações sejam normais multivariadas, porém, é necessário assumir que as matrizes de covariâncias populacionais são iguais, ou seja, $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma$. Assim, seja $\underline{\bar{\mu}}$ o vetor médio dos diversos grupos (populações),

$$\underline{\bar{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\underline{\mu}_i \tag{2.54}$$

e B_0 a matriz "Soma de produtos cruzados entre grupos populacionais" tal que:

$$B_0 = \sum_{i=1}^{g}(\underline{\mu}_i - \underline{\bar{\mu}})(\underline{\mu}_i - \underline{\bar{\mu}})' \tag{2.55}$$

A combinação linear $Y = \underline{c}'\underline{X}$ tem por esperança:

$$E(Y) = \underline{c}'E(\underline{x}\,|\pi_i) = \underline{c}'\underline{\mu}_i \tag{2.56}$$

para a população π_i e variância:

$$V(Y) = \sigma_y^2 = \underline{c}'V(\underline{X})\underline{c} = \underline{c}'\Sigma\underline{c} \tag{2.57}$$

para todas as populações. Desta forma, o valor esperado $\mu_{iy} = \underline{c}'\underline{\mu}_i$ muda quando a população da qual $\underline{X}$ é selecionado é outra. Tem-se então uma média global:

$$\bar{\mu}_y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iy} = \underline{c}'\underline{\bar{\mu}} \tag{2.58}$$

e a razão entre a "Soma dos quadrados das distâncias das populações para a média global" e a variância de Y' é $\frac{\underline{c}' B_0 \underline{c}}{\underline{c}' \Sigma \underline{c}}$ que é uma generalização multigrupal do caso de duas populações. Medindo a variabilidade entre grupos de valores (escores) Y relativamente à variabilidade comum dentro dos grupos, da mesma forma do que no caso de duas populações, pode-se selecionar $\underline{c}$ que maximiza esta razão. É conveniente normalizar $\underline{c}$ tal que $\underline{c}' \Sigma \underline{c} = 1$.

Sejam $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0$ os $s \leq \min(g - 1\ p)$ autovalores não-nulos de $\Sigma^{-1}B_0$ e $\underline{e}_1$, $\underline{e}_2, \ldots, \underline{e}_s$ os correspondentes autovetores (escalonados tal que $\underline{e}' \Sigma \underline{e} = 1$). Então, é fácil provar que o vetor de coeficientes $\underline{c}$ que maximiza a razão $\frac{\underline{c}' B_0 \underline{c}}{\underline{c}' \Sigma \underline{c}}$ é dado por $\underline{c}_1 = \underline{e}_1$. A combinação linear $\underline{c}_1' \underline{X}$ é chamada primeiro discriminante e de forma idêntica, pode-se generalizar para o k-ésimo discriminante com $\underline{c}_k = \underline{e}_k$ com $k = 1, 2, \ldots, s$. Geralmente, Σ e $\underline{\mu}_i$ não são conhecidas, toma-se amostras aleatórias de tamanhos n_i das populações π_i, $i = 1, 2, \ldots, g$ e denotando o conjunto de dados da população π_i, $i = 1, 2, \ldots, g$, por ${}_{ni}X_p$ tem-se os estimadores dos parâmetros $\underline{\mu}_i$ e $\overline{\underline{\mu}}$ dados por:

$$\overline{\underline{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{x}_{ij} \tag{2.59}$$

$$\overline{\underline{x}} = \frac{\sum_{i=1}^{g} n_i \overline{\underline{x}}_i}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} \underline{x}_{ij}}{\sum_{i=1}^{g} n_i} \tag{2.60}$$

A matriz "Soma de produtos cruzados entre grupos", B_0, é estimada por:

$$\hat{B}_0 = \sum_{i=1}^{g} (\overline{\underline{x}}_i - \overline{\underline{x}})(\overline{\underline{x}}_i - \overline{\underline{x}})' \tag{2.61}$$

e um estimador para Σ pode ser obtido com base na matriz W:

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underline{x}_{ij} - \overline{\underline{x}}_i)(\underline{x}_{ij} - \overline{\underline{x}})' = \sum_{i=1}^{g} (n_i - 1) S_i \tag{2.62}$$

Conseqüentemente,

$$\frac{W}{n_1+n_2+\ldots+n_g-g}=\frac{(n_1-1)S_1+(n_2-1)S_2+\ldots.+(n_g-1)S_g}{n_1+n_2+\ldots.+n_g-g}=S_p \qquad (2.63)$$

Assim, o mesmo $\hat{\underline{c}}$ que maximiza a razão $\frac{\hat{\underline{c}}'\hat{B}_0\hat{\underline{c}}}{\hat{\underline{c}}'S_p\hat{\underline{c}}}$ também maximiza $\frac{\hat{\underline{c}}'\hat{B}_0\hat{\underline{c}}}{\hat{\underline{c}}'W\hat{\underline{c}}}$. Logo, apresentar-se-á o otimizante $\hat{\underline{c}}$ na forma mais usual, que é o autovetor $\hat{\underline{e}}_i$ da matriz $W^{-1}B_0$, porque se $W^{-1}B_0\hat{\underline{e}} = \hat{\lambda}\hat{\underline{e}}$ então $S_p^{-1}\hat{B}_0\hat{\underline{e}} = \hat{\lambda}(n_1+n_2+\ldots+n_g-g)\hat{\underline{e}}$, portanto, concluindo que sejam $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_g > 0$ os autovalores não nulos de $W^{-1}B_0$ e $\hat{\underline{e}}_1, \hat{\underline{e}}_2, \ldots.., \hat{\underline{e}}_s$ os correspondentes autovetores, sendo $s \leq \min(g-1, p)$ e $\hat{\underline{e}}_i$ normalizado tal que $\hat{\underline{e}}_i S_p \hat{\underline{e}}_i = 1$; então o vetor de coeficientes que maximiza a razão citada acima é $\hat{\underline{c}}_1 = \hat{\underline{e}}_1$ e a combinação linear $\hat{\underline{e}}_1'\underline{x}$ é chamada primeiro discriminante amostral. Generalizando, tem-se no passo k o k-ésimo discriminante amostral $\hat{\underline{e}}_k'\underline{x}$, $k \leq s$.

2.3.3.4 Avaliação de Funções de Reconhecimento e Classificação

1. Critério *TPM* (*Total Probability of Misclassification*)

A *TPM* é dada por:

$$TPM = p_1\int_{R_2} f_1(\underline{x})d\underline{x} - p_2\int_{R_1} f_2(\underline{x})d\underline{x} \qquad (2.64)$$

onde: p_1 e p_2 são as probabilidades de uma observação pertencer a π_1 ou a π_2, respectivamente. E o menor valor para esta quantidade, obtido pela escolha adequada das regiões R_1 e R_2, é chamado de taxa ótima de erro (*optimum error rate*), *OER*,

$$OER = p_1\int_{R_2} f_1(\underline{x})d\underline{x} - p_2\int_{R_1} f_2(\underline{x})d\underline{x} \qquad (2.65)$$

com R_1 e R_2 determinados por $R_1 : \frac{f_1(\underline{x})}{f_1(\underline{x})} \geq \frac{p_2}{p_1}$ e R_2 em caso contrário.

A taxa aparente de erro (que é definida como fração das observações no treinamento amostral) é calculada da matriz de confusão que mostra a situação real das observações nos

grupos versus o reconhecimento. Para n_1 observações de π_1 e n_2 de π_2, a matriz de confusão tem a forma:

Figura 2.2: Matriz de confusão

		Predito		
		π_1	π_2	
Classificação atual	π_1	n_{1C}	n_{1M}	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Onde:

n_{1C}: número de itens de π_1 corretamente reconhecido como de π_1;

n_{1M}: número de itens π_1 misturados com de π_2;

n_{2C}: número de itens π_2 corretamente reconhecido como de π_2;

n_{2M}: número de itens π_2 misturados com de π_1.

A taxa aparente de erro (*APER*) é dada por:

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{2.66}$$

e é vista como a proporção de itens ou observações no conjunto de treinamento que são reconhecidos erroneamente.

2. Abordagem de Lachenbruch

É uma técnica para avaliar a eficiência da regra de classificação. Os passos são:

1. Comece com o grupo da população π_1. Omita uma observação deste grupo e construa uma função baseada nas $n_1 - 1$ e n_2 observações.

2. Reconheça (classifique), usando a função, a observação não incorporada.

3. Repita os passos 1 e 2 até que todas as n_1 observações de π_1 sejam classificadas. Seja $n_{1M}^{(H)}$ o número de observações reconhecidas erroneamente neste grupo.

4. Repita os passos de 1 a 3 para as n_2 observações de π_2. Seja $n_{2M}^{(H)}$ o número de observações reconhecidas erroneamente neste grupo.

Então,

$$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1} \tag{2.67}$$

e

$$\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2} \tag{2.68}$$

e a proporção total esperada de erro é:

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \tag{2.69}$$

Assim, obtém-se uma regra de reconhecimento e classificação construída com as n observações amostrais e testadas com todas referidas observações.

2.4 ANÁLISE DE REGRESSÃO LINEAR

2.4.1 Introdução

A Análise de Regressão é a técnica estatística indicada quando se estuda o relacionamento entre as variáveis (dependente e independente). Existem numerosas aplicações desta técnica que ocorrem em quase todos os campos científicos.

Na Engenharia de Avaliação, considera-se, geralmente, como variável dependente os preços à vista de mercado em oferta e efetivamente transacionados e como variáveis

independentes as características do imóvel. No modelo linear que pode representar os preços de mercado, a variável resposta (dependente) é expressa por uma combinação linear das variáveis independentes, de forma original ou transformada, e respectivas estimativas dos parâmetros populacionais, acrescidas de erro aleatório, oriundo de variações do comportamento humano.

Assim, tem-se o modelo, que representa n observações,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,\,p-1} + \varepsilon_i,\ i = 1, 2, \cdots, n \qquad (2.70)$$

2.4.2 Modelo Linear Geral de Regressão

O modelo (2.70) com n observações pode ser colocado na forma matricial. Então se tem:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad (2.71)$$

onde $\underline{Y}$ é o vetor aleatório de resposta, $\underline{\beta}$ é o vetor de parâmetros de dimensão p, X é a matriz do modelo de ordem n x p e $\underline{\varepsilon}$ é o vetor aleatório de erros de dimensão n. E, de forma detalhada,

$$\underset{n\,x\,1}{\underline{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \qquad \underset{n\,x\,p}{\underline{X}} = \begin{bmatrix} 1 & X_{11} & . & . & X_{1,p-1} \\ 1 & X_{21} & . & . & X_{2,p-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & . & . & X_{n,p-1} \end{bmatrix} \qquad \underset{p\,x\,1}{\underline{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\,x\,1}{\underline{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

A aplicação completa do modelo (2.71) é fundamentada nas seguintes suposições:

- o vetor de erros $\underline{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]$ é aleatório, ou seja, as componentes ε_i, $i = 1, 2, \ldots, n$ são variáveis aleatórias;
- a esperança de cada componente de $\underline{\varepsilon}$ é zero, ou seja, $E(\underline{\varepsilon}) = \underline{0}$ e $E(\underline{\varepsilon_i}) = \underline{0}$;
- as componentes do vetor $\underline{\varepsilon}$ não são correlacionadas, ou melhor, cov $(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$ e possuem variância constante, σ^2. Assim, a matriz de covariâncias de $\underline{\varepsilon}$ é a matriz diagonal $\sigma^2 I_n$, onde I_n é a matriz identidade de ordem n, $V(\underline{\varepsilon}) = \sigma^2 I_n$.

O modelo (2.71) com as três suposições anteriores é conhecido como Modelo Linear de Gauss Markov e o Teorema de Gauss-Markov garante que sob as três suposições e com $X'X$ não singular, os estimadores não viciados uniformemente de mínima variância (EUMV) do vetor $\underline{\beta}$ e da variância σ^2 são, respectivamente:

$$\underline{\hat{\beta}} = (X'X)^{-1}(X'\underline{Y}) \tag{2.72}$$

e

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{2.73}$$

Uma outra suposição que pode ser exigida, além das já citadas, para o modelo de regressão é a seguinte:

- a distribuição de ε_i, i =1, 2,..., n é a Normal (Gaussiana).

Levando em consideração esta suposição, tem-se o modelo de Gauss-Markov Normal e $Y_i \sim N\left(\sum_{i=1}^{p}\beta_i x_i, \sigma^2\right)$.

2.4.3 Análise da Variância da Regressão

Esta análise é uma das técnicas estatísticas cujas bases foram lançadas por Fisher. É a técnica geralmente usada para verificar se o ajuste de regressão existe. É importante construir um quadro que resuma as informações da Análise da Variância (ANOVA), para o modelo geral:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,\,p-1} + \varepsilon_i,\ i = 1, 2, \cdots, n \tag{2.74}$$

com $p \geq 2$ parâmetros. Assim, o quadro 2.1 resume as informações.

Quadro 2.1: Análise de Variância (ANOVA)

Fonte de variação	Soma de quadrados	G.L.	Quadrado médio	F
Regressão	$SQ\text{Regr} = \underline{\hat{\beta}}' X'\underline{Y} - n\bar{y}^2$	$p-1$	$\frac{SQ\,\text{Re}\,gr}{p-1}$	$\frac{SQ\,\text{Re}\,gr}{p-1} / \frac{SQR}{n-p}$
Residual	$SQR = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}' X'\underline{Y}$	$n-\text{p}$	$\frac{SQR}{n-p}$	
Total	$SQT = \underline{Y}'\underline{Y} - n\bar{y}^2$	$n-1$		

O teste feito com a estatística F (última coluna do quadro 2.1) é o da hipótese nula H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, ou seja, se existe regressão dos X's para Y, ou melhor, se existe relação linear entre a variável resposta Y e as covariáveis X_i, $i = 1, 2, \ldots, p-1$, este teste é fundamental para validade do modelo linear ajustado, $\hat{\underline{Y}} = X\hat{\underline{\beta}}$.

2.4.4 Verificação dos Pressupostos do Modelo

a) Homocedasticidade

A homocedasticidade corresponde à variância constante dos resíduos. É uma propriedade essencial e que deve ser garantida, sob pena de invalidar toda a análise estatística. Deseja-se que os erros sejam aleatórios, ou seja, não devem ser relacionados com as características dos imóveis. Quando a homocedasticidade não ocorre, há heterocedasticidade; significa dizer que as chances de ocorrerem erros grandes (ou pequenos) variam conforme o tipo de imóvel. Há tendências nos erros. As conseqüências da heterocedasticidade são que as estimativas dos parâmetros da regressão (β_0, β_1, β_2,..., β_{p-1}) não são tendenciosas, mas são ineficientes e as estimativas das variâncias são tendenciosas. Os testes "t" e F tendem a dar resultados incorretos. Neste caso, os resultados não são confiáveis, ou seja, o modelo pode parecer bom, mas ele não é adequado aos dados, na verdade.

A homocedasticidade pode ser verificada através de gráficos de resíduos (erros). Se os pontos estão distribuídos aleatoriamente em uma faixa, sem demonstrar um comportamento definido, há homocedasticidade. No entanto, se existe alguma tendência (crescimento, decrescimento ou oscilação), então há heterocedasticidade. Havendo heterocedasticidade, podem ser feitas transformações nas variáveis (geralmente logarítmicas) ou outras soluções mais complexas. O modelo então, deve ser modificado.

b) Independência serial dos resíduos (não-autocorrelação)

Ocorre autocorrelação quando os erros são correlacionados com os valores anteriores ou posteriores na série. Este fenômeno é chamado de correlação serial.

Seu surgimento pode se dar por especificação incorreta do modelo da regressão, por causa de erros na forma do modelo ou ainda, por exclusão de variáveis independentes importantes para a análise. Existindo autocorrelação, os estimadores ordinários de mínimos quadrados não são mais os melhores estimadores lineares não-tendenciosos. Neste caso, existirão outros métodos que produzem menor variância amostral nos estimadores. Além disso, em presença de correlação serial, os testes de significância (t e F) e de construção de intervalos de confiança dos coeficientes da regressão também oferecem conclusões incorretas, isto é, as regiões de aceitação e os intervalos de confiança podem ser mais largos ou mais estreitos do que os calculados, dependendo da tendência ser positiva ou negativa.

A verificação da autocorrelação pode ser realizada pela análise do gráfico dos resíduos comparados com os valores preditos, onde este deve apresentar pontos dispersos aleatoriamente, sem nenhum padrão definido ou pelo teste de Durbin-Watson.

c) Normalidade dos resíduos

A Análise de Regressão baseia-se na hipótese de que os erros seguem uma distribuição Normal (distribuição de Gauss). Em presença de falta de Gaussianidade, os estimadores são não-tendenciosos, mas os testes não têm validade, principalmente em amostras pequenas. Entretanto, pequenas fugas da Gaussianidade não causam grandes problemas. A heterocedasticidade ou a escolha de um modelo incorreto para a equação pode ser a causa da não-normalidade dos resíduos.

A verificação da Gaussianidade pode ser feita pelos testes de aderência como, por exemplo, o de Kmogorov-Sminorv ou de Shapiro-Wilks.

d) *Outliers*

Denomina-se *outlier* um dado que contém grande resíduo em relação aos demais que compõem a amostra e assim tem comportamento muito diferente dos demais (Dantas, 1998).

É de grande importância controlar os *outliers,* porque de acordo com a estimação da equação, um grande erro modifica significativamente seus somatórios, alterando os coeficientes da equação. Dessa forma, um imóvel apenas pode modificar a equação.

Geralmente, adota-se o intervalo de 2 desvios padrões em torno da média dos erros, pois não existe um limite fixo. Como a média deve ser zero, os resíduos padronizados $\left(\frac{\varepsilon_i}{dp}\right)$ devem estar entre –3 e 3. Os imóveis com erros que ultrapassam estes limites devem ser analisados cuidadosamente; a existência de *outliers* deve sempre ser interpretada como um sinal de problema na amostra.

e) Colinearidade ou multicolinearidade

Multicolinearidade é definida como a existência de relações lineares entre as variáveis "independentes", de tal forma correlacionada umas às outras, tornando-se difícil ou impossível isolarem suas influências separadas e obter uma estimativa precisa de seus efeitos relativos. Quando a relação é exata tem-se o caso da multicolinearidade perfeita.

Raramente encontram-se variáveis independentes que são perfeitamente relacionadas. Esse caso não traz problemas, pois é facilmente detectado e pode ser resolvido simplesmente eliminando uma ou mais variáveis independentes do modelo. O interesse no que se refere à multicolinearide está nos casos em que ela ocorre com alto grau, isto é, quando duas variáveis independentes estão significativamente correlacionadas ou quando há uma combinação linear entre um conjunto de variáveis independentes. Assim, a multicolinearidade é mais uma questão de grau do que de natureza (Kmenta, 1978).

O fato de muitas funções e regressões diferentes proporcionarem bons ajustes para um mesmo conjunto de dados é porque os coeficientes de regressão atendem várias amostras onde as variáveis independentes são altamente correlacionadas. "Assim, os coeficientes de regressão estimados variam de uma amostra para outra quando as variáveis independentes estão altamente correlacionadas. Isso leva a informações imprecisas a respeito dos coeficientes verdadeiros" (Neter e Wasserman, 1974).

Geralmente, a multicolinearidade é causada pela própria natureza dos dados, principalmente nas áreas de economia com variáveis que representam valores de mercado. Pode também ocorrer devido à amostragem inadequada.

Em Análise de Regressão Linear Múltipla, existe um freqüente interesse com relação à natureza e significância das relações entre as variáveis independentes e a variável dependente. "Em muitas aplicações de administração e economia, freqüentemente encontram-se variáveis independentes que estão correlacionadas entre elas mesmas e, também, com outras variáveis que não estão incluídas no modelo, mas estão relacionadas à variável dependente" (Neter e Wasserman).

A existência de multicolinearidade tendo sido detectada e considerada prejudicial, indicando que o pesquisador deve procurar soluções para suavizar seus efeitos nocivos. Várias medidas corretivas têm sido propostas, desde medidas simples às mais complexas, para suavizar os efeitos provocados pela multicolinearidade (Elian, 1988; Judge et al., 1980).

Algumas soluções para o problema da multicolinearidade são propostas, tais como: remoção de variáveis, ampliação do tamanho da amostra, adoção de técnicas estatísticas como Análise de Componentes Principais, entre outras. Neste trabalho aplicou-se a técnica das Componentes Principais para remover o problema da multicolinearidade entre as características dos imóveis.

2.4.5 Poder de Explicação do Modelo

Para se medir o quanto a variabilidade total dos dados é explicada pelo modelo de regressão, compara-se a Soma de Quadrados da Regressão com a Soma de Quadrados Total e, então, tem-se o coeficiente de determinação ou de correlação múltipla ao quadrado, R^2,

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad 0 < R^2 < 1 \qquad (2.75)$$

Quando o ajuste é bom o modelo explica boa parte da variação total e, consequentemente, o valor de R^2 é próximo de 1.

A estatística R^2 indica a qualidade do ajuste do modelo adotado.

2.4.6 Relação entre Variáveis

Na análise de um modelo de regressão o coeficiente de correlação é uma medida estatística muito importante. O grau de associação entre duas variáveis é definido

numericamente pelo Coeficiente de Correlação, parâmetro representado por ρ. Com base em n observações do par (X, Y) este parâmetro é estimado pela estatística,

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\left(X_i - \overline{X}\right)^2\left(Y_i - \overline{Y}\right)^2}} = \frac{S_{xy}}{S_x S_y} \tag{2.76}$$

onde,

$\overline{X}$ é a média da variável X;

$\overline{Y}$ é a média da variável Y;

S_{xy} é a covariância amostral entre X e Y;

S_x é o desvio padrão amostral de X;

S_y é o desvio padrão amostral de Y.

O coeficiente de correlação varia entre os limites –1 e 1 podendo ser positivo ou negativo ($-1 \leq \rho \leq 1$) e também nulo ($\rho = 0$). Quando $\rho = 0$ significa que não existe nenhum relacionamento entre as variáveis. Quando o coeficiente de correlação é igual à unidade, –1 ou + 1, tem-se um relacionamento perfeito entre elas. O grau de relacionamento entre as variáveis, definido numericamente pelo valor $\hat{\rho}$ no caso amostral, pode ser assim interpretado:

Coeficiente		**Correlação**
$\left\|\hat{\rho}\right\| = 0$	.....	relação nula
$0 < \left\|\hat{\rho}\right\| \leq 0{,}30$	.....	relação fraca
$0{,}30 < \left\|\hat{\rho}\right\| \leq 0{,}70$	.....	relação média
$0{,}70 < \left\|\hat{\rho}\right\| \leq 0{,}90$	.....	relação forte

$0{,}90 < \left|\hat{\rho}\right| \leq 0{,}99$ ..... relação fortíssima

$\left|\hat{\rho}\right| = 1$ ..... relação perfeita

Nota-se também que nem sempre uma elevada correlação entre duas variáveis representa a existência de relação de causa e efeito entre as mesmas. Esses casos dão origem às chamadas de influência no caso.

O estudo do relacionamento entre um conjunto de variáveis pode ser realizado aplicando diversas técnicas, desde os coeficientes de correlação de Pearson, de Spearman, Análise Fatorial e a Análise de Componentes Principais. A estatística (2.76) é conhecida como o coeficiente de correlação linear de Pearson e é uma medida usada no estudo da relação linear existente entre duas varáveis X e Y.

2.4.7 Seleção de Variáveis Regressoras

Um dos problemas mais freqüentes em Análise de Regressão é a seleção do conjunto de variáveis independentes a serem incluídas no modelo (Neter e Wasserman, 1974).

O pesquisador deve especificar o conjunto de variáveis independentes a ser empregado para descrever, controlar ou predizer a variável dependente. Um problema muito difícil de relacionamento que aparece na seleção de variáveis é quando uma equação de regressão é construída com o objetivo de predição e envolve muitas variáveis. Talvez, muitas delas contribuam pouco ou nada para a precisão da predição. A escolha apropriada de algumas delas fornece a melhor predição, porém quais e quantas devem ser selecionadas? (Snedecor e Cochram, 1972).

Em algumas áreas, a teoria pode ajudar na seleção das variáveis independentes a serem empregadas e na especificação da forma funcional da relação de regressão. Os experimentos podem ser controlados para fornecer dados sobre a base de que os parâmetros de regressão podem ser estimados e a forma teórica da regressão testada.

Em muitos outros campos, modelos teóricos são raros. Assim, os investigadores são freqüentemente forçados a explorar as variáveis independentes para que possam realizar estudos sobre a variável dependente. Algumas das variáveis independentes podem ser removidas seletivamente. Uma variável independente pode não ser fundamental ao problema;

pode estar sujeita a grandes erros de medidas e pode duplicar outra variável independente da lista. Assim, outras variáveis independentes, que não podem ser medidas, podem então ser excluídas ou substituídas por variáveis que estão altamente correlacionadas com estas.

Normalmente, após uma seleção inicial, o número de variáveis independentes ainda é grande. Sendo assim, o investigador geralmente desejará reduzir o número de variáveis independentes a serem usadas no modelo final, existindo razões para isto: uma delas é que um modelo de regressão com um número grande de variáveis independentes é caro para se utilizar. O problema torna-se, então, de como reduzir a lista de variáveis independentes de forma a obter a melhor seleção de variáveis independentes. Este conjunto precisa ser pequeno para que a manutenção dos custos de atualização do modelo seja manuseável e a análise facilitada, e ainda, deve ser grande o suficiente de forma que seja possível uma descrição, um controle e uma predição adequados.

Existem muitos procedimentos de seleção, mas nenhum deles pode, comprovadamente, produzir o melhor conjunto de variáveis independentes. Dentre os procedimentos, pode-se citar como os mais comumente usados: todas as regressões possíveis, *backward, forward* e *stepwise*.

Todas as regressões possíveis: consiste em ajustar todas as possíveis equações de regressão. Após a obtenção de todas as regressões, devem-se utilizar os critérios para comparação dos modelos ajustados. Alguns critérios que podem ser usados são o R^2 (coeficiente de explicação), MSE (quadrado médio dos resíduos) e C_p (estatística de Mallows). Para alguns conjuntos de variáveis, os três critérios podem levar para o mesmo "melhor" conjunto de variáveis independentes. Este não é o caso geral, pois diferentes critérios podem sugerir diferentes conjuntos de variáveis independentes. Daniel e Wood (1971) recomendam, no caso de um grande número de equações alternativas, o critério do erro quadrado total para caracterizar a equação. A principal desvantagem do procedimento de procura de todas as regressões possíveis é a quantidade de esforço computacional necessária, já que cada variável independente potencial pode ser incluída ou excluída, gerando $(2p-1)$ regressões possíveis quando existem p variáveis independentes potenciais (Elian, 1998; Draper e Smith, 1981).

- *Stepwise* (passo a passo)

É o método mais usado dos métodos de pesquisa que não requerem a computação de todas as regressões possíveis. A rotina de regressão *stepwise* permite que uma variável independente, trazida para dentro do modelo em um estágio anterior, seja removida subseqüentemente se ela não ajudar na conjunção com variáveis adicionadas nos últimos estágios. Esta rotina empregada conduz a um teste para rastrear alguma variável independente que seja altamente correlacionada com variáveis independentes já incluídas no modelo. A limitação da procura da regressão *stepwise* é que ela presume a existência de um único conjunto ótimo de variáveis independentes e busca identificá-lo. Como notado anteriormente, não existe freqüentemente um único conjunto ótimo. Outra limitação da rotina de regressão *stepwise*, é que ela algumas vezes surge com um conjunto de variáveis independentes razoavelmente fracos para predições, quando as variáveis independentes estão altamente correlacionadas (Draper e Smith, 1981).

- Seleção *forward*

É uma versão simplificada da regressão *stepwise*, omitindo o teste, se uma variável uma vez que tenha entrado no modelo deva ser retirada. Este procedimento considera, inicialmente, um modelo simples usando a variável de maior coeficiente de correlação com a variável dependente. Uma variável por vez é incorporada até que não haja mais inclusão, e as variáveis selecionadas definem o modelo.

- Eliminação *backward*

Este procedimento de procura é oposto à seleção *forward*. Ele começa com o modelo contendo todas as variáveis independentes potenciais. O procedimento *backward* requer mais cálculos do que o método de seleção *forward*, já que ele começa com o maior modelo possível. Entretanto, ele tem uma vantagem de mostrar ao analista as implicações do modelo com muitas variáveis.

3 MATERIAL E MÉTODO

3.1 MATERIAL

3.1.1 Área de Estudo

A área de estudo foi a cidade de Campo Mourão, situada ao Noroeste do Estado do Paraná, distante 87 km de Maringá, 323 km de Foz do Iguaçu, 477 km de Curitiba e 659 km de São Paulo e se constitui no maior entroncamento rodoviário do Sul do Brasil. Sede da Microrregião 12 (divisão administrativa estadual), Campo Mourão agrega 25 municípios com economia baseada inicialmente no setor primário e hoje realiza investimentos na área industrial, já em avançado estágio de implementação do setor secundário e desenvolvimento do terciário.

A fertilidade da terra permite uma grande produtividade no campo. A área cultivada de Campo Mourão ultrapassa os 50 mil hectares. As principais culturas são: soja, trigo, milho, algodão e aveia. Paralelamente à agricultura, destaca-se o parque de revendas e assistência técnica de equipamentos e insumos. Campo Mourão é sede da maior cooperativa singular da América Latina, a Coamo - Cooperativa Agropecuária Mourãoense, hoje Coamo Agroindustrial Cooperativa.

As Coordenadas geográficas do Município são 24°02'38" de Latitude Sul e 52°22'40" de Longitude Oeste do Meridiano de *Greenwich*, a uma altitude média de 630 metros sobre o nível do mar. A seguir, na figura 3.1, tem-se a localização da cidade no mapa do Estado do Paraná.

Figura 3.1: Mapa com a localização de Campo Mourão

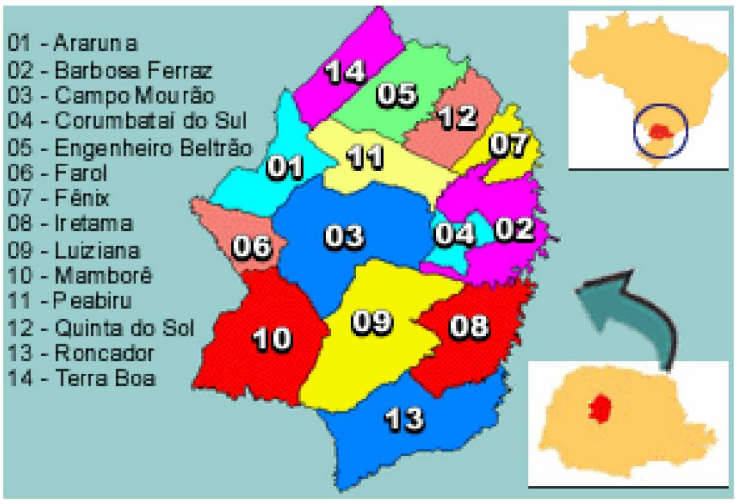
Fonte: Prefeitura Municipal de Campo Mourão, 2005.

Campo Mourão limita-se com os seguintes Municípios:

- Norte: Peabiru
- Nordeste: Barbosa Ferraz
- Sul: Luiziana
- Leste: Corumbataí do Sul
- Oeste: Farol e Mamborê
- Noroeste: Araruna

A seguir, na figura 3.2, tem-se o mapa dos municípios vizinhos de Campo Mourão.

Figura 3.2: Mapa dos municípios vizinhos à Campo Mourão



Fonte: Prefeitura Municipal de Campo Mourão, 2005.

O Município de Campo Mourão pertence à bacia hidrográfica do Rio Ivaí, sendo seu rio mais importante o Rio Mourão, que atravessa o Município de sul a norte. Outros rios, importantes por serem condicionantes físico-naturais à expansão urbana de Campo Mourão, são o Rio km 119 e Rio do Campo, este último, onde a SANEPAR coleta 80% da água que abastece o município.

A seguir têm-se outras informações:

Área da unidade territorial: 766,44 km²;

População estimada em 2004: 81.259 habitantes;

Pessoas Residentes na Área Urbana: 74.754 habitantes;

Domicílios particulares permanentes em 2004: 22.829 domicílios;

Atividades imobiliárias (aluguéis e serviços prestados às empresas): 36 empresas.

3.1.2 Limitações da Pesquisa

A falta de dados para um melhor aprimoramento do modelo, pois a imobiliária TAPOVIC, mesmo sendo a maior imobiliária da cidade tinha poucos dados a oferecer, já que para uma boa segurança e confiança no tratamento estatístico, recomenda-se pelo menos o triplo do número de variáveis para o número de informações.

Outra limitação do trabalho é o espaço de tempo. Como qualquer alteração na economia provoca modificações nos valores dos imóveis, estes estão sujeitos às influências dos governantes e das economias local, regional, nacional e global. Assim, no decorrer do tempo, existe uma flutuação dos valores dos imóveis.

3.1.3 Levantamento dos Dados

Os dados foram aproveitados da pesquisa realizada por Silvia Neide Bráulio (2004), feita de forma cautelosa, e dos dados depende o sucesso da Análise Estatística. Foi realizado um planejamento antes da coleta dos dados. Nesse planejamento contemplou-se o espaço físico, local onde está inserido o total de imóveis, população a estudar e o número de imóveis a serem pesquisados. No mercado de imóveis, é freqüente a entrada de dados novos, por isso, deve-se fazer um novo levantamento a cada nova avaliação para garantir a representação dos novos dados na amostra (Dantas, 1998).

Na determinação da oferta imobiliária existem aspectos de extensa variação e combinação de atributos constituindo a heterogeneidade do produto habitação. Essa dispersão deve estar presente na descrição completa do mercado, incluída nas faixas de preços, tamanhos dos imóveis e, ainda, nas diferentes localizações. Assim, faz-se necessário obter o maior número de dados e atributos possíveis (Bráulio, 2005).

3.1.3.1 As Variáveis Utilizadas

As variáveis explicativas (independentes) são do tipo quantitativo e qualitativo, representando as características do imóvel, e estão a seguir detalhadas.

A variável resposta (dependente) é o preço, que representa o valor de venda do imóvel em reais. As varáveis originais e as independentes estão relacionadas, classificadas e descritas nos Quadros 3.1, 3.2 e 3.3 para os tipos de imóveis, apartamentos, casas e terrenos respectivamente.

Quadro 3.1: Variáveis independentes para apartamento

Variáveis	*Categorias*	*Descrição*
Revestimento do prédio	1 a 4	Identifica o revestimento do prédio.
Andar	1 a 4	Identifica o andar que o apartamento está localizado. Sabe-se que dependendo do andar que se localiza o apartamento ele é mais ou menos valorizado.
Dependência de empregado	Sem = 0 Com = 1	Identifica a existência ou não de dependência de empregados.
Estado de conservação	1 a 4	Identifica o nível de conservação do imóvel.
Suíte	Sem = 0 Com = 1	Identifica a presença ou não de suíte, atribuindo o valor 1 mesmo quando há presença de mais de uma suíte.
Idade aparente	Até 1 ano = 6 2-5 anos =5 6-10 anos = 4 11-15 =3 15-20 anos =2 Mais de 20 =1	Idade aparente: idade aparente do edifício. Por ser uma variável contínua, a idade do imóvel dividiu-se em períodos.
Idade real	Até 1 ano = 6 2-5 anos =5 6-10 anos = 4 11-15 =3 15-20 anos =2 Mais de 20 = 1	Idade real: idade cronológica do edifício reflete o estágio tecnológico.
Proximidade	1 a 3	Identifica a quanto o imóvel se localiza próximo de escolas, supermercados, hospitais e do centro comercial.
Lavanderia	0 ou 1	Identifica a existência ou não de lavanderia.
Posição do apartamento	1 a 3	Identifica a posição do apartamento em relação ao prédio (frente, lateral ou fundo).
Padrão de acabamento	1 a 3	Identifica os vários níveis de acabamento.
Sala	Unidade	Indica o número de salas existentes no apartamento.
Pavimento	Unidade	Indica o número de pavimentos do prédio.
Garagem	Unidade	Quantifica o número de vagas para carro disponível para cada apartamento.
Dormitório	Unidades	Quantifica o número de dormitórios.
Elevador	Unidades	Identifica a quantidade de elevadores no prédio.
Área privativa	m^2	Corresponde à superfície ou área do apartamento expressa em metros quadrados, obtida do registro de imóveis.
Peças	Unidades	Quantifica as peças constituintes do imóvel.
Banheiro	Unidades	Identifica o número de banheiro social.

Fonte: Imobiliária Tapowik, 2004.

Quadro 3.2: Variáveis independentes: casas residenciais

Variáveis	*Categorias*	*Descrição*
Localização	1 a 5	Naturalmente um local é "melhor" ou "pior" do que um outro em função de diversas características, entre as quais sua infra-estrutura urbana.
Dependência de empregado	Completa = 1 Incompleta = 0,5 Inexistente = 0	Identifica a existência ou não de dependência de empregado, completa ou incompleta.
Nível de conservação	1 a 4	Identifica o nível de conservação do imóvel.
Suíte	0 ou 1	Identifica a presença ou não de suíte, atribuindo o valor 1 mesmo quando há presença de mais que uma suíte.
Idade aparente	Até 1 ano = 6 2-5 anos =5 6-10 anos = 4 11-15 =3 15-20 anos =2 Mais de 20 anos = 1	Por ser uma variável continua a idade do imóvel dividiu-se em períodos.
Garagem	0 ou 1	Identifica a presença de garagem, onde é atribuído o valor mesmo quando a mais que uma vaga.
Distância de supermercados	1 a 3	Identifica a proximidade do imóvel de grandes mercados.
Presença de lavanderia	0 ou 1	Identifica a existência ou não de lavanderia.
Edícula	0 ou 1	Identifica a presença (1) ou não (0) de edícula.
Padrão de acabamento	1 a 3	Identifica os vários níveis de acabamento.
Piscina	0 ou 1	Identifica a existência ou não de piscina.
Cobertura	1 a 4	Identifica o tipo de cobertura do imóvel.
Estrutura	1 a 5	Identifica o material de construção do imóvel.
Dormitório	Unidades	Quantifica o número de dormitórios.
Área do terreno	m^2	Identifica a área do terreno.
Área construída	m^2	Identifica a área total construída.
Peças	Unidades	Quantifica as peças constituintes do imóvel.
Banheiro	Unidades	Identifica o número de banheiro social.

Fonte: Imobiliária Tapowik, 2005.

Quadro 3.3: Variáveis independentes: terrenos

Variáveis	*Categorias*	*Descrição*
Localização	1 a 6	Variável que qualifica a localização do imóvel.
Pólo de influência	–1 ou 1	Indica se o móvel localiza-se próximo a locais que influenciam no seu valor
Plano	0 a 3	Identifica se o terreno está acima, abaixo ou ao nível da rua.
Inclinado	0 a 3	Indica o nível de inclinação do terreno.
Pavimentação	0 ou 1	Identifica a presença ou não de pavimentação na rua ou avenida onde está inserido o terreno.
Proteção	0 ou 1	Indica se o terreno possui ou não proteção (muro ou cerca).
Posição	1 ou 2	Identifica a posição do terreno na quadra (meio ou esquina).
Frente	1 a 3	Identifica a largura do terreno. Sabendo que um terreno de frente com maior metragem possui uma melhor valorização.
Ponto Comercial	0 a 3	Sabendo que os terrenos localizados em zona de comércio ou de moradia, o terreno é mais ou menos valorizado. Esta variável identifica os vários níveis de localização.
Área do terreno	m^2	Quantifica a área do terreno.

Fonte: Imobiliária Tapowik, 2005.

3.1.3.2 Amostra

A amostra foi constituída por 119 imóveis. Sendo 44 apartamentos, 51 casas e 24 terrenos localizados na área urbana da Cidade de Campo Mourão – PR, dos quais 80 estão localizados na área central.

3.2 METODOLOGIA PARA O DESENVOLVIMENTO DA PESQUISA

A metodologia aqui proposta procura determinar classes homogêneas de apartamentos, casas e terrenos através de uma Análise de Agrupamento aplicada à amostra considerada. Utilizaram-se as técnicas de inferência, no nível de avaliação rigorosa, de acordo com NB-502/89 (avaliação de imóveis urbanos), com o desenvolvimento de um programa em MAT LAB, para o processamento dos dados e o *Software Excel* ®, por ser um aplicativo de

uso quase universal.

A Análise de Agrupamento foi aplicada para juntar imóveis semelhantes. Utilizou-se a Distância Euclidiana e a ligação pelo método de Ward. Então, a partir dos grupos formados (*clusters*), aplicou-se o método de Componentes Principais, procurando eliminar o problema da multicolinearidade que pode ocorrer na regressão do preço. Assim, conservaram-se as primeiras componentes e as que constituem um resumo de informação mais importante da estrutura de covariância. Com a finalidade de alcançar um dos objetivos deste, que é a obtenção do modelo de precisão, foi desenvolvido um estudo com a técnica da Regressão Linear Múltipla, para prever dentro de cada agrupamento o valor de um novo imóvel. As primeiras ferramentas descritas são para atender o objetivo de estudo das variáveis que participam da construção do modelo de Regressão Linear Múltipla.

3.2.1 Considerações Para a Construção do Modelo

As etapas, ou roteiro, que são necessárias para construir um modelo matemático através de critérios multivariados usando a Regressão Linear Múltipla com a finalidade de estimar valores de imóveis urbanos em Campo Mourão são apresentados a seguir.

3.2.1.1 Identificação das Variáveis Independentes

Uma das dificuldades existentes na avaliação de imóveis é a determinação das variáveis que influenciam no seu valor. São muitos os fatores que devem ser levados em consideração, mas nem sempre se pode desenvolver um único modelo representativo da realidade do conjunto do mercado de imóveis. Um dos aspectos mais importantes na avaliação de imóveis é a seleção das variáveis independentes que possam ser utilizadas na regressão, que são aquelas que têm influência na formação do preço, pois são variáveis importantes na formação de valor de uma determinada categoria ou subconjuntos de imóveis, não necessariamente são as mesmas que para outro subconjunto, inclusive dentro de uma mesma região. As possíveis variáveis independentes (ou explicativas) que podem influenciar no preço de um imóvel devem ser listadas a priori. A definição das variáveis explicativas preliminarmente economiza tempo e diminui o custo de execução da pesquisa. Em algumas ocasiões é necessário desconsiderar alguns dos elementos da amostra coletada pelo fato de serem elementos diferenciados do resto, razão pela qual sua presença afeta fortemente os valores globais da equação de regressão, não permitindo então a sua consideração no modelo de avaliação. As variáveis explicativas para a avaliação de imóveis são aquelas referentes a

todas as características físicas e locacionais do imóvel. No entanto, dentre todas as características físicas e locacionais relacionadas a um imóvel, nem todas são relevantes à formação de seu preço.

De forma geral e preliminar, podem-se citar como relevantes à formação do preço, as seguintes variáveis explicativas.

Apartamentos: Área total, área útil, número de dormitórios, número de suíte, número de carros na garagem, dependências de empregados, idade do imóvel, elevador, estado de conservação, padrão de acabamento, região de valorização imobiliária, distância à escola, etc.

Casas residenciais: Estado de conservação, área construída, área do terreno, localização, número de suítes, dependência de empregados, estrutura, padrão de acabamento, entre outras.

Terrenos: Área total, comprimento frontal (frente), localização, área comercial, etc.

Essa lista de variáveis explicativas tende a variar de município para município, dependendo das características de cada um. Para a cidade de Campo Mourão – PR, a variável explicativa mais relevante é a de proximidade do centro comercial.

3.2.1.2 Transformações de Variáveis

As variáveis que são definidas para a caracterização e localização de um imóvel são do tipo quantitativas ou qualitativas. Geralmente, estas variáveis precisam sofrer transformações para que então possam ser realizadas as análises. As variáveis qualitativas devem ser quantificadas através de uma codificação adequada. Em muitas situações são atribuídas para as variáveis qualitativas o valor 0 (zero) quando não tem a característica e 1 (um) caso contrário. Então, tem-se uma variável do tipo *dummy*, pronta para ser utilizada para análise. E outras variáveis que se referem às características qualitativas dos imóveis, como a conservação do imóvel (péssimo, regular, bom e ótimo); classificação do imóvel (baixo, normal e alto) e outras, são casos que são resolvidos dando pesos para a característica. Geralmente esses pesos são na ordem crescente, da situação menos favorável para a mais favorável. E ainda, quando uma variável pode vir a gerar um número muito grande de modalidades, algumas vezes, ela pode ser definida por uma escala numérica, atribuindo-se também pesos às modalidades, (por exemplo, a idade do imóvel).

As variáveis originais apresentadas nos Quadros 3.1, 3.2 e 3.3, foram transformadas utilizando a técnica multivariada Análise de Componentes Principais.

3.2.1.3 Análise Exploratória

Para o estudo de relacionamento entre as variáveis pode ser utilizado, entre outros, o coeficiente de correlação de linear de Pearson para as variáveis quantitativas e as qualitativas, sendo que, no segundo caso, elas devem ser transformadas em *dummy*. O coeficiente de correlação indica a existência ou não de relação linear significativa entre as variáveis independentes e a variável dependente, informação necessária para uso da regressão linear. Esses coeficientes, quando apresentam valores altos entre as variáveis independentes, indicam a possível existência de multicolinearidade, e ainda, o valor do determinante da matriz ($X'X$), quando é próximo de zero, também indica a existência de multicolinearidade. Apesar do coeficiente de correlação linear de Pearson e do determinante da matriz ($X'X$) indicar a existência da multicolinearidade, eles não a quantificam.

3.2.1.4 Análise dos Resíduos

A investigação da adequação do modelo é uma etapa do procedimento necessário na análise dos dados, tão importante quanto a sua construção. A plotagem dos resíduos é o instrumento usado para examinar o modelo. A análise gráfica dos resíduos é necessária para examinar o ajuste do modelo, ou seja, para confirmar se ele tem uma boa aproximação do verdadeiro sistema e para verificar se as suposições da regressão por mínimos quadrados não foram violadas (Montgomery, 1997).

3.2.1.5 Verificação da Adequação do Modelo

Um último passo que deve ser realizado antes de adotar o modelo para avaliação de imóveis, é verificar sua aplicabilidade. Inicialmente, deve-se fazer a Análise de Variância para testar a significância do modelo ajustado, no entanto, isto por si só não garante a qualidade das predições. A qualidade do ajuste pode ser testada comparando os valores preditos com os valores observados. O ajuste é tão bom, quanto maior for a quantidade de valores preditos próximos dos valores observados, isto é, com pequeno erro de predição. O valor do coeficiente de correlação linear R^2 é importante para definir a qualidade do modelo adotado.

3.3 ESTUDO DE CASO

Neste trabalho, efetuou-se um estudo no mercado imobiliário da Cidade de Campo Mourão – PR, restringindo-se ao segmento de imóveis urbanos, cujo objetivo será modelar este mercado através da Análise de Regressão, pautando-se da Análise Multivariada e estimar ou calcular o valor de venda de apartamentos, casas e terrenos, de forma absolutamente objetiva, sem qualquer "opinião" originária da subjetividade intrínseca do ser humano.

O *Software* utilizado para a construção da tabela de dados e as devidas transformações foi o *Excel*.

As matrizes de dados resultantes (apartamentos, casas e terrenos) foram submetidas aos tratamentos estatísticos descritos no segundo capítulo. Todos os resultados estatísticos foram obtidos através do programa desenvolvido em *Matlab*, versão 7.0 e comparados com os resultados obtidos no *Software Statgraphics*.

Em primeiro lugar as matrizes de dados foram submetidas à Análise de Agrupamentos (*clusters*) hierárquicos, utilizando-se a Distância Euclidiana, sendo que os agrupamentos foram feitos por meio da ligação de Ward para formar as classes homogêneas. Após várias simulações, ficou claro que o número ótimo de classes a considerar para o caso de residências seria quatro e para os apartamentos três, enquanto para os terrenos apenas duas classes. Após a formação das classes homogêneas realizou-se uma Análise Discriminante para avaliar a consistência das classes obtidas. Em seguida realizou-se uma Análise de Componentes Principais para cada classe formada para os tipos de imóveis. Com a obtenção dos escores das Componentes Principais para explicar a variação total, substituiu-se as variáveis explicativas originais. Por fim foi desenvolvido um modelo de Regressão Linear Múltipla para cada uma das classes de cada tipo de imóvel. Considerou-se como variável resposta o preço de venda à vista, que se denominou valor.

4 APLICATIVO AMI DESENVOLVIDO

Para resolver os problemas de trabalho manual, necessário no *Software Statgraphics,* foi desenvolvido um programa denominado AMI (Análise Multivariada de Imóveis) em *Matlab 7.0*. O programa desenvolvido está no Anexo 1.

4.1 ALGORITMO DO PROGRAMA AMI

O algoritmo usado para o desenvolvimento do aplicativo foi o seguinte:

1. Leitura de dados da planilha *Excel*. Esta base de dados contém as variáveis em colunas e as observações em linhas, estando a variável dependente na última coluna.

2. Após a leitura, o programa mostra as três opções de cálculo para a regressão com os seguintes tratamentos das observações:

   a. Análise de Agrupamentos e Análise de Componentes Principais.

   b. Análise Fatorial.

   c. Regressão Múltipla Simples (sem nenhum tratamento).

a1. Formação de grupos (*clusters*), nesta etapa o usuário seleciona um tipo de distância dentre as seguintes:

   $d = 1$, EUCLIDEAN (distância euclidiana).

   $d = 2$, SEUCLIDEAN (quadrado da distância euclidiana).

   $d = 3$, CITYBLOCK (distância Manhattan)

   $d = 4$, MAHALANOBIS (distância estatística).

   $d = 5$, MINKOSWSKI (usando $n = 3$)

assim como também os tipos de ligação entre os grupos, tais como:

$t = 1$, SINGLE (Vizinho mais próximo).

$t = 2$, COMPLETE (Vizinho mais distante)

$t = 3$, AVERAGE (Média das distancias).

$t = 4$, WARD (Método de Ward).

O programa apresenta um dendrograma para a avaliação de formação de K grupos, o programa "pergunta" o número de grupos a formar.

a2. Com os K grupos formados o programa reagrupa as n observações em K matrizes para a análise das componentes principais.

a3. Para o grupo i são verificadas as análises das componentes principais, mostrando uma tabela com a proporção da variância explicada pelos autovalores da matriz covariância e o programa pede ao usuário o numero de p componentes principais a serem usados.

a4. Para o grupo i, tendo as componentes principais, é feita a análise de regressão múltipla, mostrando a função de regressão, o parâmetro R^2 (coeficiente de correlação linear) e os parâmetros da avaliação da variância F e p.

a5. No caso de calcular o valor de um imóvel, as observações (variáveis independentes) devem ser colocadas na última linha com a variável dependente igual a zero. Neste caso os processos 1, a1 e a2 são iguais e será analisado unicamente o grupo j, onde a última observação pertence. Em continuação são realizados os processos a3 e a4 para o grupo j *e é* calculado o valor da variável dependente com a função regressão, mostrando os parâmetros das estatísticas.

Para o caso da opção 2 (Análise Fatorial), tem-se a seguinte seqüência do algoritmo:

b1. O AMI avalia os autovalores da matriz de correlação das observações e mostra a proporção da variância explicada acumulada para o usuário selecionar o número N de fatores a serem analisados.

b2. O programa faz a regressão dos escores rotacionados (varimax) de todas as observações (sem a última coluna que é o preço) para serem usados na regressão.

b3. Para os escores de todas as observações com n componentes são calculados os coeficientes da equação de regressão múltipla, mostrando a equação de regressão, o parâmetro R^2 (coeficiente de correlação linear) e os parâmetros da avaliação da variância F e p.

b4. Se o preço da última observação for zero, será estimado o valor da variável dependente (preço) da mencionada observação.

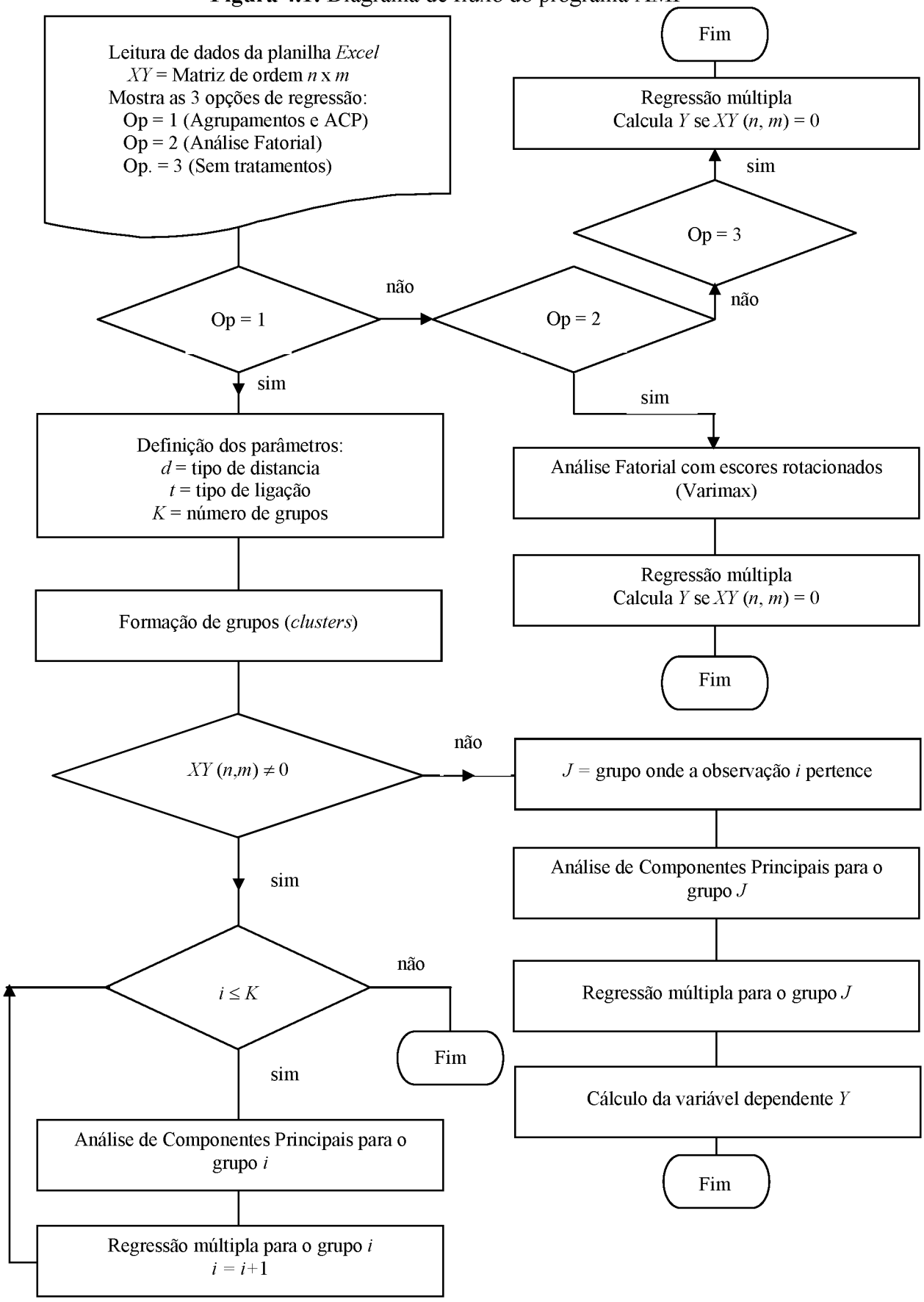
Para o caso da opção 3 (Regressão Múltipla Simples), a seqüência do algoritmo é a seguinte:

c1. Com base nas observações (com todas as variáveis) são calculados os coeficientes da equação de regressão múltipla, mostrando a equação de regressão, o parâmetro R^2 (coeficiente de correlação linear) e os parâmetros da avaliação da variância F e p.

c2. Se o preço da última observação for zero, será estimado o valor da variável dependente (preço) da mencionada observação.

Na figura 4.1 a seguir, é apresentado o diagrama de fluxo do programa AMI.

Figura 4.1: Diagrama de fluxo do programa AMI



Fonte: O Autor, 2005.

4.2 AVALIAÇÃO DO AMI

Os resultados foram comparados com os resultados do *Software Statgraphics Plus 5.0*, e a seqüência para realizar o objetivo (regressão múltipla) neste aplicativo é a seguinte:

1) Abrir o programa e na planilha própria colar a base de dados do *Excel*. Na seqüência faz-se a análise de agrupamentos com as opções de K grupos, variáveis padronizadas.

2) Uma vez formados os grupos, o programa indica simplesmente quais observações pertencem aos K grupos. No *Excel* faz-se a redistribuição das observações em K grupos, que seria em K sub-pastas. Este trabalho é feito manualmente demandando tempo e cuidado, pois se fossem centenas ou milhares de observações, o tempo seria de algumas horas ou dias.

3) Com as observações enquadradas em cada grupo feitas no *Excel*, colar na planilha própria os dados. Para este grupo analisam-se as componentes principais com as opções de número das componentes principais (i). Os resultados são as i componentes do grupo, salvar os resultados na planilha própria do Statgraphics.

4) Na seqüência realiza-se a regressão múltipla tendo como dados as componentes principais e verificam-se os parâmetros das estatísticas da regressão.

Neste processo, o tempo de execução total dos itens 1, 2, 3 e 4 é de várias horas, dependendo da prática com o programa *Statgraphics* e da segurança de que tudo o que é feito manualmente não tenha erros.

Usando o programa AMI, o tempo é quase nulo, pois tudo é realizado pelo programa. Os resultados são idênticos aos resultados do *Statgraphics*.

4.3 DESCRIÇÃO DO PROGRAMA AMI

A seqüência das janelas apresentadas pelo AMI é:

Figura 4.2: Entrada de dados usando um arquivo do *Excel*

```
-----------------------------------------------------------------------
 *  ACESSANDO OS DADOS: Selecione o arquivo de dados no excel       *
-----------------------------------------------------------------------
```

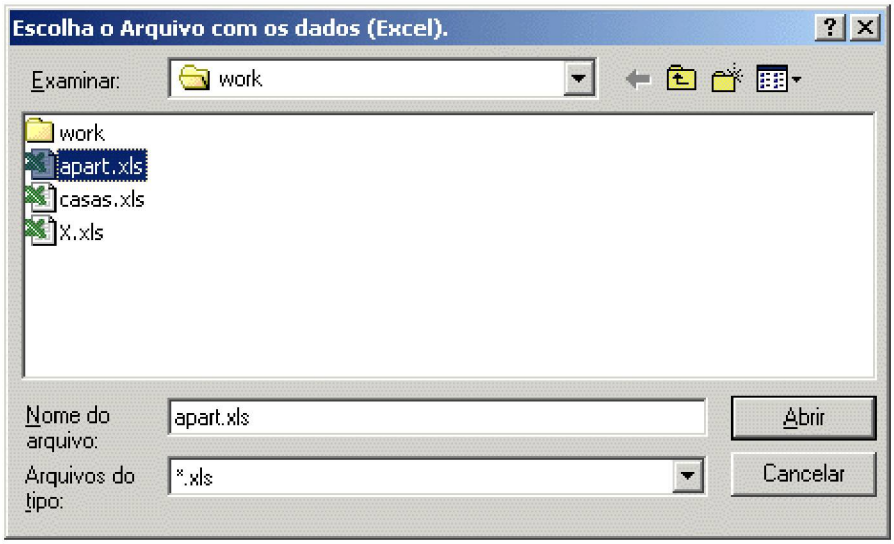


Figura 4.3: Mostrando os tipos de opções de cálculo para a regressão

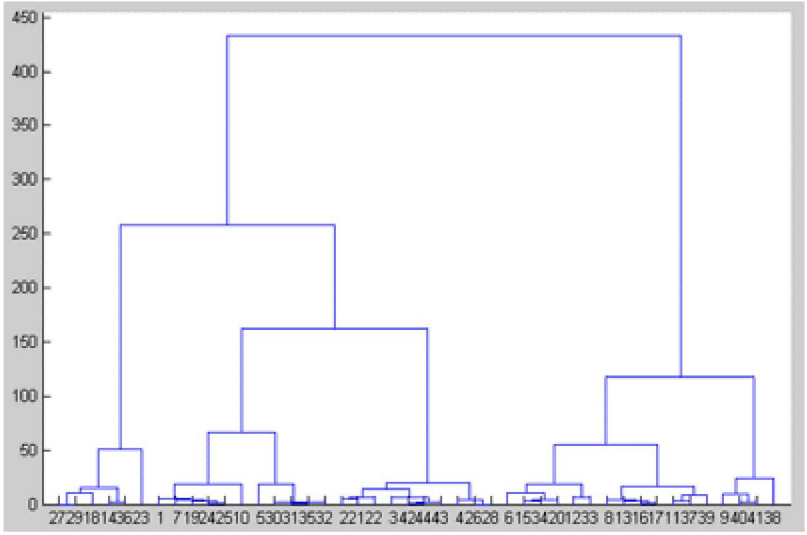
```
-----------------------------------------------------------------------
 *  ACESSANDO OS DADOS: Selecione o arquivo de dados no excel       *
-----------------------------------------------------------------------


arquivo XLS = C:\valdir\apart.xls

total de observações = 44
total de variáveis = 22

-----------------------------------------------------------------------
 *         OPÇÕES DE ANÁLISE PARA A REGRESSÃO                        *
***********************************************************************
-----------------------------------------------------------------------
 * (1) Análise por agrupamentos e componentes principais *
-----------------------------------------------------------------------
 * (2) Análise fatorial                                  *
-----------------------------------------------------------------------
 * (3) Análise simples de Regressão                                   *
-----------------------------------------------------------------------

ENTRAR COM O TIPO DE OPÇÃO  = 1
```

Figura 4.4: Com a opção 1 mostrando os tipos de distância e ligações

```
-------------------------------------------------------------------
 *  PROCESSO 1 : AGRUPANDO - CLUSTERING                            *
*******************************************************************
-------------------------------------------------------------------
 *             OPÇÃO DE TIPOS DE DISTÂNCIA                         *
-------------------------------------------------------------------
 *   d = 1, EUCLID    (distância euclidiana)                       *
 *   d = 2, SEUCLID (quadrado da distância euclidiana)             *
 *   d = 3, CITYBLOCK  (distância Manhattan)                       *
 *   d = 4, MAHAL (distância estatistica)                          *
 *   d = 5, MINKOSWSKI  (usando n=3)                               *
-------------------------------------------------------------------

ENTRAR COM O TIPO DE DISTÂNCIA d = 1
melhor agrupamento = "average" sendo o cophenet = 0.74361

-------------------------------------------------------------------
 *             OPÇÃO DE TIPOS DE LIGAÇÃO                           *
-------------------------------------------------------------------
 *   t = 1, SINGLE   (Vizinho mais próximo)    *  Cophenet = 0.72535
 *   t = 2, COMPLETE (Vizinho mais distante)   *  Cophenet = 0.73425
 *   t = 3, AVERAGE  (Média das distâncias)    *  Cophenet = 0.74361
 *   t = 4, WARD     (Método de Ward)          *  Cophenet = 0.6391
-------------------------------------------------------------------

ENTRAR COM O TIPO DE LIGAÇÃO t = 4
```

Este gráfico mostra que a melhor ligação entre os grupos é “average”, ou seja, que a média das distâncias com índice cofenético é igual a 0,74361.

Figura 4.5: Dendrograma apresentado para a escolha de número de grupos



Neste processo podem ser selecionados 2, 3 ou 5 grupos, mas foram selecionados 2 grupos.

Figura 4.6: Grupos formados

```
Digite o número de grupos K = 2

-----------------------------------------------------------------
 *  FIM DO PROCESSO 1:   GRUPOS FORMADOS                         *
-----------------------------------------------------------------
GRUPO 1 = 6 8 9 11 12 13 15 16 17 20 33 34 37 38 39 40 41 (17)
GRUPO 2 = 1 2 3 4 5 7 10 14 18 19 21 22 23 24 25 26 27 28 29 30 31 32 35 36 42 43 44 (27)
-----------------------------------------------------------------
Tecle ENTER para continuar.....
```

Figura 4.7: Processo 2: análise de Componentes Principais para o grupo 1

```
-----------------------------------------------------------------
 *  PROCESSO 2:   COMPONENTES PRINCIPAIS                         *
-----------------------------------------------------------------

ANALISANDO O GRUPO 1
*****************************************************************
A variável N° [3] foi eliminada por possuir valores iguais


* PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *
*    AUTOVALORES DA MATRIZ CORRELAÇÃO    *

ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL.
        LORES     (EM %)    ACUM. (%)
----------------------------------------
  1     6.1608    29.34       29.34
  2     3.6422    17.34       46.68
  3     3.5396    16.86       63.54
  4     1.6849     8.02       71.56
  5     1.4289     6.80       78.36
  6     1.1578     5.51       83.88
  7     1.0491     5.00       88.87
  8     0.7977     3.80       92.67
```

Nesta etapa foi excluída a coluna 3 (variável 3) por possuir valores iguais, uma característica deste grupo. O programa pergunta o número de componentes a serem usadas, neste caso, o programa indica que tem que ser menor a 17 componentes, pois a condição é que o número de observações para uma regressão tem que ser maior ao número de componentes.

Figura 4.8: Perguntando o número de componentes e o processo de regressão

```
 17      0.0000      0.00       100.00
 18      0.0000      0.00       100.00
 19     -0.0000     -0.00       100.00
 20     -0.0000     -0.00       100.00
 21     -0.0000     -0.00       100.00

----------------------------------------
Digite o número de componentes  < 17 = 10

------------------------------------------------------------------------
 *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                                    *
------------------------------------------------------------------------

REGRESSÃO PARA O GRUPO 1
------------------------------------------------------------------------
Y = -231.1999*CP1  + 1782.8403*CP2 -9175.1104*CP3 -3518.4216*CP4 + 10:
------------------------------------------------------------------------

ESTATÍSTICA DO MODELO 1
------------------------------------------------------------------------
R2          F          p
----------------------------------------
0.9006      5.4375    0.02527
```

O número de componentes usado neste caso será de 10 com uma variância explicada acumulada de 97,22 %. A figura 4.8 também mostra a equação de regressão para o grupo 1:

Y = -231,1999*CP1 + 1 782,8403*CP2 – 9 175,1104*CP3 – 3 518,4216*CP4 + + 10 126,8349*CP5 + 507,8941*CP6 + 10 383,084*CP7 + 1 744,384*CP8 + + 5 895,8829*CP9 + 11 933,8859*CP10 + 21 461,2404, com R^2 = 0,9006 e a análise da variância com F = 5,4375 e com valor da significância de 0,02527, o que significa que se rejeita a hipótese de não haver regressão.

Analogamente os resultados são mostrados para o grupo 2:

Y = -12 140,2077*CP1 + 7 491,9091*CP2 – 7 889,13*CP3 + 2 967,7093*CP4 + + 22 871,6474*CP5 – 7 286,1078*CP6 – 9 574,2814*CP7 + 10 397,2659*CP8 + + 1 479,4856*CP9 -23 796,0254*CP10 – 117 191,3768, com R^2 = 0,7857, F = 5,8658 e p = 0,00095.

Ambos os grupos têm $R^2 > 0,75$ indicando boa correlação.

No caso de calcular o valor da variável dependente de uma nova observação, coloca-se esta observação na última linha na planilha do *Excel* com o preço igual a zero. Veja o exemplo na continuação:

Figura 4.9: Calculando o valor do 45º apartamento

	A	B	C	V	W
1	1	3	2	2	130000
2	2	3	1	6	85000
3	3	3	1	6	80000
40	41	3	1	4	30000
41	42	3	1	4	40000
42	43	3	1	6	100000
43	44	1	1	6	90000
44	45	3	1	6	0

O valor do mercado é de R$110 000,00, mas, para testar esta observação, foi colocado um valor de 0.

Figura 4.10: Estimando o valor da última observação

```
Digite o número de componentes  < 17 = 10

-------------------------------------------------------------------
 *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                               *
-------------------------------------------------------------------

REGRESSÃO PARA O GRUPO 2
-------------------------------------------------------------------
Y = -11814.6919*CP1  + 7419.3319*CP2 -7810.4737*CP3 + 3892.5137*CP4 + 23118.6218*CP5 -779
-------------------------------------------------------------------

ESTATÍSTICA DO MODELO 2
-------------------------------------------------------------------
R2         F          p
-----------------------------------------
0.7827     5.4044   0.00186

-----------------------------------------

Componentes principais da última observação  (cp1,cp2,...)
-50.5967 -6.5033 21.4963 29.7843 67.3294 13.1197 -34.1355 -14.4874 -91.3018 72.6822

Y(variável dependente) da última observação  = 132609.8185
*******************************************************************
FIM DO PROGRAMA
```

A última observação pertence ao segundo grupo e o programa simplesmente analisou este grupo e o valor estimado para esta observação é de 132 609.81, com um coeficiente de regressão de 0,7827 com 10 componentes.

Na seqüência será analisado por Análise Fatorial, digitando 2 na opção de análise:

Figura 4.11: Analisando os apartamentos por Análise Fatorial

```
------------------------------------------------------------------------
 *  ACESSANDO OS DADOS: Selecione o arquivo de dados no excel       *
------------------------------------------------------------------------


arquivo XLS = C:\valdir\apart.xls

total de observações = 44
total de variáveis = 22

------------------------------------------------------------------------
 *          OPÇÕES DE ANÁLISE PARA A REGRESSÃO                        *
************************************************************************
------------------------------------------------------------------------
 * (1) Análise por agrupamentos e componentes principais *
------------------------------------------------------------------------
 * (2) Análise fatorial                                  *
------------------------------------------------------------------------
 * (3) Análise simples de Regressão                                    *
------------------------------------------------------------------------

ENTRAR COM O TIPO DE OPÇÃO  = 2
```

Após mostrar os autovalores e a proporção da variância acumulada, o programa pergunta o número de fatores a serem considerados.

Figura 4.12: AMI perguntando o número de fatores

```
* PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *
*    AUTOVALORES DA MATRIZ CORRELAÇÃO    *

ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL.
       LORES     (EM %)    ACUM. (%)
-----------------------------------------
  1    6.6781    31.80       31.80
  2    2.7531    13.11       44.91
  3    2.6123    12.44       57.35
  4    1.9697     9.38       66.73
 15    0.1890     0.90       97.81
 16    0.1655     0.79       98.60
 17    0.1082     0.52       99.11
 18    0.0669     0.32       99.43
 19    0.0604     0.29       99.72
 20    0.0421     0.20       99.92
 21    0.0167     0.08      100.00

---------------------------------------

Digite o número de fatores  < 21 = 10
```

Depois de digitado o número de componentes para a análise da regressão com 10 componentes, o programa mostra o seguinte:

Figura 4.13: Regressão com Análise Fatorial

```
-------------------------------------------------------------------------
 *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                                     *
-------------------------------------------------------------------------

REGRESSÃO
-------------------------------------------------------------------------
Y = 44065.0343*F1  + 8908.6589*F2 -17686.0306*F3 + 7885.606*F4 -5654.08
-------------------------------------------------------------------------

ESTATÍSTICA DO MODELO
-------------------------------------------------------------------------
R2          F          p
-----------------------------------------
0.7916     12.1583    0.00000

-----------------------------------------
*************************************************************************

Y(variável dependente) da última observação  = 116002.8105
*************************************************************************
*FIM DO PROGRAMA *
```

A equação da regressão com 10 fatores é:

Y = 44 065,0343*F1 + 8 908,6589*F2 – 17 686,0306*F3 + 7 885,606*F4 -- 5 654,0828*F5 + 11 501,3934*F6 + 6 105,4323*F7 + 15 266,1667*F8 + 16 396,1855*F9 -- 14 413.7115*F10 + 125 068,2457.

O valor da última observação é de 116 002,81 com um coeficiente de regressão de 0,7916, o que significa que a análise da variância rejeita a hipótese de não haver regressão.

Da mesma forma pode-se optar pela terceira opção e fazer uma regressão diretamente com as variáveis, sem nenhum tratamento.

Figura 4.14: Regressão considerando todas as variáveis

```
-------------------------------------------------------------------------
 *  ACESSANDO OS DADOS: Selecione o arquivo de dados no excel       *
-------------------------------------------------------------------------


arquivo XLS = C:\valdir\apart.xls

total de observações = 44
total de variáveis = 22

-------------------------------------------------------------------------
 *          OPÇÕES DE ANÁLISE PARA A REGRESSÃO                          *
*************************************************************************
-------------------------------------------------------------------------
 * (1) Análise por agrupamentos e componentes principais *
-------------------------------------------------------------------------
 * (2) Análise fatorial                                  *
-------------------------------------------------------------------------
 * (3) Análise simples de Regressão                                     *
-------------------------------------------------------------------------

ENTRAR COM O TIPO DE OPÇÃO  = 3
```

Figura 4.15: Resultados da regressão

```
     -------------------------------------------------------------------------
      *  PROCESSO 1:   REGRESSÃO MÚLTIPLA                                    *
     -------------------------------------------------------------------------

     REGRESSÃO
     -------------------------------------------------------------------------
     Y = 11165.4093*V1  + 1756.9593*V2 + 26018.5638*V3 + 112463.3882*V4 + 418.4
     -------------------------------------------------------------------------

     ESTATÍSTICA DO MODELO
     -------------------------------------------------------------------------
     R2         F         p
     ----------------------------------------
     0.8936     8.4010    0.00000

     ----------------------------------------
     *************************************************************************

     Y(variável dependente) da última observação  = 98949.7482
     *************************************************************************
*FIM DO PROGRAMA *
```

A equação da regressão é de:

Y = 11 165,4093*V1 + 1 756,9593*V2 + 26 018,5638*V3 + 112 463,3882*V4 + + 418,4552*V5 – 1 647,8111*V6 + 2 186,2769*V7 – 13 924,2677*V8 + 14 916,216*V9 + + 43 800,5883*V10 – 32 571,4002*V11 + 22 319,6361*V12 + 52 736,861*V13 - - 1 830,7984*V14 + 3 151,1696*V15 – 15 814,0065*V16 + 5 499,5626*V17 + + 11 482,7732*V18 -13 218,9298*V19 + 13 260,6722*V20 + 4 652,535*V21 -164 341,852

Como se pode observar, os parâmetros são bons e o valor da última observação estimada é de 98 949,7482.

Analisando os três casos, quem mais se aproxima do valor é a opção 2:

Tabela 4.1: Comparação dos valores estimados pelas 3 opções

opções	valor	estimado	diferença	diferença %
opção 1	110 000,00	132 609,81	-22 609,81	20,55
opção 2	110 000,00	116 002,81	-6 002,81	5,457
opção 3	110 000,00	98 949,75	11 050,25	10,046

Neste caso pode-se falar que o preço de R$110 000,00 é um estimativo do valor de mercado, cujas regras de decisão variam.

4.4 COMPARAÇÃO DOS RESULTADOS ENTRE O PROGRAMA STATGRAPHICS E O AMI

i) Na formação de grupos (*cluster*) ambos com distância euclidiana e agrupamento de Ward.

No *Statgraphics* os 2 grupos formados são:

GRUPO 1 = 1 4 5 6 7 10 13 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 35 36 (27)

GRUPO 2 = 2 3 8 9 11 12 15 33 34 37 38 39 40 41 42 43 44 (17)

No AMI os 2 grupos formados são:

GRUPO 1 = 6 8 9 11 12 13 15 16 17 20 33 34 37 38 39 40 41 (17)

GRUPO 2 = 1 2 3 4 5 7 10 14 18 19 21 22 23 24 25 26 27 28 29 30 31 32 35 36 42 43 44 (27)

Na avaliação dos resultados, as observações comuns são:

Os grupos têm as mesmas quantidades com os nomes dos grupos diferentes:

Similaridade entre os Grupos com 27 observações = 22 de 27 = 81,5%

Similaridade entre os Grupos com 17 observações = 12 de 17 = 70,6%

Essa diferença existe quando as distâncias são iguais, ficando o programa com a decisão de formação dos grupos.

ii) Na análise de componentes principais, para este caso usam-se os grupos formados pelo programa AMI e rodando no programa *Statgraphics*.

Com os escores das 10 componentes principais (col1, col2,..., col10, col11 = preço) extraídos a partir da matriz de correlação das observações, a regressão obtida no *Statgraphics* é:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	7.58821E9	10	7.58821E8	5.44	0.0253
Residual	8.37324E8	6	1.39554E8		
Total (Corr.)	8.42553E9	16			

R-squared = 90.0621 percent
R-squared (adjusted for d.f.) = 73.4988 percent
Standard Error of Est. = 11813.3
Mean absolute error = 5300.13
The equation of the fitted model is

Col_11 = 21 461,3 – 231,192*Col_1 + 1 782,82*Col_2 – 9 175,1*Col_3 -
- 3 518,4*Col_4 + 10 126,8*Col_5 + 507,89*Col_6 + 10 383,1*Col_7 +
+ 1 744,42*Col_8 + 5 895,85*Col_9 + 11 933,9*Col_10

No AMI para o grupo 1:
REGRESSÃO MÚLTIPLA
Y = - 231,1999*CP1 + 1 782,8403*CP2 - 9 175,1104*CP3 - 3 518,4216*CP4 +
+ 10 126,8349*CP5 + 507,8941*CP6 + 10 383,084*CP7 + 1 744,384*CP8 +
+ 5 895,8829*CP9 + 11 933,8859*CP10 + 21 461,2404

ESTATÍSTICA DO MODELO
R^2 = 0,90062 F = 5,4375 p = 0,025266

Como se pode observar, os resultados da regressão são iguais.

Para o grupo 2, os resultados no *Statgraphics* são:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	6.45385E10	10	6.45385E9	5.79	0.0010
Residual	1.78356E10	16	1.11472E9		
Total (Corr.)	8.23741E10	26			

R-squared = 78.3481 percent
R-squared (adjusted for d.f.) = 64.8156 percent
Standard Error of Est. = 33387.5
Mean absolute error = 18693.2
The equation of the fitted model is

Col_11 = - 11 3123,0 – 11 708,1*Col_1 + 7 395,67*Col_2 – 7 784,78*Col_3 +
+ 4 194,68*Col_4 + 23 199,4*Col_5 – 7 958,46*Col_6 – 7 123,08*Col_7 +
+ 12 618,3*Col_8 + 2 370,04*Col_9 – 21 388,6*Col_10

E no AMI tem-se:
REGRESSÃO MÚLTIPLA
Y = -11 708,2947*CP1 + 7 395,6095*CP2 -7 784,6064*CP3 + 4 194,7926*CP4 +
+ 23 199,3472*CP5 -7 958,6098*CP6 -7 123,0317*CP7 + 12 618,4508*CP8 +
+ 2 369,774*CP9 -21 389,0761*CP10 -113 121,4588

ESTATÍSTICA DO MODELO
R^2 = 0,78348 F = 5,7897 p = 0,0010212.

Analisando os resultados, pode-se afirmar que são aproximadamente iguais. Com valores de R^2 >75 %.

4.4.1 Resumo dos Resultados para Apartamentos

Dados com 44 observações e 22 variáveis:

GRUPO 1 = 6 8 9 11 12 13 15 16 17 20 33 34 37 38 39 40 41 (17)

Y = -231,1999*CP1 + 1 782,8403*CP2 – 9 175,1104*CP3 – 3 518,4216*CP4 +
+ 10 126,8349*CP5 + 507,8941*CP6 + 10 383,084*CP7 + 1 744,384*CP8 +
+ 5 895,8829*CP9 + 11 933,8859*CP10 + 21 461,2404.

Neste grupo foi desconsiderada a variável 3 (elevador = 1) por possuir todos os valores iguais a 1 e R^2 = 0.9006.

GRUPO 2 = 1 2 3 4 5 7 10 14 18 19 21 22 23 24 25 26 27 28 29 30 31 32 35 36 42 43 44 (27)

Y = -11 708,2947*CP1 + 7 395,6095*CP2 – 7 784,6064*CP3 + 4 194,7926*CP4 + + 23 199,3472*CP5 – 7 958,6098*CP6 – 7 123,0317*CP7 + 12 618,4508*CP8 + + 2 369,774*CP9 -21 389,0761*CP10 – 113 121,4588.

Neste grupo foram desconsideradas as variáveis 4 (localização = 1) e 11 (suíte = 1) por possuir todos os valores iguais e R^2 = 0.7835.

4.4.2 Resumo dos Resultados para Casas

Dados com 51 observações e 19 variáveis:

Opção 1: (Análise de Agrupamentos e Componentes Principais)

Tipo de distância euclidiana e tipo de ligação ward, formada em 4 grupos.

GRUPO 1 = 8 12 13 14 24 25 26 28 31 34 44 48 (12)

Equação da regressão com 10 componentes principais:

Y = 2 352,8701*CP1 + 14 682,8764*CP2 – 26 680,0908*CP3 – 2 143,8816*CP4 - -15 706,917*CP5 – 16 068,2388*CP6 – 807,695*CP7 + 2 837,1778*CP8 + + 30 340,9987*CP9 + 7 161,036*CP10 + 335 172,8204.

Neste grupo foi desconsiderada a variável piscina por possuir todos os valores iguais à zero (sem piscina) e R^2 = 0.9123.

GRUPO 2 = 1 3 4 5 6 9 10 15 16 17 19 20 22 29 30 32 36 38 40 43 45 46 47 49 51 (25)

Equação da regressão com 10 componentes principais:

Y = 12 220,9925*CP1 + 307,181*CP2 – 3 553,3992*CP3 + 8 534,9659*CP4 - - 21 673,6631*CP5 + 1 399,9938*CP6 – 406,8441*CP7 + 24 863,9387*CP8 + + 3 076,8517*CP9 + 7 133,0792*CP10 – 135 246,2812.

Neste grupo não foi desconsiderada nenhuma variável, pois possuem valores diferentes e R^2 = 0.9123.

GRUPO 3 = 2 7 11 21 35 37 39 (7)

Equação da regressão com 05 componentes principais:

Y = 3 381,742*CP1 + 25 395,7101*CP2 – 3 085,5499*CP3 + 2 998,7434*CP4 + + 35 063,3867*CP5 + 38 609,3956

Neste grupo foram desconsideradas as variáveis 2 (garagem = 1) e 16 (lavanderia = 1), pois possuem valores iguais e R^2 = 0. 9509.

GRUPO 4 = 18 23 27 33 41 42 50 (7)

Equação da regressão com 05 componentes principais:

Y = -8 719,292*CP1 – 4 037,9719*CP2 – 4 617,7801*CP3 – 1 899,7625*CP4 - - 4 786,5778*CP5 – 69 709,5748

Neste grupo foram desconsideradas as variáveis 11 (estrutura = 3) e 13(piscina = 0), pois possuem valores iguais e R^2 = 0.9994.

Opção 2: (Análise Fatorial)

Equação da regressão com 10 fatores:

Y = 17 160,2815*F1 + 18 097,7924*F2 + 3 788,0114*F3 + 11 702,6414*F4 + + 21 321,633*F5 + 5 647,0059*F6 + 11 399,342*F7 + 21 162,3526*F8 + 30 948,3984*F9 + + 5 791,8663*F10 + 96 745,098 e R^2 = 0.7847.

Opção 3: (Regressão Múltipla Simples, com todas as variáveis).

Y = 9 613,4525*V1 + 22 164,1341*V2 + 671,2289*V3 + 4 524,6907*V4 - - 930,9592*V5 + 6 691,6853*V6 + 225,8114*V7 + 85,0273*V8 + 12 495,6409*V9 - - 10 415,1484*V10 + 21 678,8472*V11 – 1 476,9568*V12 + 17 263,8745*V13 - - 18 058,2558*V14 + 6 853,9381*V15 – 20 153,3818*V16 + 5 775,6852*V17 - - 1 684,3853*V18 -94 861,0962 e R^2 = 0. 8366.

Colocando zero no preço da última observação (preço = 40000).

Tabela 4.2: Comparação dos valores estimados pelas 3 opções

opções	valor	estimado	diferença	diferença %
opção 1	40 000,00	42 734,52	-2 734,52	6,836
opção 2	40 000,00	47 381,33	-7 381,33	18,45
opção 3	40 000,00	58 096,46	-18 096,46	45,24

4.4.3 Resumo dos Resultados para Terrenos

Dados com 24 observações e 11 variáveis:

Opção 1: (Análise de Agrupamento e Componentes Principais)

Tipo de distância euclidiana e tipo de ligação ward, formada em 4 grupos.

GRUPO 1 = 1 2 3 4 5 8 9 10 11 12 16 18 19 20 21 22 23 24 (18)

Y = 15 508,3498*CP1 + 6 408,5873*CP2 + 19 369,2796*CP3 + 14 705,3493*CP4 - - 4 078,808*CP5 – 4 154,5414*CP6 + 3 939,7138*CP7 + 5 781,7527*CP8 + 5 864,1289*CP9 - - 16 556,4398*CP10 – 50 239,7229 e R^2 = 0.9799.

GRUPO 2 = 6 7 13 14 15 17 (6)

As variáveis de números 3 (pólo = 1), 4 (frente = 3), 8 (inclinado = 0) e 10 (pavimentação = 1) foram eliminadas por possuírem valores iguais. Dessa forma, a equação é dada por:

Y = 66 082,5699*CP1 – 2 285,7628*CP2 + 12 034,3121*CP3 + 21 812,3321*CP4 - - 28 360,0695*CP5 – 28 291,8593 e R^2 = 1.0.

Opção 2: (Análise Fatorial) com 5 fatores

Y = 21 269,8524*F1 – 5 876,4286*F2 + 22 094,3272*F3 + 55 872,7757*F4 + + 19 952,1654*F5 + 74 458,3333 e R^2 = 0. 8491.

Opção 3: (Regressão Múltipla Simples)

Y = 14 982,0119*V1 + 42 917,3384*V2 – 2 101,3467*V3 + 1 628,9076*V4 + + 11,1196*V5 + 4 029,0109*V6 + 5 702,7019*V7 – 2 524,9226*V8 + 16 158,2245*V9 - - 3 402,4047*V10 – 65 237,4996 e R^2 = 0. 8643.

Colocando zero na variável preço na última observação: (preço = 30.000,00).

Tabela 4.3: Comparação dos valores estimados pelas 3 opções

opções	valor	estimado	diferença	%
opção 1	30 000,00	21 264,52	8 735,48	29,118
opção 2	30 000,00	18 216,87	11 783,13	39,277
opção 3	30 000,00	21 275,44	8 724,56	29,082

5 CONSIDERAÇÕES FINAIS

Em 1918, em nosso país mostram-se os primeiros estudos de avaliação de imóveis e já em 1923 foram introduzidos novos métodos de avaliação de terrenos, que a partir de 1929 começaram a ser sistematicamente aplicados. Nesse contexto a engenharia de avaliação no Brasil tomou corpo e continua crescendo e evoluindo nas técnicas de avaliação. Atualmente um grande número de profissionais e entidades desenvolve estudos nesse campo, visando dar ao tema o suporte científico necessário aos métodos técnicos até então utilizados (Fiker, 1997).

Este trabalho mostra a importância de usar um programa computacional para a avaliação de imóveis em qualquer cidade desde que se tenha uma base de dados atualizados e pode ser usado por engenheiros, arquitetos e agrônomos, cada um atuando em sua habilitação profissional, conforme normas e regulamentos do CREA, CONFEA, ABNT, leis municipais, estaduais e federais.

5.1 CONCLUSÕES

As conclusões pertinentes ao trabalho são:

- Com o tratamento estatístico dos dados das observações evitou-se problema de multicolinearidade e valores das características iguais podem ser eliminados para evitar divisões por zero no cálculo da correlação. Ainda, reduziu-se a quantidade de variáveis para a regressão, ou seja, a matriz de dados ficou com a ordem mais baixa.

- Já com a Análise Fatorial também pode se fazer uma boa regressão evitando a multicolinearidade através do rotacionamento dos escores sem perda significativa de informação e obtendo, assim, variáveis não correlacionadas.

- O programa AMI, não é exclusivo para a cidade de Campo Mourão, pode ser usada em qualquer cidade, basta possuir os dados atualizados obedecendo ao formato de entrada num arquivo de Excel, onde as colunas são as variáveis e as

observações são as linhas. Recomenda-se que o número de observações seja maior que o número de variáveis, e se usar a opção 1 (agrupamentos) o número de observações ainda muito maior (na ordem de 3 vezes maior que o número de variáveis).

- Nos três casos de avaliação têm-se regressões com R^2 maiores que 75 %, o que garante a consistência da regressão para estimativa dos preços dos imóveis. Isto se deve também à boa coleta de dados representativos.

- De acordo com os apêndices 5, 6 e 7 a melhor opção para apartamentos, casas e terrenos e a opção 1, que usa as técnicas multivariadas de agrupamentos homogêneos e para cada grupo usa a Análise de Componentes Principais para reduzir as variáveis e evitar variáveis altamente correlacionadas, desta forma evita-se a multicolinearidade.

- De acordo com os apêndices 5, 6 e 7, a opção menos recomendada é a opção 3, pois os dados não sofrem tratamento estatístico, dessa forma têm-se problemas de multicolinearidade.

- De acordo com o apêndice 6 para o grupo 1 (12 observações) os parâmetros das estatísticas foram $R^2 = 0.99675$ $F = 30.7138$ $p = 0.13959$, o valor de p=13,96%, de maneira que se aceita a hipótese nula, a de não existir regressão, da mesma forma para o grupo 3 com p = 36,69%, de acordo a estes resultados pode-se afirmar que para o uso da opção 1, o número de observações deve ser maior que número de variáveis de cada grupo, dessa forma, se usar os agrupamentos, o número de observações deve ser maior pelo menos o número de vezes o dos grupos formados.

- A metodologia multivariada aplicada neste trabalho se mostrou viável, atingiram-se resultados com alto nível de precisão e a metodologia pode ser aplicada de modo geral em imóveis de outras cidades, ainda melhorando o questionário ao nível de quantificação das respostas.

Dessa forma, o presente trabalho procurou oferecer uma contribuição na área de Engenharia de Avaliação.

5.2 SUGESTÕES PARA FUTURAS PESQUISAS

1) Construir um programa em linguagem visual para uso de pessoas leigas e com poucas instruções para o cálculo (donos de imobiliárias) sem precisar do programa *Matlab*.

2) Comparar com modelos de redes neurais e algoritmos genéticos.

REFERÊNCIAS BIBLIOGRÁFICAS

ABNT (Associação Brasileira de Normas Técnicas). **Avaliação de imóveis urbanos** (NBR 5676 e NBR 502). Rio de Janeiro: ABNT, 2004.

ABUHNAMAN, S. A. **Curso básico de engenharia legal e de avaliações**. São Paulo: Pini, 1998.

BARBOSA FILHO, D. S. **Técnicas avançadas de engenharia de avaliações**. Caixa Econômica Federal, 1998.

BOUROCHE, J. M. SAPORTA, G. **Análise de dados**. Rio de Janeiro: Zahar, 1982.

BRAÚLIO, S. N. **Proposta de uma metodologia para a avaliação de imóveis urbanos baseado em métodos estatísticos multivariados**. Dissertação de Mestrado. UFPR: Curitiba, 2005.

CUADRAS, C. M. **Métodos de análisis multivariante**. Universidade de Barcelona, 1981.

CRIVISQUI, E. M. **Análisis factorial de correspondencias, un instrumento de investigación en ciencias sociales.** Laboratoire de Méthodologie du Traitement des Données, Université Libre de Bruxelles. Edición: Universidad Católica de Asunción, Asunción, Paraguay, 1993.

DANIEL, C.; WOOD, T. E. **Fitting equations to data**. New York: John Wiley & Sons, Inc, 1971.

DANTAS, R. A. **Engenharia de avaliações**: introdução à metodologia científica. São Paulo: Pini, 1998.

DANTAS, R.A. **Engenharia de avaliações uma introdução à metodologia científica**. [S.l.] São Paulo: Pini, 2000.

DRAPER, N. R. & SMITH, H. **Applied regression analysis**. New York: Jhon Wiley & Sons, Inc, 1981.

ELIAN, S. N. **Análise de regressão**. São Paulo: IME,1998.

FIKER, J. **Avaliação de imóveis urbanos.** 5. ed. São Paulo: Pini, 1997.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. **Análise conceitual das dificuldades na determinação de modelos de formação de preços através da análise de regressão**. Engenharia Civil – UM, 8: 65-75, 2000.

GONZÁLEZ, M. A. S.. **A engenharia de avaliações na visão inferencial**. São Leopoldo: Unisinos,1998.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4. ed. Nova Jersey: Prentice Hall, Inc., 1998.

JOHNSTON, J. **Métodos econométricos**. São Paulo: Atlas, 1986.

JUDGE, G. G.; GRIFFITHS, E. W.; HILL, R. C.; LEE, T. **The theory and practice of econometrics**. New York: John Wiley & Sons,1980.

KMENTA, J. **Elementos de econometria**. São Paulo: Atlas, 1978.

LACHENBRUCH, P. A. **Discriminant analysis**. New York: Hafner Press, 1975.

LEBART, L.; MORINEAU, A.; FÉNELON, J. **Tratamiento estadístico de datos**. Barcelona: Marcombo Boixareu, 1995.

MOLINA, M. A. **El catastro en España**. Valência: UPV, 1999.

MONTENEGRO DUARTE, A. **Modelo geral de valores isento de subjetividade**: caso de apartamentos na cidade de Belém. Dissertação de Mestrado. Porto Alegre e Valência, 1999.

MONTGOMERY, D. C. **Design and analysis of experiments**. 4. ed. USA: John Wiley, 1997.

MOREIRA, A. L. **Princípios de engenharia de avaliações**. São Paulo: Pini, 1997.

MOREIRA FILHO, I. I.; FRAINER, J. I.; MOREIRA, R. M. I. **Avaliação de bens por estatística inferencial e regressões múltiplas**. Porto Alegre: Avalien, 1993.

MOSCOVITCH, S. K. **Qualidade de vida urbana e valores de imóveis**: um estudo de caso para Belo Horizonte. Nova Economia, número especial: 247-279, 1997.

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

PEREIRA, R. S. **Estatística e suas aplicações**. São Paulo: Grafosul, 1970.

PLA, L. E. **Análisis multivariado**: método de componentes principales. Secretaria general de la organización de estados americanos. Washington, 1986.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**. 6. ed., Iowa: Ames, 1972.

WORZALA, E.; LENK M.; SILVA A. **An exploration of neural networks and iténs application to real estate valuation**. The Journal of Real Estate Research, 10 (2): 185-201, 1995.

APÊNDICES

APÊNDICE 1: PROGRAMA AMI (PARA MATLAB 7.0)

```
function AMI

%    ************************************************************
%    *    ANÁLISE DE AGRUPAMENTO, COMPONENTES PRINCIPAIS E REGRESSÃO
%    *              DE DADOS MULTIVARIADOS                  *
%    ************************************************************
%    *     Função programada para a dissertação de mestrado       *
%    *                  Valdir Alves                       *
%    ************************************************************

%    lê a matriz XY sendo X matriz de dados e Y vetor coluna de preços
%    a primeira coluna e a ordem das observações e a última os preços,
%    XY e uma planilha de Excel NOME. xls na pasta Work do Matlab7.
%    o programa clusteriza em K grupos, calcula os componentes principais
%    por grupo e faz a regressão linear multivariada para cada grupo mostrando
%    a equação de regressão. Para a Análise de Agrupamento podem ser
%    usadas as distâncias:

%    1)'euclid'; 2)'seuclid'; 3)'cityblock'; 4)'mahal'; 5)'minkowski'
%    e os tipos de ligação:
%    1)single, 2)complete, 3)average, 4)ward

%    Para saber o preço de um novo imóvel coloque os dados do imóvel
%    na última linha da planilha Excel com o preço = 0.
%    ao final do programa será mostrado o preço calculado para o novo imóvel.

%    para a regressão é importante saber que o número de observações
%    tem que ser maior ao número de componentes principais

%    Ao final serão gerados os arquivos com as respostas:
%     nome_G1, nome_G2.....se for agrupamentos e ACP.
%     Nome_F se for por Análise Fatorial
%     Nome_S se for por  Regressão Múltipla Simples


clear
clc
disp('      ------------------------------------------------------------------')
disp('       *  ACESSANDO OS DADOS: Selecione o arquivo de dados no Excel     *')
disp('      ------------------------------------------------------------------')
disp(' ')
[ARQUIVO,CAMINHO] = uigetfile('*.xls','Escolha o Arquivo com os dados (Excel).');
nomef=[CAMINHO ARQUIVO];
nf1=[ARQUIVO];
ncar=size(nf1);
ncar2=ncar(2);
ncar1=ncar2-3;
nf1(ncar1:ncar2)=[];
nomeA=nf1;% nome sem a extensão
disp(' ')
tt=['       arquivo XLS = ' nomef ];
disp(tt)
disp(' ')
%XY=dados;v6
XY = xlsread(nomef);%v7
[n,m]=size(XY);
ncol=m;
```

```
ncol=ncol-1;%tirando a ordem das observações
tt=['       total de observações = ' num2str(n)];
disp(tt)
tt=['       total de variáveis = ' num2str(ncol)];
disp(tt)
disp(' ')
X_Y=XY;
disp('       ---------------------------------------------------------------')
disp('        *        OPÇÕES DE ANÁLISE PARA A REGRESSÃO                  *')
disp('       ***************************************************************')
disp('       ---------------------------------------------------------------')
tt=['        * (1) Análise por Agrupamentos e Componentes Principais *'];
disp(tt)
disp('       ---------------------------------------------------------------')
tt=['        * (2) Análise Fatorial                          *'];
disp(tt)
disp('       ---------------------------------------------------------------')
tt=['        * (3) Análise de Regressão  Simples             *'];
disp(tt)
disp('       ---------------------------------------------------------------')
disp(' ')
OP=input('       ENTRAR COM O TIPO DE OPÇÃO  = ');
disp(' ')
if OP==2
   fatorim(XY, nomeA)
   return
elseif OP==3
   Rsimples(XY, nomeA)
   return
elseif OP>3
   return
end

tt=['        * Análise por Agrupamentos e Componentes Principais *'];
disp(tt)
disp('       ***************************************************************')
%X_Y matriz X sem o preço e a ordem
X_Y(:,m)=[];
X_Y(:,1)=[];
disp(' ')
d5={'euclidean';'seuclidean';'cityblock';'mahalanobis';'minkowski'};
t4={'single';'complete';'average';'ward'};
%v6
%d5={'euclid';'seuclid';'cityblock';'mahal';'minkowski'};
%t5={'single';'complete';'average';'centroid';'ward'};
%format long
% EXTRAÇÃO DOS DADOS
disp('       ---------------------------------------------------------------')
disp('        *  PROCESSO 1 : AGRUPANDO - CLUSTERING                  *')
disp('       ***************************************************************')
disp('       ---------------------------------------------------------------')
disp('        *           OPÇÃO DE TIPOS DE DISTÂNCIA                 *')
disp('       ---------------------------------------------------------------')
tt=['        *   d = 1, EUCLID   (distância euclidiana)          *'];
disp(tt)
tt=['        *   d = 2, SEUCLID (quadrado da distância euclidiana)  *' ];
disp(tt)
tt=['        *   d = 3, CITYBLOCK  (distância Manhattan)          *' ];
disp(tt)
tt=['        *   d = 4, MAHAL (distância estatística)             *'];
```

```
disp(tt)
tt=['        *   d = 5, MINKOSWSKI  (usando n=3)                  *'];
disp(tt)
disp('       ----------------------------------------------------------------------')
disp(' ')
td=input('       ENTRAR COM O TIPO DE DISTÂNCIA d = ');
disp('')
%matriz de distância
aa=d5{td};
if td==5
   Y=pdist(X_Y,aa,3);
else
   Y=pdist(X_Y,aa);
end
%verificar o melhor agrupamento
cmax=0;
imax=0;
c=[];
for i=1:4
   aa=t4{i};
   Z=linkage(Y,aa);
   c(i)=cophenet(Z,Y);
   if cmax<c(i) ;
      cmax=c(i);
      imax=i;
   end
end
aa=t4{imax};
Z=linkage(Y,aa);
tt=['       melhor agrupamento = "' num2str(aa)  '" sendo o cophenet = ' num2str(cmax)];
disp(tt)
disp(' ')
disp('       ----------------------------------------------------------------------')
disp('        *            OPÇÃO DE TIPOS DE LIGAÇÃO                        *')
disp('       ----------------------------------------------------------------------')
tt=['        *   t = 1, SINGLE   (Vizinho mais próximo)   *  Cophenet = ' num2str(c(1))];
disp(tt)
tt=['        *   t = 2, COMPLETE (Vizinho mais distante)  *  Cophenet = ' num2str(c(2))];
disp(tt)
tt=['        *   t = 3, AVERAGE  (Média das distâncias)   *  Cophenet = ' num2str(c(3))];
disp(tt)
tt=['        *   t = 4, WARD     (Método de Ward)         *  Cophenet = ' num2str(c(4))];
disp(tt)
disp('       ----------------------------------------------------------------------')
disp(' ')
tip=input('       ENTRAR COM O TIPO DE LIGAÇÃO t = ');
disp('')
aa=t4{tip};
Z=linkage(Y,aa);%fazendo as ligações
aa=dendrogram(Z,n);%VISUALIZANDO OS GRUPOS DENDROGRAMA
% K grupos
K=input('       Digite o número de grupos K = ');
T=cluster(Z,K);
disp(' ')
% VETORES DE GRUPOS
VG=[];%vetor de  ordem K indicando o número de elementos por grupo
disp('       ----------------------------------------------------------------------')
disp('        *  FIM DO PROCESSO 1:   GRUPOS FORMADOS                    *')
disp('       ----------------------------------------------------------------------')
if K>1;
```

```
   for i=1:K;
      tt=['       GRUPO ' num2str(i) ' = ' ];
      ne=0;
      for j=1:n;
         if T(j) == i;
            tt=[tt  num2str(j) ' '];
            ne=ne+1;
         end
      end
      VG=[VG ne];
      tt=[tt  '(' num2str(ne) ')'];
      disp(tt)
   end
elseif K==1;
   tt=['       GRUPO 1 = 1 .....' num2str(n)  ];
   disp(tt)
   T=diag(eye(n));
   VG(1)=n;
end
disp('       ------------------------------------------------------------------------')
zz=input('       Tecle ENTER para continuar.....','s');
disp(' ')
disp(' ')
%guardando os grupos em texto
   %gera X1, X2...grupos
for i=1:K;
   CC=[];%CC matriz Curinga
   for j=1:n;
      if T(j) == i;
        CC=[CC;XY(j,:)];
      end
   end
   [n1,m1]=size(CC);
   nomearq=[nomeA '_G' num2str(i) '.txt' ];
   arq=fopen(nomearq,'w');
   tt=['ANÁLISE DO GRUPO ' num2str(i)];
   fprintf(arq,tt);
   tt=['--- observações ' ];
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
   for i=1:n1
      for j=1:m1
         fprintf(arq,' %d    ',CC(i,j));
      end
      fprintf(arq,'\r\n');
   end
   fclose(arq);
end
%avalia todos se na última linha de XY tem preço se não avalia somente o grupo da
última linha
if XY(n,m)==0;% pega somente o grupo da n obs
   KG=T(n);%grupo da n
   CC=[];%CC matriz Curinga do grupo K
   for j=1:n;%formando a matriz do grupo KG
      if T(j) == KG;
        CC=[CC;XY(j,:)];
      end
   end
   %eliminando colunas com valores iguais
   [n1,m1]=size(CC);
```

```
CE=[];%vetor de colunas eliminadas
NCC=[];
for j=1:m1
   V=CC(:,j);
   if min(V)==max(V);
      CE=[CE j];
   else
      NCC=[NCC CC(:,j)];
   end
end
[n1,m1]=size(NCC);
YY=NCC(:,m1);%vetor de preços
NCC(:,m1)=[];%tirando a última coluna
a=isempty(CE);
if a==1;%caso null
   tt=['       Não existem variáveis  com valores iguais'];
else
   tt=['       As variáveis Nº. ' num2str(CE-1) ' foram eliminadas por possuir valores iguais' ];
end
disp(tt)
disp(' ')
CC=NCC;
nomearq=[nomeA '_G' num2str(KG) '.txt' ];
arq=fopen(nomearq,'a');
fprintf(arq,'\r\n');
fprintf(arq,tt);
fprintf(arq,'\r\n');
%fim da eliminaçao
disp('       **************************************************************************')
disp('       NA SEQÜÊNCIA SERÁ CALCULADA O VALOR (Y) DA ÚLTIMA OBSERVAÇÃO')
disp('       **************************************************************************')
disp(' ')
tt=['       A última observação ' num2str(n) ' pertence ao grupo = ' num2str(KG)];
disp(tt)
disp(' ')
disp('       **************************************************************************')
tt=['       ANÁLISE DO GRUPO = ' num2str(KG)];
disp(tt)
disp('       **************************************************************************')
disp(' ')
disp('       ---------------------------------------------------------------------')
disp('        *  PROCESSO 2:   COMPONENTES PRINCIPAIS                        *')
disp('       ---------------------------------------------------------------------')
%covariância
S=corrcoef(CC);
%S=cov(CC);
r1=eig(S);
r1=flipud(sort(r1));
m1=length(r1);
j1=(1:m1)';
t1=sum(r1);
r2=(r1/t1)*100;
r3=(cumsum(r1)/t1)*100;
r=[j1 r1 r2 r3];
disp('')
disp('       * PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *')
disp('       *    AUTOVALORES DA MATRIZ CORRELAÇÃO    *')
disp(' ')
disp('       ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL. ')
disp('              LORES     (EM %)    ACUM. (%)  ')
```

```
disp('        ----------------------------------------')
disp(sprintf('   %8.0f %10.4f %8.2f %11.2f\n',r'))
disp('        ----------------------------------------')
tt=['       Digite o número de componentes  < ' num2str(VG(1)) ' = '];
nc=input(tt);
%salvando as informações
if nc>m1;
   nc=m1;
end
%E autovetores, e D Matriz Diagonal de autovalores
[E,D]=eig(S);
[dd,ind]=sort(diag(D)); %dd=vetor autovalor, ind=vetor de índices
dd=flipud(dd)';
ind=flipud(ind)';
%ordena de acordo a os índices e ind flipado
E=E(:,ind);
n2=length(dd);
%componentes principais
 [n1,m1]=size(E);
 %nomearq=[nomeA '_G' num2str(KG) '.txt' ];
 %arq=fopen(nomearq,'a');
 fprintf(arq,'\r\n');
 tt=['autovalores  ' ];
 fprintf(arq,tt);
 fprintf(arq,'\r\n');
 for j=1:m1
    fprintf(arq,' %d    ',dd(j));
 end
 fprintf(arq,'\r\n');
 fprintf(arq,' ');
 fprintf(arq,'\r\n');
 tt=['Componentes principais - autovetores  ' ];
 fprintf(arq,tt);
 fprintf(arq,'\r\n');
 for ii=1:n1
    for j=1:m1
       fprintf(arq,' %d    ',E(ii,j));
    end
    fprintf(arq,'\r\n');
 end
 %fclose(arq);
%escores
ESCR=CC*E;
%controlando o número de comp principais
CP=[];%MATRIZ DE COMP PRINC tirando as outras colunas
for ii=1:nc;
   CP=[CP ESCR(:,ii)];
end
[n1,m1]=size(CP);
 %nomearq=[nomeA '_G' num2str(KG) '.txt' ];
 %arq=fopen(nomearq,'a');
 fprintf(arq,'\r\n');
 tt=['Escores  com ' num2str(nc) ' componentes'];
 fprintf(arq,tt);
 fprintf(arq,'\r\n');
 for ii=1:n1
    for j=1:m1
       fprintf(arq,' %d    ',CP(ii,j));
    end
    fprintf(arq,'\r\n');
```

```
 end
 %fclose(arq);
disp(' ')
disp('       ------------------------------------------------------------------')
disp('        *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                              *')
disp('       ------------------------------------------------------------------')
disp(' ')
[n3,m3]=size(CP);
CPY=CP(n3,:);%comp prin da última linha
CP(n3,:)=[];%Tira a última linha sem preço
YY(n3,:)=[];% Tira a última linha no YY
n3=n3-1;
%acrescer uma coluna de um
uno=diag(eye(n3));
CP=[CP uno];
[B,BINT,R,RINT,STATS] = regress(YY,CP,95);
%B=coeficientes y=Bx sendo a última a constante
    ttY=['       Y = ' num2str(B(1)) '*CP1 ' ];
    for ii=2:nc;
       if B(ii)>0;
          ttY=[ttY ' + ' num2str(B(ii)) '*CP' num2str(ii) ];
       else
          ttY=[ttY ' ' num2str(B(ii)) '*CP' num2str(ii) ];
       end
    end
    ii=nc+1;
    if B(ii)>0;
      ttY=[ttY ' + ' num2str(B(ii))  ];
    else
      ttY=[ttY ' ' num2str(B(ii)) ];
    end
    tt1=['       REGRESSÃO PARA O GRUPO ' num2str(KG) ];
    disp(tt1)
    disp('       ------------------------------------------------------------------')
    disp(ttY)
    disp('       ------------------------------------------------------------------')
    disp(' ')
    tt1=['       ESTATÍSTICA DO MODELO ' num2str(KG) ];
    disp(tt1)
    disp('       ------------------------------------------------------------------')
    disp('       R2         F         p   ')
    disp('       ----------------------------------------')
    STATS1=STATS;
    STATS1(4)=[];
    disp(sprintf('      %8.4f %10.4f  %8.5f\n',STATS1'))
    disp('       ----------------------------------------')
    disp(' ')
    tt1=['       Componentes principais da última observação  (cp1,cp2,...)'  ];
    disp(tt1)
    tt1=['       '];
    for ii=1:nc;
       tt1=[tt1 num2str(CPY(ii)) ' '];
    end
    disp(tt1)
    disp(' ')
    nc1=nc+1;
    B0=B(nc1);
    B(nc1)=[];
    Y0=CPY*B+B0;
    tt2=['       Y(variável dependente) da última observação  = ' num2str(Y0) ];
```

```
    disp(tt2)
    disp('       **************************************************************************')

    disp('       FIM DO PROGRAMA');
    %nomearq=[nomeA '_G' num2str(KG) '.txt' ];
    %arq=fopen(nomearq,'a');
    fprintf(arq,'\r\n');
    tt1=['REGRESSÃO MÚLTIPLA ' ];
    fprintf(arq,tt1);
    fprintf(arq,'\r\n');
    fprintf(arq,ttY);
    fprintf(arq,'\r\n');
    fprintf(arq,' ');
    fprintf(arq,'\r\n');
    fprintf(arq,tt2);
    fprintf(arq,'\r\n');
    tt1=['ESTATÍSTICA DO MODELO '];
    fprintf(arq,tt1);
    fprintf(arq,'\r\n');
    tt1=['R2 = ' num2str(STATS(1)) '    F = ' num2str(STATS(2)) '    p = ' num2str(STATS(3))];
    fprintf(arq,tt1);
    fprintf(arq,'\r\n');
    fclose(arq);
    return
else%pega todos os grupos para avaliação
  disp(' ')
  disp('       **************************************************************************')
  tt=['       NA SEQÜÊNCIA SERÃO ANALIZADOS OS ' num2str(K) ' GRUPOS'];
  disp(tt)
  disp('       **************************************************************************')
  disp(' ')
  disp('       ---------------------------------------------------------------------')
  disp('        *  PROCESSO 2:   COMPONENTES PRINCIPAIS                        *')
  disp('       ---------------------------------------------------------------------')
  disp(' ')
  for i=1:K
    tt=['       ANALISANDO O GRUPO ' num2str(i) ];
    disp(tt)
    disp('       **************************************************************************')
    CC1=[];%CC1 matriz Curinga DO GRUPO
    for j=1:n;
      if T(j) == i;
        CC1=[CC1;XY(j,:)];
      end
    end
    %eliminando colunas com valores iguais
    [n1,m1]=size(CC1);
    CE=[];%vetor de colunas eliminadas
    NCC=[];
    for j=1:m1
      V=CC1(:,j);
      if min(V)==max(V);
        CE=[CE j];
      else
        NCC=[NCC CC1(:,j)];
      end
    end
    [n1,m1]=size(NCC);
    a=isempty(CE);
    if a==1;%caso null
```

```
      tt=['       Não existem variáveis  com valores iguais'];
   else
      [n7,m7]=size(CE);
      if m7>1
         tt=['       As variáveis Nº [' num2str(CE-1) '] foram eliminadas por possuir valores iguais' ];
      else
         tt=['       A variável Nº [' num2str(CE-1) '] foi eliminada por possuir valor igual ];
      end
   end
   disp(tt)
   disp(' ')
   nomearq=[nomeA '_G' num2str(i) '.txt' ];
   arq=fopen(nomearq,'a');
   fprintf(arq,'\r\n');
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
   CC1=NCC;
%fim da eliminação
   [n1,m1]=size(NCC);
   YY=CC1(:,m1);%vetor de preços
   CC1(:,m1)=[];%tirando a última coluna
   %covariância
   %S=cov(CC1);
   S=corrcoef(CC1);
   r1=eig(S);
   r1=flipud(sort(r1));
   m1=length(r1);
   j1=(1:m1)';
   t1=sum(r1);
   r2=(r1/t1)*100;
   r3=(cumsum(r1)/t1)*100;
   r=[j1 r1 r2 r3];
   disp(' ')
   disp('       * PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *')
   disp('       *    AUTOVALORES DA MATRIZ CORRELAÇÃO    *')
   disp(' ')
   disp('       ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL. ')
   disp('              LORES     (EM %)    ACUM. (%)  ')
   disp('       ----------------------------------------')
   disp(sprintf('   %8.0f %10.4f %8.2f %11.2f\n',r'))
   disp('       ----------------------------------------')
   tt=['       Digite o número de componentes  < ' num2str(VG(i)) ' = '];
   nc=input(tt);
   if nc>m1;
      nc=m1;
   end
   %E autovetores, e D Matriz Diagonal de autovalores
   [E,D]=eig(S);
   [dd,ind]=sort(diag(D)); %dd=vetor autovalor, ind=vetor de índices
   dd=flipud(dd)';
   ind=flipud(ind)';
   %ordena de acordo a os índices e ind flipado
   E=E(:,ind);
   n2=length(dd);
   [n1,m1]=size(E);
%nomearq=[nomeA '_G' num2str(i) '.txt' ];
%arq=fopen(nomearq,'a');
fprintf(arq,' ');
fprintf(arq,'\r\n');
tt=['autovalores  ' ];
```

```
fprintf(arq,tt);
fprintf(arq,'\r\n');
for j=1:m1
   fprintf(arq,' %d    ',dd(j));
end
fprintf(arq,'\r\n');
fprintf(arq,' ');
fprintf(arq,'\r\n');
tt=['Componentes principais - autovetores  ' ];
fprintf(arq,tt);
fprintf(arq,'\r\n');
for ii=1:n1
   for j=1:m1
      fprintf(arq,' %d    ',E(ii,j));
   end
   fprintf(arq,'\r\n');
end
%fclose(arq);
   %escores
   ESCR=CC1*E;
   %controlando o número de comp principais
   CP=[];%MATRIZ DE COMP PRINC
   for ii=1:nc;
      CP=[CP ESCR(:,ii)];
   end
   [n2,m2]=size(CP);
   %nomearq=[nomeA '_G' num2str(i) '.txt' ];
   %arq=fopen(nomearq,'a');
   %fprintf(arq,' ');
   fprintf(arq,'\r\n');
   tt=['Escores  com ' num2str(nc) ' componentes' ];
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
   for ii=1:n2
      for j=1:m2
         fprintf(arq,' %d    ',CP(ii,j));
      end
      fprintf(arq,'\r\n');
   end
   %fclose(arq);
   disp(' ')
   disp('       ------------------------------------------------------------------------')
   disp('        *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                                  *')
   disp('       ------------------------------------------------------------------------')
   disp(' ')
   [n3,m3]=size(CP);
   %acrescer uma coluna de um
   uno=diag(eye(n3));
   CP=[CP uno];
   [B,BINT,R,RINT,STATS] = regress(YY,CP,95);
   %B=coeficientes y=Bx sendo a última a constante
   ttY=['       Y = ' num2str(B(1)) '*CP1 ' ];
   for ii=2:nc;
      if B(ii)>0;
         ttY=[ttY ' + ' num2str(B(ii)) '*CP' num2str(ii) ];
      else
         ttY=[ttY ' ' num2str(B(ii)) '*CP' num2str(ii) ];
      end
   end
   ii=nc+1;
```

```
if B(ii)>0;
  ttY=[ttY ' + ' num2str(B(ii))  ];
else
  ttY=[ttY ' ' num2str(B(ii)) ];
end
tt1=['       REGRESSÃO PARA O GRUPO ' num2str(i) ];
disp(tt1)
disp('       ---------------------------------------------------------------------')
disp(ttY)
disp('       ---------------------------------------------------------------------')
disp(' ')
tt1=['       ESTATÍSTICA DO MODELO ' num2str(i) ];
disp(tt1)
disp('       ---------------------------------------------------------------------')
disp('       R2        F        p   ')
disp('       ----------------------------------------')
STATS1=STATS;
STATS1(4)=[];
disp(sprintf('     %8.4f %10.4f  %8.5f\n',STATS1'))
disp('       ----------------------------------------')
disp('       ************************************************************************')
disp(' ')
%nomearq=[nomeA '_G' num2str(i) '.txt' ];
%arq=fopen(nomearq,'a');
%fprintf(arq,' ');
fprintf(arq,'\r\n');
tt1=['REGRESSÃO MÚLTIPLA ' ];
fprintf(arq,tt1);
fprintf(arq,'\r\n');
fprintf(arq,ttY);
fprintf(arq,'\r\n');
fprintf(arq,' ');
fprintf(arq,'\r\n');
tt1=['ESTATÍSTICA DO MODELO '];
fprintf(arq,tt1);
fprintf(arq,'\r\n');
tt1=['R2 = ' num2str(STATS(1)) '    F = ' num2str(STATS(2)) '    p = ' num2str(STATS(3))];
fprintf(arq,tt1);
fprintf(arq,'\r\n');
fprintf(arq,'\r\n');
%avaliando o grupo
Naval=CP*B;
tt1=['Avaliando o grupo '];
fprintf(arq,tt1);
fprintf(arq,'\r\n');
for i=1:n2
   fprintf(arq,' %d    ',CC1(i,1));
   fprintf(arq,' %d    ',YY(i));
   %tt1=[num2str(Naval(i)) ];
   fprintf(arq,' %d    ',Naval(i));
   fprintf(arq,'\r\n');
end
fclose(arq);

if i==K
   disp('       FIM DO PROGRAMA');
else
   zz=input('       Tecle ENTER para continuar.....','s');
   disp(' ')
end
```

```
   end
end
```

SUB-PROGRAMA ANÁLISE FATORIAL

```
function fatorim(XY, nomeA)
%nome A = nome do arquivo sem a extensão
%XY = arquivo do Excel
[n1,m1]=size(XY);
   nomearq=[nomeA '_F.txt' ];
   arq=fopen(nomearq,'w');
   tt=['ANÁLISE FATORIAL ' ];
   fprintf(arq,tt);
   tt=['--- Observações  ' ];
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
   for i=1:n1
      for j=1:m1
         fprintf(arq,' %d    ',XY(i,j));
      end
      fprintf(arq,'\r\n');
   end
X_Y=XY;
[n,m]=size(XY);
YY=XY(:,m);
%X_Y matriz X sem o preço e a ordem
X_Y(:,m)=[];
X_Y(:,1)=[];
disp(' ')
X=X_Y;
k=1;%matriz de dados
 R=corrcoef(X);

%Autovalores e autovetores de R
   [E2,D2]=eig(R);
   [dd2,i2]=sort(diag(D2));
   dd2=flipud(dd2)';
   i2=flipud(i2)';
   E2=E2(:,i2);
   ddt=(dd2/(sum(dd2)))*100;
   ddacum=cumsum(ddt);
    disp('       ------------------------------------------------------------------------')
    disp('        *  PROCESSO 1:   Autovalores e variância explicada acumulada      *')
    disp('       ------------------------------------------------------------------------')
    disp(' ')

%Matriz de pesos L
r1=eig(R);
   r1=flipud(sort(r1));
   m1=length(r1);
   j1=(1:m1)';
   t1=sum(r1);
   r2=(r1/t1)*100;
   r3=(cumsum(r1)/t1)*100;
r=[j1 r1 r2 r3];
   disp('       * PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *')
   disp('       *    AUTOVALORES DA MATRIZ CORRELAÇÃO    *')
   disp(' ')
```

```
  disp('       ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL. ')
  disp('              LORES     (EM %)    ACUM. (%)  ')
  disp('       ----------------------------------------')
  disp(sprintf('   %8.0f %10.4f %8.2f %11.2f\n',r'))
  disp('       ----------------------------------------')

  disp(' ')
  tt=['       Digite o número de fatores  < ' num2str(m1) ' = '];
  N=input(tt);
  nc=N;
disp(' ')
tt=['número de fatores ' num2str(N) ];
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
%disp(f1)%ESCORES FATORIAIS
[Lambda,Psi,T,stats,F] = factoran(X,N,'scores','regression');
%[Lambda,Psi,T,stats,F] = factoran(X,N,'rotate','varimax','scores','regression');
disp('       ---------------------------------------------------------------')
disp('        *  PROCESSO 2:   ESCORES FATORIAIS ROTACIONADOS - VARIMAX                *')
disp('       ---------------------------------------------------------------')
disp(' ')
disp(F)%ESCORES FATORIAIS
CP=F;
disp(' ')
disp('       ---------------------------------------------------------------')
disp('        *  PROCESSO 3:   REGRESSÃO MÚLTIPLA                              *')
disp('       ---------------------------------------------------------------')
disp(' ')
%para calcular Y
[n3,m3]=size(CP);
if XY(n,m)==0
  CPY=CP(n3,:);%comp prin da última linha
  CP(n3,:)=[];%Tira a última linha sem preço
  YY(n3,:)=[];% Tira a última linha no YY
  n3=n3-1;
end

      %acrescer uma coluna de um
      uno=diag(eye(n3));
      CP=[CP uno];
      [B,BINT,R,RINT,STATS] = regress(YY,CP,100);
      %B=coeficientes y=Bx sendo a última constante
      ttY=['       Y = ' num2str(B(1)) '*F1 ' ];
      for ii=2:nc;
         if B(ii)>0;
            ttY=[ttY ' + ' num2str(B(ii)) '*F' num2str(ii) ];
         else
            ttY=[ttY ' ' num2str(B(ii)) '*F' num2str(ii) ];
         end
      end
      ii=nc+1;
      if B(ii)>0;
        ttY=[ttY ' + ' num2str(B(ii))  ];
      else
        ttY=[ttY ' ' num2str(B(ii)) ];
      end
      tt1=['       REGRESSÃO' ];
      disp(tt1)
      disp('       ---------------------------------------------------------------')
      disp(ttY)
```

```
      disp('      ----------------------------------------------------------------------')
      disp(' ')
      tt1=['      ESTATÍSTICA DO MODELO '  ];
      disp(tt1)
      STATS1=STATS;
      STATS1(4)=[];
      disp('      ----------------------------------------------------------------------')
      disp('      R2        F        p   ')
      disp('      ----------------------------------------')
      disp(sprintf('     %8.4f %10.4f  %8.5f\n',STATS1'))
      disp('      ----------------------------------------')
      disp('      **********************************************************************')
      disp(' ')
      if XY(n,m)==0
         nc1=nc+1;
         B0=B(nc1);
         B(nc1)=[];
         Y0=CPY*B+B0;
         tt2=['      Y(variável dependente) da última observação  = ' num2str(Y0) ];
         disp(tt2)
         disp('      **********************************************************************')
      end
      fprintf(arq,'\r\n');
      tt1=['REGRESSÃO MÚLTIPLA ' ];
      fprintf(arq,tt1);
      fprintf(arq,'\r\n');
      fprintf(arq,ttY);
      fprintf(arq,'\r\n');
      fprintf(arq,' ');
      fprintf(arq,'\r\n');
      tt1=['ESTATÍSTICA DO MODELO '];
      fprintf(arq,tt1);
      fprintf(arq,'\r\n');
      tt1=['R2 = ' num2str(STATS(1)) '    F = ' num2str(STATS(2)) '    p = ' num2str(STATS(3))];
      fprintf(arq,tt1);
      fprintf(arq,'\r\n');
      fclose(arq);

disp('   *FIM DO PROGRAMA *')
return
```

SUB-PROGRAMA REGRESSÃO SIMPLES

```
function Rsimples(XY, nomeA)
[n1,m1]=size(XY);
   nomearq=[nomeA '_S.txt' ];
   arq=fopen(nomearq,'w');
   tt=['regressão multivariada ' ];
   fprintf(arq,tt);
   tt=['---  Observações  ' ];
   fprintf(arq,tt);
   fprintf(arq,'\r\n');
   for i=1:n1
      for j=1:m1
         fprintf(arq,' %d    ',XY(i,j));
      end
      fprintf(arq,'\r\n');
   end
X_Y=XY;
[n,m]=size(XY);
YY=XY(:,m);
%X_Y matriz X sem o preço e a ordem
X_Y(:,m)=[];
X_Y(:,1)=[];
disp(' ')
CP=X_Y;
disp(' ')
disp('       ------------------------------------------------------------------------')
disp('        *  PROCESSO 1:   REGRESSÃO MÚLTIPLA                              *')
disp('       ------------------------------------------------------------------------')
disp(' ')
%para calcular Y

if XY(n,m)==0
  CPY=CP(n,:);%variáveis da última linha
  CP(n,:)=[];%Tira a última linha sem preço
  YY(n,:)=[];% Tira a última linha no YY
  n1=n-1;
end

      %acrescer uma coluna de um
      uno=diag(eye(n1));
      CP=[CP uno];
      [B,BINT,R,RINT,STATS] = regress(YY,CP,100);
      %B=coeficientes y=Bx sendo a última constante
      ttY=['       Y = ' num2str(B(1)) '*V1 ' ];
      nc=m-2;
      for ii=2:nc;
         if B(ii)>0;
            ttY=[ttY ' + ' num2str(B(ii)) '*V' num2str(ii) ];
         else
            ttY=[ttY ' ' num2str(B(ii)) '*V' num2str(ii) ];
         end
      end
      ii=nc+1;
      if B(ii)>0;
        ttY=[ttY ' + ' num2str(B(ii))  ];
      else
        ttY=[ttY ' ' num2str(B(ii)) ];
      end
      tt1=['       REGRESSÃO' ];
```

```
        disp(tt1)
        disp('       ------------------------------------------------------------------------')
        disp(ttY)
        disp('       ------------------------------------------------------------------------')
        disp(' ')
        tt1=['       ESTATÍSTICA DO MODELO '  ];
        disp(tt1)
        STATS1=STATS;
        STATS1(4)=[];
        disp('       ------------------------------------------------------------------------')
        disp('       R2         F         p   ')
        disp('       ----------------------------------------')
        disp(sprintf('      %8.4f %10.4f  %8.5f\n',STATS1'))
        disp('       ----------------------------------------')
        disp('       ************************************************************************')
        disp(' ')
        if XY(n,m)==0
           nc1=nc+1;
           B0=B(nc1);
           B(nc1)=[];
           Y0=CPY*B+B0;
           tt2=['       Y(variável dependente) da última observação  = ' num2str(Y0) ];
           disp(tt2)
           disp('       ************************************************************************')
        end
        fprintf(arq,'\r\n');
        tt1=['REGRESSÃO MÚLTIPLA ' ];
        fprintf(arq,tt1);
        fprintf(arq,'\r\n');
        fprintf(arq,ttY);
        fprintf(arq,'\r\n');
        fprintf(arq,' ');
        fprintf(arq,'\r\n');
        tt1=['ESTATÍSTICA DO MODELO '];
        fprintf(arq,tt1);
        fprintf(arq,'\r\n');
        tt1=['R2 = ' num2str(STATS(1)) '    F = ' num2str(STATS(2)) '    p = ' num2str(STATS(3))];
        fprintf(arq,tt1);
        fprintf(arq,'\r\n');
        fclose(arq);

disp('   *FIM DO PROGRAMA *')
return
```

APÊNDICE II - MATRIZ DE DADOS REFERENTE AOS APARTAMENTOS

apartamento	posição do apto	elevador	garagem	localização	área privativa	pavimento	andar	pecas	sala	dormitório	suíte	banheiro	dep. de emp.
1	3	2	1	1	222	15	3	9	1	2	1	1	1
2	3	1	1	1	162,44	7	1	9	1	2	1	1	1
3	3	1	1	1	176	8	2	9	1	2	1	1	1
4	3	2	2	1	179,2	14	3	10	2	2	1	1	1
5	3	2	2	1	279,4	20	3	8	1	2	1	1	1
6	3	1	1	1	120	12	3	12	2	3	1	2	1
7	2	1	1	1	220	16	2	12	3	3	1	2	1
8	3	1	1	1	107	8	1	9	3	2	1	1	0
9	2	0	1	1	50	4	2	4	1	1	0	1	0
10	3	2	3	1	240	15	3	12	3	2	1	1	1
11	3	0	1	0	100	3	2	7	2	3	0	2	0
12	3	0	1	1	147	6	1	7	1	2	1	1	0
13	3	1	1	1	108	7	3	11	2	2	1	1	1
14	2	2	2	1	311	16	3	10	3	2	1	1	1
15	2	1	1	1	132	7	2	8	2	2	1	1	0
16	3	2	1	1	107	8	3	8	2	2	1	1	0
17	3	2	1	1	107	8	4	8	2	2	1	1	0
18	3	2	4	1	330	15	3	17	3	3	1	3	1
19	3	2	2	1	220	15	3	13	2	3	1	1	1
20	3	1	1	1	130	6	3	11	2	3	1	1	1
21	3	2	2	1	164	10	3	11	2	3	1	1	1
22	3	2	2	1	160	15	3	11	2	3	1	2	1
23	3	2	1	1	374	15	3	14	3	2	1	3	1
24	3	1	1	1	220	16	4	14	3	3	1	2	1
25	2	1	1	1	220	16	3	14	3	3	1	2	1
26	3	1	2	1	180	13	3	13	2	3	1	1	1
27	3	1	2	1	320	13	3	11	1	3	1	1	1
28	3	1	2	1	180	13	3	13	2	3	1	1	1
29	3	1	2	1	320	13	3	11	1	3	1	1	1

30	1	2	2	1	260	14
31	2	2	2	1	260	14
32	3	2	2	1	260	14
33	3	1	1	1	140	8
34	1	1	1	1	130	7
35	1	2	2	1	260	14
36	1	2	2	1	310	16
37	3	0	1	1	100	4
38	3	0	1	1	70	3
39	3	0	1	1	90	3
40	3	1	1	1	38	8
41	3	1	1	1	40	8
42	3	1	1	1	170	7
43	1	1	1	1	170	7
44	3	1	1	1	170	7

apartamento	dist. de escola	dist. de hospital	dist. de supermercado	acabamento
1	2	2	3	3
2	3	1	3	3
3	3	2	3	3
4	3	2	3	3
5	3	2	3	3
6	3	2	3	3
7	2	2	2	3
8	2	2	2	2
9	3	2	3	2
10	3	3	3	3
11	3	1	1	2
12	3	2	3	2
13	3	3	3	2
14	1	1	1	3
15	3	2	1	2
16	3	3	3	2

3	11	2	2	1	1	1
2	11	2	2	1	1	1
1	11	2	2	1	1	1
2	6	1	2	1	1	0
3	7	2	2	1	1	0
1	11	2	2	1	1	1
2	11	2	2	1	1	1
3	7	1	3	0	2	0
2	5	1	2	0	1	0
2	6	1	2	0	2	0
2	4	1	1	0	1	0
3	5	1	2	0	1	0
2	9	2	2	1	1	0
1	9	2	2	1	1	0
3	9	2	2	1	1	0

revestimento do prédio	conservação	idade real	idade aparente	valor (R$)
4	2	2	2	130000
1	5	4	6	85000
1	5	4	6	80000
4	5	4	6	115000
4	5	2	4	150000
2	5	3	4	110000
4	4	2	2	120000
1	3	3	5	68000
1	4	4	4	40000
4	4	2	3	170000
1	3	2	3	50000
2	5	4	5	60000
4	3	3	3	93000
4	5	4	4	250000
1	4	5	5	65000
4	4	3	3	65000

17	3	3	3	2
18	2	1	2	2
19	3	3	3	2
20	3	3	3	2
21	3	3	3	2
22	3	3	3	2
23	1	1	1	2
24	3	3	3	3
25	3	3	3	3
26	3	1	3	3
27	3	3	3	3
28	3	1	3	3
29	3	3	3	3
30	3	3	3	2
31	3	3	3	2
32	3	3	3	2
33	2	2	3	2
34	3	3	3	2
35	3	3	3	2
36	1	1	1	3
37	3	3	3	2
38	3	3	3	2
39	3	3	3	2
40	3	3	3	2
41	3	3	3	2
42	3	3	3	2
43	3	3	3	2
44	3	3	3	2

4	4	3	3	71000
4	4	3	2	220000
4	4	5	5	200000
3,5	4	3	4	90000
3,5	4	3	4	140000
4	4	3	4	250000
4	4	4	4	250000
4	4	3	4	150000
4	4	3	4	120000
4	3	4	4	180000
4	4	5	5	250000
4	3	4	4	180000
4	4	5	5	250000
4	3	2	2	115000
4	3	2	2	120000
4	3	2	2	140000
4	3	5	5	90000
1	3	4	4	65000
4	3	2	2	120000
4	4	4	4	210000
1	3	1	1	100000
1	3	1	1	70000
1	2	1	1	95000
1	3	3	4	30000
1	3	3	4	40000
4	4	4	6	100000
4	4	4	6	90000
4	4	4	6	110000

APÊNDICE III - MATRIZ DE DADOS REFERENTE A CASAS RESIDENCIAIS

casa	bairro	garagem	suíte	banheiro	edícula	dist. supermercado	área construída	área do terreno	acabamento	cobertura	estrutura	conservação	piscina
1	5	1	1	2	1	3	183	500	1	4	3	1	0
2	4	1	1	3	0	2	216	977	1	4	3	1	0
3	4	1	1	3	0	2	155	420	2	4	3	1	0
4	2	1	1	2	0	1	170	490	2	4	3	1	0
5	4	1	1	2	1	1	160	480	2	4	3	3	0
6	5	1	1	2	1	3	160	500	2	4	3	3	0
7	5	1	1	2	1	3	134	1000	2	4	3	2	0
8	3	1	0	1	0	3	70	300	1	2	1	1	0
9	3	0	0	1	0	2	198	350	1	1	1	3	0
10	3	1	0	1	0	2	113	400	2	2	1	2	0
11	5	1	0	2	1	3	238	1200	2	4	3	2	0
12	3	1	0	2	0	2	158	315	2	1	1	2	0
13	3	1	0	1	0	3	124	300	2	4	3	2	0
14	5	0	0	1	0	3	95	340	2	2	1	1	0
15	4	1	1	1	1	3	187	490	3	4	3	2	0
16	4	1	1	1	1	1	242	490	3	4	3	2	0
17	3	0	0	2	0	1	100	480	2	4	3	2	0
18	3	1	1	1	0	1	180	786	3	4	3	3	0
19	5	1	1	3	1	1	380	480	2	3	3	2	0
20	4	1	1	1	1	1	240	490	3	4	3	2	1
21	3	1	1	1	1	2	184	1000	2	4	3	3	1
22	3	1	0	2	0	3	140	446	2	1	2	2	0
23	5	1	1	1	0	3	400	600	2	4	3	2	0
24	4	1	0	1	0	3	110	270	2	4	3	3	0
25	2	0	0	1	0	3	120	300	2	4	3	1	0
26	2	1	0	1	0	3	150	300	2	4	3	3	0
27	5	1	1	2	0	3	400	750	2	4	3	3	0
28	5	1	0	1	0	3	130	300	2	2	1	1	0

29	4	1	1	1	0	3
30	4	1	1	2	1	1
31	5	1	1	1	1	1
32	5	1	1	2	1	1
33	5	1	1	1	1	3
34	5	1	1	1	0	3
35	5	1	0	1	0	3
36	5	1	1	2	1	3
37	5	1	0	1	0	3
38	4	1	0	1	0	2
39	5	1	0	1	0	3
40	5	1	0	1	0	3
41	3	0	0	1	0	2
42	3	1	1	1	0	2
43	2	0	1	2	0	2
44	3	1	1	1	0	1
45	3	1	1	1	0	1
46	3	1	1	2	1	3
47	5	1	1	2	0	3
48	2	1	0	1	0	2
49	5	1	1	3	0	3
50	2	1	0	1	0	3
51	3	1	0	1	0	2

casa	dormitório	dep.de empregados	lavanderia	peças
1	2	1	1	8
2	3	1	1	13
3	3	1	1	12
4	2	0	1	7
5	2	0	1	9
6	3	0	1	8
7	4	0	1	10

200	400	2	4	3	3	0
244	500	2	4	3	2	0
113	300	2	4	3	2	0
220	500	4	4	3	2	1
92	650	2	4	3	3	0
123	300	2	4	3	2	0
150	1000	1	2	1	1	0
200	400	2	4	4	2	0
100	1000	1	1	1	1	0
70	490	1	1	1	1	0
100	950	2	2	1	1	0
80	475	2	2	1	1	0
70	600	1	3	3	3	0
180	640	2	4	3	3	0
70	480	2	4	3	3	0
115	250	2	4	3	3	0
160	450	2	4	3	3	1
325	225	3	4	4	3	0
320	450	3	4	4	3	0
116	300	2	3	3	2	0
180	500	2	4	3	3	0
100	800	2	3,5	3	2	0
80	390	2	2	3	1	0

idade aparente	valor (R$)
2	120.000
2	160.000
2	110.000
3	70.000
3	85.000
3	100.000
4	160.000

8	3	0	0	5
9	2	0	0	5
10	3	0	0	8
11	4	0	1	9
12	4	0	1	12
13	3	0	0	8
14	2	0	0	6
15	2	1	1	12
16	3	1	1	13
17	3	0	1	8
18	2	1	1	10
19	3	0	1	14
20	2	1	1	11
21	2	1	1	13
22	3	0	1	6
23	5	1	1	14
24	2	0	1	5
25	3	0	1	6
26	3	0	1	6
27	3	0	1	11
28	3	0	0	5
29	3	0	0	9
30	3	1	1	13
31	2	1	0	9
32	3	1	1	15
33	2	0	1	7
34	2	0	1	7
35	2	0	1	7
36	2	1	1	7
37	2	0	1	5
38	2	0	1	5
39	3	0	1	7
40	3	0	1	7

1	30.000
3	28.000
2	35.000
1	150.000
1	45.000
1	30.000
1	28.000
4	140.000
3	130.000
5	50.000
3	150.000
5	150.000
2	135.000
3	180.000
2	40.000
3	150.000
4	60.000
3	15.000
5	45.000
3	165.000
3	95.000
6	100.000
6	130.000
4	90.000
3	185.000
5	90.000
5	15.000
1	150.000
2	150.000
1	140.000
1	30.000
1	60.000
1	45.000

41	2	0	0	5
42	2	1	1	10
43	1	0	1	8
44	2	0	1	8
45	2	1	1	9
46	4	1	1	16
47	2	0	0	13
48	3	0	0	6
49	3	0,5	1	12
50	2	0	0	5
51	3	0	1	6

5	30.000
6	90.000
5	50.000
4	70.000
1	130.000
6	180.000
4	290.000
3	50.000
5	98.000
4	65.000
2	40.000

APÊNDICE IV - MATRIZ DE DADOS REFERENTE AOS TERRENOS

terreno	localização	setor comercial	pólo	frente	área do terreno	proteção	plano	inclinado	posição	pavimentação	valor (R$)
1	6	2	1	2	650	3	0	1	1	1	100000
2	2	0	0	3	640	3	1	0	2	1	25000
3	5	0	1	1	500	3	1	0	1	1	43000
4	3	0	1	2	390	3	2	0	1	1	33000
5	5	0	1	3	262	3	2	0	2	1	40000
6	5	2	1	3	1000	3	2	0	1	1	140000
7	4	1	1	3	950	3	2	0	1	1	90000
8	4	0	1	1	475	3	0	1	1	1	38000
9	5	1	1	3	470	3	3	0	2	1	70000
10	6	3	1	1	500	0	3	0	1	1	145000
11	2	0	0	2	420	0	3	0	2	0	15000
12	2	0	0	2	420	0	3	0	1	0	13000
13	3	2	1	3	2000	0	2	0	2	1	100000
14	4	1	1	3	950	3	2	0	1	1	90000
15	6	3	1	3	1000	3	2	0	2	1	250000
16	5	0	1	1	242	0	2	0	1	1	38000
17	6	3	1	3	940	2	3	0	2	1	300000
18	5	0	1	1	500	2	1	0	1	1	47000
19	5	0	1	3	350	0	2	0	2	1	60000
20	5	0	1	1	500	0	0	1	1	1	50000
21	3	0	0	2	336	0	3	0	1	1	15000
22	3	0	1	2	336	0	3	0	1	1	30000
23	4	0	1	2	300	0	2	0	1	1	25000
24	3	0	1	1	450	3	3	0	1	1	30000

APÊNDICE V – QUADRO DE PREÇOS PREDECIDOS DE APARTAMENTOS

apartamento	valor	Avaliação					
		A_Agrupam	Dif.(%)	A_Fatorial	Dif.(%)	A_Simples	Dif.(%)
1	130.000,00	139.620,30	-7,40%	142.334,10	-9,49%	132.327,90	-1,79%
2	85.000,00	73.230,46	13,85%	69.097,90	18,71%	80.008,87	5,87%
3	80.000,00	91.097,49	-13,87%	105.933,70	-32,42%	90.129,81	-12,66%
4	115.000,00	162.978,70	-41,72%	132.372,70	-15,11%	147.583,00	-28,33%
5	150.000,00	165.618,90	-10,41%	163.176,50	-8,78%	157.204,80	-4,80%
6	110.000,00	108.145,20	1,69%	121.467,10	-10,42%	89.394,28	18,73%
7	120.000,00	103.441,50	13,80%	147.293,40	-22,74%	156.971,50	-30,81%
8	68.000,00	68.270,63	-0,40%	69.892,33	-2,78%	57.743,62	15,08%
9	40.000,00	34.495,38	13,76%	26.403,62	33,99%	34.817,09	12,96%
10	170.000,00	138.725,60	18,40%	159.508,70	6,17%	160.250,10	5,74%
11	50.000,00	59.842,96	-19,69%	50.407,09	-0,81%	50.000,00	0,00%
12	60.000,00	70.007,96	-16,68%	80.267,51	-33,78%	54.584,42	9,03%
13	93.000,00	94.332,59	-1,43%	89.401,90	3,87%	84.663,29	8,96%
14	250.000,00	216.105,20	13,56%	206.583,00	17,37%	225.102,90	9,96%
15	65.000,00	47.884,77	26,33%	83.912,75	-29,10%	86.226,03	-32,66%
16	65.000,00	68.497,66	-5,38%	58.406,38	10,14%	59.516,79	8,44%
17	71.000,00	68.903,35	2,95%	70.054,15	1,33%	62.539,01	11,92%
18	220.000,00	238.017,00	-8,19%	278.144,20	-26,43%	256.561,30	-16,62%
19	200.000,00	229.705,80	-14,85%	193.097,70	3,45%	183.489,80	8,26%
20	90.000,00	90.232,79	-0,26%	115.127,10	-27,92%	125.127,60	-39,03%
21	140.000,00	145.853,40	-4,18%	150.750,60	-7,68%	159.736,60	-14,10%
22	250.000,00	160.269,20	35,89%	168.790,30	32,48%	177.219,10	29,11%
23	250.000,00	250.538,90	-0,22%	208.643,70	16,54%	236.247,70	5,50%
24	150.000,00	162.987,40	-8,66%	159.246,20	-6,16%	155.352,90	-3,57%
25	120.000,00	143.434,80	-19,53%	148.260,80	-23,55%	141.149,00	-17,62%
26	180.000,00	176.995,20	1,67%	148.774,20	17,35%	163.370,40	9,24%
27	250.000,00	227.360,70	9,06%	234.573,50	6,17%	245.999,50	1,60%
28	180.000,00	177.764,00	1,24%	148.774,20	17,35%	163.370,40	9,24%
29	250.000,00	228.129,50	8,75%	234.573,50	6,17%	245.999,50	1,60%
30	115.000,00	145.532,50	-26,55%	141.209,40	-22,79%	124.130,10	-7,94%
31	120.000,00	136.580,90	-13,82%	128.899,30	-7,42%	132.289,60	-10,24%
32	140.000,00	127.629,30	8,84%	116.589,10	16,72%	140.449,10	-0,32%
33	90.000,00	84.140,32	6,51%	119.424,30	-32,69%	130.124,10	-44,58%
34	65.000,00	77.515,85	-19,26%	72.862,38	-12,10%	57.095,17	12,16%
35	120.000,00	118.181,50	1,52%	117.913,90	1,74%	118.085,70	1,60%
36	210.000,00	221.732,20	-5,59%	231.330,10	-10,16%	195.785,10	6,77%
37	100.000,00	93.798,20	6,20%	103.761,90	-3,76%	100.447,40	-0,45%
38	70.000,00	66.740,97	4,66%	48.566,10	30,62%	48.822,15	30,25%
39	95.000,00	91.840,82	3,33%	61.654,72	35,10%	78.847,79	17,00%
40	30.000,00	36.676,36	-22,25%	37.858,92	-26,20%	39.372,14	-31,24%
41	40.000,00	40.674,19	-1,69%	75.418,86	-88,55%	72.693,41	-81,73%
42	100.000,00	107.639,10	-7,64%	91.639,18	8,36%	99.511,42	0,49%
43	90.000,00	82.785,98	8,02%	81.316,13	9,65%	74.125,76	17,64%
44	110.000,00	123.044,30	-11,86%	103.287,00	6,10%	102.533,60	6,79%

Estatísticas:	Grupo1	R^2 = 0.90062	F = 5.4375	p = 0.025266
	Grupo2:	R^2 = 0.78569	F = 5.8658	p = 0.00094988
	A Fatorial	R^2 = 0.845	F = 10.1761	p = 1.1396e-007
	A Simples	R^2 = 0.89329	F = 8.7702	p = 1.8882e-006

APÊNDICE VI – QUADRO DE PREÇOS PREDECIDOS DE CASAS

casas	Valor	Avaliando por opções					
n		Agrupam.	Dif(%)	A Fatorial	Dif(%)	Simples	Dif(%)
1	120.000,00	110.594,00	7,84%	113.853,70	5,12%	115.887,50	3,43%
2	160.000,00	165.112,10	-3,20%	163.207,30	-2,00%	163.868,00	-2,42%
3	110.000,00	104.565,00	4,94%	104.270,90	5,21%	109.453,30	0,50%
4	70.000,00	83.244,11	-18,92%	57.014,11	18,55%	68.990,58	1,44%
5	85.000,00	82.956,91	2,40%	106.292,10	-25,05%	92.775,60	-9,15%
6	100.000,00	100.432,40	-0,43%	104.080,60	-4,08%	93.639,03	6,36%
7	160.000,00	153.527,80	4,05%	128.716,00	19,55%	123.567,30	22,77%
8	30.000,00	29.620,70	1,26%	30.117,58	-0,39%	6.944,65	76,85%
9	28.000,00	24.752,10	11,60%	50.731,80	-81,19%	33.394,77	-19,27%
10	35.000,00	38.939,61	-11,26%	36.217,29	-3,48%	45.126,94	-28,93%
11	150.000,00	151.772,10	-1,18%	174.500,10	-16,33%	162.663,40	-8,44%
12	45.000,00	46.174,35	-2,61%	56.274,48	-25,05%	49.576,45	-10,17%
13	30.000,00	29.383,03	2,06%	65.388,98	-117,96%	70.011,60	-133,37%
14	28.000,00	26.780,53	4,36%	55.580,93	-98,50%	49.383,38	-76,37%
15	140.000,00	132.490,30	5,36%	126.013,80	9,99%	145.050,70	-3,61%
16	130.000,00	142.401,60	-9,54%	144.192,00	-10,92%	133.488,70	-2,68%
17	50.000,00	18.330,54	63,34%	56.810,92	-13,62%	21.983,31	56,03%
18	150.000,00	150.006,90	0,00%	136.267,60	9,15%	135.228,30	9,85%
19	150.000,00	164.577,60	-9,72%	174.180,50	-16,12%	175.935,70	-17,29%
20	135.000,00	159.128,00	-17,87%	157.660,00	-16,79%	158.492,30	-17,40%
21	180.000,00	181.586,70	-0,88%	179.143,70	0,48%	182.183,40	-1,21%
22	40.000,00	44.249,19	-10,62%	48.134,95	-20,34%	66.740,72	-66,85%
23	150.000,00	151.414,40	-0,94%	185.436,60	-23,62%	159.611,20	-6,41%
24	60.000,00	62.882,68	-4,80%	52.656,80	12,24%	47.960,57	20,07%
25	15.000,00	15.791,10	-5,27%	23.717,74	-58,12%	3.734,20	75,11%
26	45.000,00	42.827,20	4,83%	27.195,82	39,56%	26.349,99	41,44%
27	165.000,00	164.211,70	0,48%	192.324,10	-16,56%	187.348,60	-13,54%
28	95.000,00	94.460,51	0,57%	30.719,73	67,66%	48.847,10	48,58%
29	100.000,00	127.596,00	-27,60%	91.153,03	8,85%	101.837,50	-1,84%
30	130.000,00	117.345,10	9,73%	126.399,80	2,77%	121.766,50	6,33%
31	90.000,00	90.366,90	-0,41%	100.070,50	-11,19%	98.746,20	-9,72%
32	185.000,00	166.023,60	10,26%	177.669,10	3,96%	184.820,20	0,10%
33	90.000,00	90.626,75	-0,70%	97.081,05	-7,87%	95.427,06	-6,03%
34	15.000,00	15.620,85	-4,14%	72.547,27	-383,65%	75.075,57	-400,50%
35	150.000,00	134.092,50	10,61%	110.897,80	26,07%	113.211,80	24,53%
36	150.000,00	146.523,20	2,32%	131.727,10	12,18%	138.145,40	7,90%
37	140.000,00	149.735,80	-6,95%	92.802,93	33,71%	100.785,00	28,01%
38	30.000,00	12.861,83	57,13%	35.934,48	-19,78%	34.341,63	-14,47%
39	60.000,00	64.173,03	-6,96%	89.826,86	-49,71%	92.107,29	-53,51%
40	45.000,00	51.735,79	-14,97%	50.189,58	-11,53%	47.203,08	-4,90%
41	30.000,00	30.220,23	-0,73%	64.558,87	-115,20%	44.906,37	-49,69%
42	90.000,00	87.620,67	2,64%	107.719,00	-19,69%	111.957,20	-24,40%
43	50.000,00	50.693,19	-1,39%	55.709,54	-11,42%	47.597,98	4,80%
44	70.000,00	67.648,81	3,36%	58.913,22	15,84%	42.390,55	39,44%
45	130.000,00	139.631,10	-7,41%	121.947,70	6,19%	104.504,20	19,61%
46	180.000,00	178.216,50	0,99%	141.623,20	21,32%	152.411,00	15,33%
47	290.000,00	243.132,70	16,16%	194.438,90	32,95%	226.028,60	22,06%
48	50.000,00	51.443,33	-2,89%	34.171,09	31,66%	47.394,97	5,21%

49	98.000,00	139.043,20	-41,88%	123.253,40	-25,77%	126.771,80	-29,36%
50	65.000,00	65.899,25	-1,38%	61.383,24	5,56%	98.377,96	-51,35%
51	40.000,00	41.536,25	-3,84%	46.411,45	-16,03%	49.954,79	-24,89%

Estatísticas

grupo1	R^2 = 0.99675	F = 30.7138	p = 0.13959
grupo 2	R^2 = 0.91233	F = 14.5695	p = 9.5501e-006
grupo 3	R^2 = 0.95092	F = 3.8749	p = 0.36694
grupo 4	R^2 = 0.99938	F = 321.2337	p = 0.04233
A Fatorial	R^2 = 0.78469	F = 14.5778	p = 1.9472e-010
A Simples	R^2 = 0.8366	F = 9.1021	p = 5.0106e-008

APÊNDICE VII – QUADRO DE PREÇOS PREDECIDOS DE TERRENOS

terreno	preço	Avaliação por opções					
n		agrupamento	Dif.(%)	A Fatorial	Dif.(%)	A Simples	Dif.(%)
1	100.000,00	103.178,10	-3,18%	148.873,60	-48,87%	141.191,40	-41,19%
2	25.000,00	24.905,15	0,38%	27.449,32	-9,80%	23.433,55	6,27%
3	43.000,00	48.369,13	-12,49%	28.653,89	33,36%	45.305,46	-5,36%
4	33.000,00	28.632,82	13,23%	34.573,59	-4,77%	21.449,89	35,00%
5	40.000,00	37.169,79	7,08%	58.727,13	-46,82%	67.777,74	-69,44%
6	140.000,00	140.000,00	0,00%	151.669,00	-8,34%	145.660,40	-4,04%
7	90.000,00	90.000,00	0,00%	94.309,39	-4,79%	87.205,11	3,11%
8	38.000,00	34.325,81	9,67%	24.052,34	36,70%	21.817,83	42,58%
9	70.000,00	77.048,69	-10,07%	109.140,50	-55,92%	118.710,70	-69,59%
10	145.000,00	140.033,70	3,43%	192.995,10	-33,10%	188.357,90	-29,90%
11	15.000,00	19.406,38	-29,38%	11.801,99	21,32%	22.079,11	-47,19%
12	13.000,00	8.133,56	37,43%	8.011,58	38,37%	5.920,89	54,45%
13	100.000,00	100.000,00	0,00%	128.921,60	-28,92%	130.887,20	-30,89%
14	90.000,00	90.000,00	0,00%	94.309,39	-4,79%	87.205,11	3,11%
15	250.000,00	250.000,00	0,00%	213.707,10	14,52%	219.718,00	12,11%
16	38.000,00	41.672,51	-9,66%	30.302,31	20,26%	36.052,28	5,13%
17	300.000,00	300.000,00	0,00%	213.253,30	28,92%	220.724,50	26,43%
18	47.000,00	42.870,81	8,79%	27.946,71	40,54%	41.276,45	12,18%
19	60.000,00	51.262,84	14,56%	55.042,66	8,26%	56.669,23	5,55%
20	50.000,00	51.225,11	-2,45%	24.814,66	50,37%	24.990,80	50,02%
21	15.000,00	15.444,18	-2,96%	17.205,92	-14,71%	16.566,45	-10,44%
22	30.000,00	34.402,67	-14,68%	32.598,87	-8,66%	14.465,10	51,78%
23	25.000,00	31.039,01	-24,16%	37.378,36	-49,51%	23.344,11	6,62%
24	30.000,00	27.879,77	7,07%	21.261,69	29,13%	26.190,86	12,70%

Estatísticas

Grupo 1	R^2 = 0.97995	F = 34.208	p = 5.391e-005
Grupo 2	R^2 = 1	F = NaN	p = NaN
A Fatorial	R^2 = 0.84911	F = 20.2585	p = 8.0357e-007
A Simples	R^2 = 0.8643	F = 8.2799	p = 0.00037022